

doi: 10.12052/gdutxb.220170

基于方向控制的差分隐私轨迹数据发布方法

李 杨, 周 莹

(广东工业大学 计算机学院, 广东 广州 510006)

摘要: 随着差分隐私研究及其应用的不断拓展, 其在轨迹数据发布的隐私保护领域应用受到了广泛关注, 现有研究方法大多采用Kmeans聚类方法对轨迹进行聚类划分, 但由于差分隐私约束下的轨迹数据集受到噪声的扰动, 导致现有的聚类方法无法保证最后的收敛效果。本文提出了一种基于方向控制的差分隐私保护轨迹数据发布方法。首先, 提出了基于SKmeans||聚类的轨迹泛化算法, 在聚类迭代过程中针对质心的更新, 加入方向控制机制, 设计指数机制中的打分函数控制质心的收敛, 保证高维数据聚类的质量。其次, 设计了一个基于有界阶梯噪声机制的轨迹数据发布算法, 其中的有界阶梯噪声机制保证了在隐藏轨迹点真实计数的同时, 提高了发布后轨迹数据的可用性。最后, 通过实验验证了本文所提出方法的有效性。

关键词: 差分隐私; 聚类; 轨迹数据发布; 方向控制; 有界噪声

中图分类号: TP391

文献标志码: A

文章编号: 1007-7162(2023)05-0056-08

Differential Privacy Trajectory Data Publishing Based on Orientation Control

Li Yang, Zhou Ying

(School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: With the continuous expansion of differential privacy and its applications, its application in the privacy protection field of trajectory data release has received extensive attention. However, most existing research methods use the Kmeans to cluster the trajectory, which cannot guarantee the final convergence due to the fact that the trajectory datasets under differential privacy constraints are usually disturbed by noise. To address this, this paper proposes an orientation control-based differential privacy-preserving trajectory data publishing method. Firstly, a trajectory generalization algorithm based on SKmeans|| clustering is proposed, which updates the centroid via a direction control mechanism during iterative process of clustering, and designs a scoring function in the index mechanism to control the convergence of the centroid, such that the quality of high dimensional data clustering can be improved. Secondly, a trajectory data publishing algorithm based on bounded noise mechanism is designed, which improves the availability of trajectory data after publishing. Meanwhile, the bounded noise mechanism ensures the true count of the hidden trajectory. Finally, the effectiveness of the method proposed in this paper is evaluated by experiments.

Key words: differential privacy; clustering; trajectory publishing; orientation control; bounded noise

随着物联网、智能电子设备以及定位系统的快速发展, 交通探头、交通刷卡机等智能设备更容易生成和收集轨迹数据。这些轨迹由一系列位置组成, 代表着用户的日常轨迹活动。虽然轨迹数据具有很大

的应用价值, 但轨迹数据中包含的敏感信息^[1], 例如家庭住址、工作地点等, 一旦泄露, 对用户的生命财产安全会构成极大的威胁。然而传统的数据收集方法, 例如k-匿名^[2], 攻击者可以通过获取额外的背景

收稿日期: 2022-11-16

基金项目: 国家自然科学基金资助项目(61972102); 广东省重点研发项目(2021B0101220004)

作者简介: 李杨(1980-), 女, 教授, 博士, 主要研究方向为多智能体、差分隐私保护

通信作者: 周莹(1998-), 女, 硕士研究生, 主要研究方向为差分隐私保护, E-mail: joying314@163.com

知识来分析用户在某段时间内的轨迹数据、行动路线等,导致无法抵御背景知识攻击^[3]。传统的k-匿名方法即使采用了唯一标志符,仍无法阻止用户的行为被追溯。因此,迫切需要实现隐私保护的轨迹数据发布方案。

Dwork^[4]于2006年提出了差分隐私的概念,通过在查询或分析结果添加噪声的方式,防止数据库中的信息被泄露。由于差分隐私严格的数学统计模型与可量化的噪声机制,研究者们开始将其应用在轨迹发布领域,并取得了良好的效果。Chen等^[5]首次提出基于前缀树的数据发布模型,随后扩展了前缀树的工作,提出了基于可长度 n -gram的方法^[6]。Zhao等^[7]也提出了一种 r 树结构,由轨迹序列组成 r 树,从而保证轨迹的时空完整性。上述工作都需要满足轨迹必须要有相同前缀的假设,这并不符合大多数轨迹的特性。

因此,Hua等^[8]提出了一种基于聚类的轨迹泛化方法,Li等^[9]拓展了该方法,提出了一种具有有界拉普拉斯噪声的轨迹发布方法,但是Kmeans聚类受初始值影响较大,且在有限的迭代次数中无法提供终止聚类算法有效解。并且,已有工作表明,在复杂高维数据面前,拉普拉斯机制并不是最优的噪声添加机制。Quan等^[10]提出一种优化的阶梯机制,与拉普拉斯机制相比,在同等隐私预算下,阶梯机制的噪声添加量会更小。现有的轨迹隐私保护工作,Kmeans方法受初始点和离群点的影响,迭代过程无法保证质心的收敛,从而结果不稳定,对随后的轨迹映射的准确性也会造成影响。同时,在提高发布轨迹的统计,保证发布的轨迹数据的隐私效用时,会在一定程度上忽略掉数据的可用性。

为此,本文提出了一种差分隐私轨迹数据发布方法(Differential Privacy Data Publishing with SKmeans||, SKTDP),集成了方向控制机制的SKmeans||聚类轨迹泛化算法和阶梯有界噪声机制的轨迹发布算法。首先,结合SKmeans||聚类和指数机制对轨迹点进行区域划分以及泛化映射。其次,提出了一种方向控制机制来保证聚类的迭代收敛,实验结果表明,本文的方法提高了聚类的质量。

本文的主要工作与贡献如下:

(1) 提出了一种基于SKmeans||聚类的轨迹泛化算法(Trajectory Generalization Based on SKmeans|| Clustering, TGSC),通过在SKmeans||算法中添加方向控制机制,确保了聚类泛化质量,消除了现有机制对

隐式假设的依赖,提高了聚类效果和数据效用。

(2) 设计了基于阶梯有界噪声机制的轨迹发布算法(Differentially Private Publishing with Bounded Staircase Noise, SNDP),该算法相较于传统的拉普拉斯机制,添加的噪声量更小。阶梯有界噪声机制基于一般的效用最大化或成本最小化框架,在具有阶梯分布的机制中添加查询输出的噪声是最优的,且该分布在原点附近对称,在 $x \geq 0$ 时是单调递减且几何衰减的阶梯形概率密度函数。

(3) 在真实轨迹数据集上验证了该方法的可行性,发布后轨迹数据具有较好的可用性。

1 相关工作与问题描述

1.1 相关工作

差分隐私^[4]的提出是为了解决隐私保护的数据分析问题。采用基于差分隐私的隐私保护模型,不仅可以有效防止背景知识攻击,还在一定程度保证了数据的统计学特性。Chen等^[5]首次提出了基于混合粒度前缀树结构的SeqPT模型,用前缀树表示轨迹序列数据,并利用前缀树的节点来保存轨迹子序列的噪声计数,对轨迹数据有很好的可用性。后来,Chen等^[6]拓展了基于前缀树的工作,提出了一种可变长度 n -gram模型来处理序列数据。Li等^[11]提出的前缀树结构,取消了每一层的分类树,并设计了增量隐私预算分配模型和时空维度缩减方法,提高了算法的运行效率和数据效用。Gursoy等^[12]提出了当前较为先进的DP-Star框架,利用特征提取、特征学习、噪声扰动和合成来实现隐私保护。然而,上述工作假设原始轨迹包含相同的前缀或长度为 n 的公共子序列,这在许多应用中都很难满足。

为了消除文献^[5-6]中的轨迹数据必须要有相同前缀的假设,Hua等^[8]提出了基于Kmeans聚类与指数机制结合的位置泛化算法。随后,Li等^[9]拓展了该方法,提出了一种具有有界拉普拉斯噪声的轨迹发布方法。由于Kmeans聚类受初始值影响较大,且在有限的迭代次数中无法提供终止聚类算法的有效解,Ni等^[13]提出了一种新的聚类算法DP-KCCM,通过添加自适应噪声以及合并聚类,显著提高了聚类的效用。而Lu等^[14]提出了一种基于交互式差分隐私的Kmeans聚类方法,该方法在质心迭代中的受控方向上应用ExpDP方法,然后注入有界拉普拉斯噪声,以保护数据隐私免受推理攻击,从而保证了聚类的收敛。与该文献不同的是,本文采用非交互式差分隐

私,同时为了适应高维数据的特征,在SKmeans||聚类中设计指数机制来保证收敛。Liu等^[15]在提出了两种基于Kmeans||聚类的轨迹合并方案,使用Kmeans||聚类算法对定位区域进行聚类,聚类中的所有点都被聚类的中心所取代以及利用楼梯机制来扰乱集群中心,以提高隐私保护的水平。Joonas等^[16]提出了两种新的Kmeans||聚类初始化方法,通过采样、子采样和降维来减少数据处理,适用于初始化大规模问题。

此外,随着机器学习的不断发展,越来越多机器学习与差分隐私的融合技术产生。Xu等^[17]提出了一种差异隐私潜在轨迹组发现方案(DP-LTOD),将原始轨迹序列模糊为差异隐私保证的轨迹序列,以保持轨迹隐私,并通过对上传的轨迹序列组进行聚类来发现潜在的轨迹。陈思等^[18]通过建立新的基于时间泛化和空间分割的高效采样模型,并利用Kmeans聚类算法进行抽样数据处理,同时借助差分隐私保护机制对轨迹数据进行双重扰动,有效解决了具有强大背景知识的攻击者窃取用户隐私的问题。Zhao等^[19]提出了一种基于聚类的差异隐私轨迹保护方法,首先将拉普拉斯噪声添加到集群中的轨迹位置计数中,然后在聚类中的轨迹位置数据中加入有限半径的拉普拉斯噪声,以避免过度的噪声影响聚类。

1.2 理论基础

差分隐私是建立在数学理论上,严格且可证明的数学模型。对于两个相邻数据集,经过差分隐私的处理,可以发布两个不存在明显差异的数据集,即无法识别数据集中的单条轨迹。因此,可以有效防止背景知识攻击,保护用户的隐私。

定义1^[4] 对相邻数据集 D_1, D_2 ,一种随机算法 A , $\text{Range}(A)$ 表示算法 A 所有可能的输出集合,有 $S \in \text{Range}(A)$ 。若对于所有的 S ,算法 A 输出 S 的概率为 $\Pr[A(D_1) = S]$,都有

$$\Pr[A(D_1) = S] \leq e^\epsilon \times \Pr[A(D_2) = S] \quad (1)$$

则随机算法 A 满足 ϵ -差分隐私。式中:隐私预算 ϵ 越小,噪声添加量越大,隐私级别越高;隐私预算越大则相反; e 为自然对数的底。

定义2^[4] 用敏感度表示查询结果变化的最大值,对函数 $f: D_1 \rightarrow R^m$,输入为数据集 D_1 ,输出为 m 维向量。则敏感度 Δf 为

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\| \quad (2)$$

式中: D_1, D_2 互为相邻数据集, $\|\cdot\|$ 表示一阶范式。

定理1^[10](阶梯机制) 对一个多维查询函数

$f: D \rightarrow R^d$,staircase-shaped机制会输出

$$A(D) = f(D) + \text{staircase-shaped}(\Delta f, \epsilon, \gamma)^d \quad (3)$$

令 $\text{staircase-shaped}(\Delta f, \epsilon, \gamma)$ 表示要添加的噪声,它是具有阶梯概率密度函数的随机变量,阶梯概率密度函数表示为

$$f_\gamma(x) = \begin{cases} a(\gamma), & x \in [0, \gamma \Delta f] \\ e^{-\epsilon} a(\gamma), & x \in [\gamma \Delta f, \Delta f] \\ e^{-h\epsilon} f_\gamma(x - h\Delta f), & x \in [h\Delta f, (h+1)\Delta f] \\ f_\gamma(-x), & x < 0 \end{cases} \quad (4)$$

式中: h 为比例系数, Δf 为敏感度, $a(\gamma)$ 是一个归一化常数, γ 是损失函数的最优值,计算方法为

$$a(\gamma) \triangleq \frac{1 - e^{-\epsilon}}{2\Delta f(\gamma + e^{-\epsilon}(1 - \gamma))} \quad (5)$$

$$\gamma = -\frac{b}{1-b} + \frac{(b-2b^2+2b^4-b^5)^{\frac{1}{5}}}{2^{\frac{1}{5}}(1-b)^2}, b = e^{-\epsilon} \quad (6)$$

定理2(指数机制) 设得分函数 $q(D, R)$,输入数据集 D ,输出为一实体集 R ,若对一随机算法 A 以正比于 $\exp(\epsilon q(D, R)/(2\Delta q))$ 的概率输出结果 $y \in R$,则称随机算法 A 满足 ϵ -差分隐私。

1.3 问题描述

本文研究的问题是在给定的轨迹数据集 D 和隐私预算 ϵ 下,构建差分隐私下基于SKmeans||聚类的轨迹泛化算法,然后构建基于有界阶梯噪声机制的轨迹数据发布算法,使得发布的数据集 D 在满足差分隐私的同时,保证聚类最后的集群质量,同时保证发布轨迹集具有较高的数据可用性。数据效用的衡量主要是以下两个方面,一是衡量原始数据集与重构轨迹数据集之间的相似性,二是范围查询准确度。在本文中,数据集相似性通过豪斯多夫聚类来衡量,而范围查询准确度则通过F-measure分数来衡量。最后,在这里介绍关于轨迹的定义。

定义3 轨迹是按时间顺序排列的时间地点戳对,即 $T = (t_1, l_1) \rightarrow \dots \rightarrow (t_i, l_i) \rightarrow \dots \rightarrow (t_{|T|}, l_{|T|})$,其中 $l_i(1 \leq i \leq |T|)$ 是由经度和纬度坐标表示的离散空间点, $|T|$ 是轨迹的数量。

2 方法描述及分析

本文提出的满足差分隐私保护的轨迹数据发布方法包括两个算法,基于Skmeans||聚类的轨迹泛化算法(TGSC)和基于阶梯有界噪声机制的轨迹发布算法(SNDP)。

由于传统Kmeans聚类在初始点选择和质心迭代的过程中存在局限性,无法在有限的迭代次数内

提供终止聚类的有效解,从而影响泛化的质量。因此,TGSC算法使用SKmeans||算法,先获取一组候选质心集合,然后基于该集合再生成 k 个质心,形成 k 个分组,保证了选取的初始点距离较远且都落在密集区域。在聚类迭代过程中加入了方向控制机制,设计了指数机制中的打分函数选择偏移后的质心,确保聚类一定能够收敛,从而提高聚类的质量和泛化的效果。

其次,传统的拉普拉斯机制在轨迹等高维数据中存在噪声添加量过大的问题,SNDP算法结合阶梯有界噪声添加机制,设计了噪声添加阈值,在同等的隐私保护级别下添加的噪声量更小,保护数据隐私的同时提高了数据的可用性。

实验结果表明,该方法发布的重构数据集在满足差分隐私保护的同时,有效降低了轨迹合并的时间成本,具有较高的数据可用性。

2.1 基于SKmeans||聚类的轨迹泛化算法(TGSC)

TGSC算法使用SKmeans||算法对每个时间戳下的轨迹点进行聚类划分。为提高聚类的效果,在迭代质心更新的过程中,加入了方向控制机制,使用指数机制对质心进行距离和方向偏移,期间设计了指数机制的打分函数,更契合高维轨迹数据。

算法1: TGSC算法流程

输入: 原始数据集 D , 时间域 D_t , 候选划分数量 φ , 隐私预算 ε_1

输出: 泛化后的轨迹集 D_1

```

01.  $D_1 = []$ 
02. For  $i = 1:\text{len}(D_t)$  do
03.    $D_i = [T_i \text{ for } T \text{ in } D]$ 
04.    $\text{init\_k} = \text{SKmeans}||(\mathcal{D}_i, k)$ 
05.   For  $t = 1:\text{max\_iter}$  do
06.     For cluster  $i$  in  $t$  do
07.        $c[i] = \text{each } D_i \text{ closest with cent}[t-1]$ 
08.        $\text{cent\_k}[t] = \text{means of } c[i]$ 
09.        $\text{Szone} = \text{TSZG}(\text{cent\_k}[t], \text{cent\_k}[t-1])$ 
10.        $\text{New\_k} = \text{ExpDP}(\text{Szone}, \varepsilon_1)$ 
11.     End for
12.   End for
13. End for
14.  $\text{Temp} = \text{map}(D, \text{New\_k})$ 
15. For  $T_i$  in  $\text{temp}$ 
16.   while  $\text{Haud}(T_i[i], T_i[j]) \leq \text{threshold}$ 
17.      $D_1.\text{add}(T_i)$ 
18.   End while
19. End for
20. return  $D_1$ 

```

算法1首先对每一个时间戳下的轨迹点聚类(02~13行),先通过SKmeans||生成初始点(04行),其次在聚类的每轮迭代中(05~12行),采用文献[14]中的TSZG算法,设计概率函数得到采样区的中心点以及半径,建立收敛区和采样区(09行)。设本轮迭代得到的簇质心为 c_t ,上一轮迭代得到的簇质心为 c_{t-1} ,通过指数机制偏移得到的质心为 c'_t ,可以重新设计得到距离偏移函数和角度偏移函数为

$$d_t = \frac{\|c_t - c'_t\|}{\|c_t - c_{t-1}\|} \quad (7)$$

$$\theta_t = \angle c'_t c_t c_{t-1} \quad (8)$$

式中: $d_t \in (0, 1)$, $\theta_t \in (-\pi/2, \pi/2)$ 。对聚类当前迭代的任意一个簇 t ,可以得到打分函数为

$$\text{score}(d_t, \theta_t) = (1 - d_t) + \left(1 - \frac{2\theta_t}{\pi}\right) \quad (9)$$

通过打分函数得到簇中的偏移质心(10行),与文献[14]不同的是没有在当前迭代过程中再利用聚类重新迭代,去除了冗余的操作,缩短了迭代时间,从而有效提高了聚类的效率,具有更好的聚类质量。然后将簇中的其他轨迹点映射为质心,再将不同时间戳下的轨迹连接起来(14行),得到重构的泛化轨迹。为了保持轨迹集的种类一致,以及连续两个位置之间的距离受到速度的限制,将不满足速度限制的 n 条轨迹筛选掉(15~19行)。

2.2 基于阶梯有界噪声机制的轨迹发布算法(SNDP)

SNDP算法包括有界阶梯噪声添加模块和一致性约束模块。首先,有界阶梯噪声添加模块假设有 n 个包含原始轨迹的泛化轨迹,通过在真实计数上添加有界阶梯分布的噪声,得到相应的噪声计数。然后,通过一致性约束模块,对轨迹进行后处理并发布,使轨迹更符合真实情况。

算法2: SNDP算法流程

输入: 泛化轨迹 D_1 , 隐私预算 ε_2 , 敏感度 Δf , 每条轨迹的真实计数 $\{t_1, t_2, \dots, t_n\}$, α, β

输出: 重构发布的轨迹数据集 D'

```

01.  $\Delta f = \max_{p,q} \{|t_p - t_q|\}$ 
02. For  $p, q$  from 1 to  $n$  do
03.    $\mu = \text{avg}(t_p), \eta = \min\{t_p\}$ 
04. End for
05.  $\alpha = \eta/2, \beta = 2\mu$ 
06. For  $i$  in  $N$  do
07.   if  $i < n$ :
08.      $s_i = t_i + \text{staircase\_shaped}(\Delta f, \varepsilon_2, \gamma)$ 
09.   else if

```

```

10.    $s_i = \text{staircase\_shaped}(\Delta f, \varepsilon_2, \gamma)$ 
11.   End if
12.   if  $s_i \geq \beta$  then  $s_i = \beta$ 
13.   if  $s_i \leq \alpha$  then  $s_i = \alpha$ 
14. End for
15.  $\text{NC} = \langle s_1, s_2, \dots, s_N \rangle$ 
16. Sorted NC and  $S = \langle S'_1, S'_2, \dots, S'_N \rangle$ 
17.  $Q_m = \max_{i \in [1, m]} \min_{j \in [i, n]} \frac{\sum_{m=1}^j tc(v_m)}{j-i+1}$ 
18.  $S' = \langle Q_1, Q_2, \dots, Q_N \rangle$ 
19. For  $i$  in  $N$  do
20.    $s'_i = Q_i$ 
21. End for
22. for  $i=1:N$  do
23.    $D'.\text{add}(T)$  based on  $P(s'_i/\text{sum}(S'))$ 
24. End for
25. Return  $D'$ 

```

算法2包括有界阶梯噪声添加和一致性约束两个模块。前半部分(01~11行)为噪声添加模块,先计算出敏感度(01行),由敏感度定义可知这里敏感度为1。假设轨迹范围属于 $[\alpha, \beta]$,而阶梯机制噪声属于 $[\gamma, \eta]$,噪声计数等于 t 加阶梯噪声 sn ,如果 $t + \text{sn} < 0$,则噪声计数为负,这会是一个毫无意义的轨迹发布。虽然噪声计数可以超过轨迹上限 β ,但如果 $t + \text{sn} > \beta$,则这种泛化轨迹已经很拥挤了。因此,噪声阈值设置模块(02~04行)先计算轨迹真实计数的平均值和最小值(03行),并将噪声上界设为平均值的2倍,噪声下界设为最小值的1/2(05行)。有界阶梯噪声添加模块(06~14行)通过定理1给出的staircase-shaped机制对轨迹的真实计数进行加噪,若噪声计数超过阈值,则用阈值替换噪声计数。

假设一对邻近轨迹数据集 D 和 D' 不同的那条轨迹为 T_0, T'_0 为 T_0 对应的泛化轨迹,则从序列 $s_i(D)$ 中输出候选项 $o = \{o_1, o_2, \dots, o_{|\Omega|}\}$ 的概率为 $\Pr[s(D) = o] = \prod_i \Pr[s_i(D) = o_i]$;对于任何泛化轨迹 $T'_i \neq T'_0$,有 $\Pr[s_i(D) = o_i] = \Pr[s_i(D') = o_i]$;对于任何 $T'_i \neq T'_0 \wedge T'_0 \in D', T'_0$ 的噪声计数等于真实计数加staircase-shaped噪声,有 $\Pr[s(D) = o] < e^{\varepsilon_2} \cdot \Pr[s(D') = o]$;对任意 $T'_i \neq T'_0 \wedge T'_0 \notin D' \wedge o \neq \emptyset$,有 $\Pr[s(D) = o] = \Pr[s(D') = o]$,因此有界的staircase-shaped噪声是符合差分隐私定义的。

后半部分(15~21行)为一致性约束模块,在无噪声图中,对于任意的两条边 v, u ,如果边 v 的轨迹多于边 u ,则有 $t(v) > t(u)$ 。因此,对于有噪声的轨迹,也通过图来映射轨迹集,基于文献[9]的方法,先生成图 G 中每条边的噪声计数,并组成序列 S ,令 $\text{mean}[i, j]$ 表示其中一条子序列的平均数

$$\text{mean}[i, j] = \frac{\sum_{m=1}^j t(v_m)}{j-i+1} \quad (10)$$

$$Q_m = \max_{i \in [1, m]} \min_{j \in [i, |E(G)|]} \text{mean}[i, j] \quad (11)$$

然后可以得到预期序列 S'

$$S' = \langle Q_1, Q_2, \dots, Q_{|E(G)|} \rangle \quad (12)$$

最后,按照噪声计数与序列和的比值输出与原轨迹同样数量的重构轨迹并发布。

2.3 隐私分析

在本文提出的方法中,主要包括两个算法。TGSC算法对每个时间戳下的地点域进行聚类,运用指数机制获取更符合实际情况的聚类质心,然后映射得到泛化轨迹。SNDP算法采用有界的staircase-shaped机制进行加噪,设计噪声阈值,防止添加噪声过大或者过小。可以得到以下定理。

定理3 TGSC算法满足 ε_1 -差分隐私。

证明 对于得到聚类中心的指数机制,有

$$\frac{\Pr[c'_i = c]}{\Pr[c''_i = c]} = \frac{\Pr[c, c'_i, c''_i] \times \exp\left(\frac{\varepsilon'_i q(d'_i, \theta'_i)}{2\Delta q}\right)}{\Pr[c, c'_i, c''_i] \times \exp\left(\frac{\varepsilon'_i q(d''_i, \theta''_i)}{2\Delta q}\right)} = \exp\left(\varepsilon'_i \cdot \frac{q(d'_i, \theta'_i) - q(d''_i, \theta''_i)}{2\Delta q}\right) \leq \exp(\varepsilon'_i) \quad (13)$$

式中: c'_i 为第 i 次迭代 t 簇的质心, c''_i 为第 i 次迭代 t 簇的加噪质心。在一轮迭代中,均满足 ε'_i -差分隐私。因为在每次迭代中,所有地点域都被分成 k 个簇,基于顺序组合特性,泛化方法满足 ε_1 -差分隐私,其中 $\varepsilon_1 = \sum_{i=1}^T \max_{i=1}^k \{\varepsilon'_i\}$ 。

定理4 SNDP算法满足 ε_2 -差分隐私。

证明 在轨迹发布过程,隐私分析集中在加噪声的发布。根据并行组合特性,消耗总预算为 ε_2 ,有

$$\frac{\Pr[s(D) = s]}{\Pr[s(D') = s]} = \frac{\frac{1}{\beta-\alpha} \int_{\alpha}^{\beta} \text{pdf}(x) dx}{\frac{1}{\beta-\alpha} \int_{\alpha}^{\beta} \text{pdf}(x) dx} = 1 \leq e^{\varepsilon_2} \cdot \Pr[s(D') = s] \quad (14)$$

式中: $\text{pdf}(x)$ 为有界噪声函数, α, β 为噪声的上下界。因此,SKTDP方法总体满足 $\varepsilon_1 + \varepsilon_2$ -差分隐私。

3 实验分析

本文提出的方法基于Python实现。实验的硬件环境为双核Intel Core i5 1.60 GHz,内存为16 GB。本文使用的数据集为微软提供的公开数据集T-Drive^[20],

包含了北京地区10 357辆出租车一周的出行轨迹数据,由出租车ID、时间、经纬度位置组成,从中选取850条轨迹,并进行预处理。选取每天上午8:30到下午14:30的时间段,并离散为32个时间点,其中任意相邻的两个时间点相差大概10 min。在接下来的实验中,用SKTDP表示本文提出的方法,通过文献[9]和文献[15]的IS17方法和DP21方法,对比验证本文提出方法的有效性。

3.1 评估指标

3.1.1 轨迹相似度

豪斯多夫距离是度量空间中任意两个集合之间定义的一种距离,本文使用豪斯多夫距离来衡量原始轨迹集与重构轨迹集之间的轨迹相似度。

$$H(D) = \max \{h(D, D'), h(D', D)\} \quad (15)$$

式中:

$$h(D, D') = \max_{T \in D} \{\min_{T' \in D'} \{\text{Distance}(T, T')\}\} \quad (16)$$

式中:Distance(T, T')为距离函数。如果豪斯多夫距离越小,则说明两个数据集的相似度越高,对应数据可用性也越高。

3.1.2 计数查询误差

为了评价重构数据集的可用性,可以通过计数查询来定义相对误差,计数查询的结果是轨迹集中拥有相同前缀的子轨迹的数量。因此,本文规定,计数查询 $q(D)$,在地图上任取一点为圆心作圆,计算轨迹集 D 中经过该圆的轨迹数。则相对误差定义为

$$e = \frac{|q(D') - q(D)|}{\max \{q(D), \eta\}} \quad (17)$$

式中: η 为阈值,为原轨迹数据集数量的0.1%。实验运行100次,并取平均值为最终结果。

3.1.3 范围查询精度

因为发布重构轨迹集的目的是为了对其进行查询,采用窗口范围查询对轨迹数据集 D 和 D' 的查询结果进行比较。给定一个范围查询RQ, R_d 为原始轨迹集的查询结果, $R_{d'}$ 为重构轨迹集的查询结果,则范围查询的查询精度和召回率定义为

$$\text{prec} = \frac{|R_d \cap R_{d'}|}{R_{d'}} \quad (18)$$

$$\text{recall} = \frac{|R_d \cap R_{d'}|}{R_d} \quad (19)$$

将二者进行加权,可以得到

$$\text{F-measure} = \frac{2 \times \text{prec} \times \text{recall}}{\text{prec} + \text{recall}} \quad (20)$$

使用F-measure作为指标,随机运行100次,取平均值为最终结果。

3.2 结果分析

3.2.1 方法评估

图1为本文提出的方法在不同隐私预算下的计数查询误差和F-measure度量。从图1(a)可以看出,随着分组数 k 和隐私预算的增加,计数查询误差是逐渐降低的。从图1(b)可以看出,随着隐私预算的增加,对于不同的查询半径 r , F-measure度量也随之增加,但随着查询窗口的增加而减小。从两个图中可以看出,随着隐私预算的增加,计数查询误差呈减小的趋势, F-measure度量呈增加的趋势,都满足差分隐私的定义。

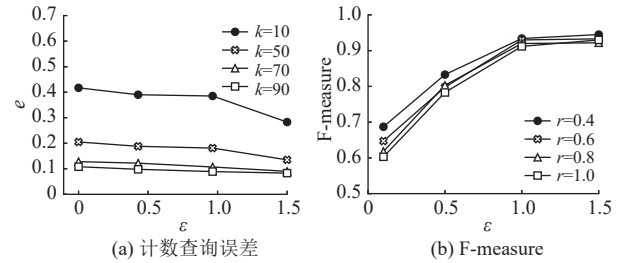


图1 不同隐私预算下的指标

Fig.1 Proxies with different epsilon

3.2.2 对比分析

本文在不同隐私预算和不同聚类分组数下,对比方法IS17和DP21进行实验,并分析评价指标的差异。

图2为隐私预算为0.5和1,组数为10、20、30、40、50、60、70、80、90、100的情况下,本方法与方法IS17和DP21的豪斯多夫距离对比图。可以看出,在大多数的隐私预算和分组数下,本方法得到的豪斯多夫距离小于方法IS17和DP21,且随着差分隐私预算的增加,豪斯多夫距离越来越低,符合差分隐私的定义,所以本方法得到的数据可用性较高。

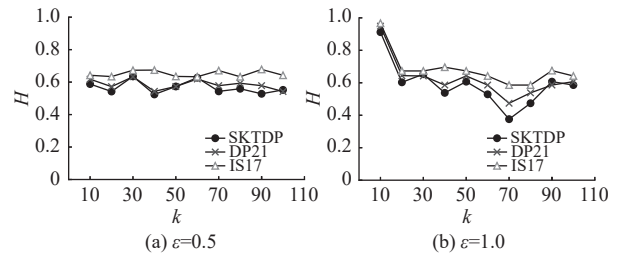


图2 不同隐私预算下的豪斯多夫距离

Fig.2 Hausdorff distances under different epsilon

图3分别为隐私预算0.5和1.0,组数为10、20、30、40、50、60、70、80、90、100的情况下,本方法与方法

IS17和DP21的计数查询误差对比图。在隐私预算为0.5时,分组数为70时计数查询误差最小,在隐私预算为1.0,分组数为100时计数查询误差最小,且均小于方法IS17和DP21。可以看出,在不同的隐私预算下,本方法都具有更低的计数查询误差,这说明本方法可以更好地拟合原始轨迹数据的稀疏性,更大程度上保存了轨迹数据信息。

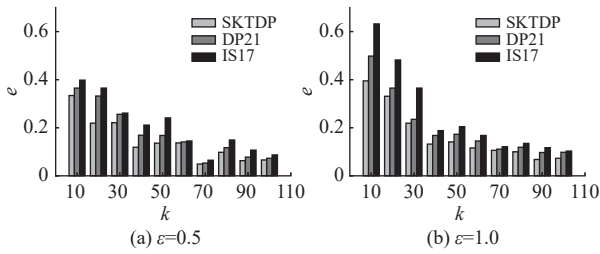


图 3 计数查询误差
Fig.3 Average relative error

图4分别为在同一隐私预算($\epsilon=1.0$)下,查询半径分别为0.4、0.6、0.8、1.0 km的F-measure得分。可以看出,在不同的分组数和查询半径下,本方法比方法IS17和DP21具有更高的数据效用。此外,从图中可以发现,随着分组数的增加,数据效用也呈现一个增加的趋势,这是因为分组数的增加会导致合并集群中的轨迹数量减少,从而提高数据效用。

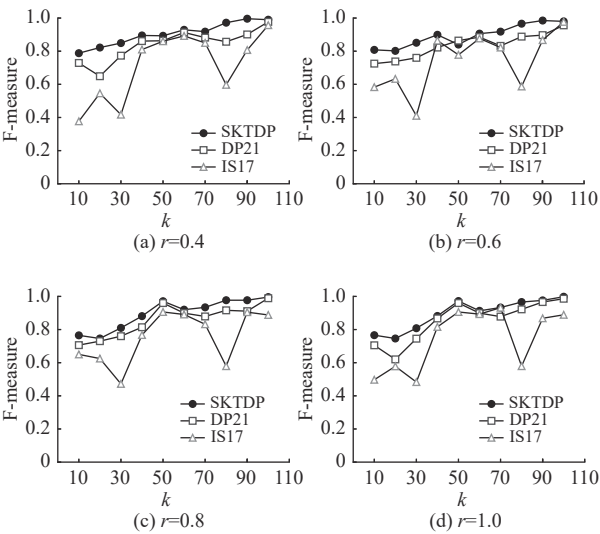


图 4 F-measure($\epsilon=1.0$)
Fig.4 F-measure($\epsilon=1.0$)

4 结论

本文针对时间序列轨迹数据发布存在的数据可用性低和聚类效果质量差的问题,提出一种基于

SKmeans||聚类和最优噪声机制的轨迹重构方法,在提高数据可用性的同时保证差分隐私。本文的方法包括基于Skmeans||聚类的轨迹泛化算法(TGSC)和基于有界噪声机制的轨迹发布算法(SNDP)。在TGSC算法中,在聚类中加入方向控制机制来控制聚类的收敛,保证聚类的质量,并针对质心的更新设计了新的指数机制的打分函数,然后将聚类得到的簇心映射到该簇内所有的点。对比常用的Kmeans聚类,本文提出的方法具有更好的聚类效果,也更适用于高维稀疏的轨迹集。实验表明,本文提出的方法在保证差分隐私的同时,具有更好的数据可用性。下一步将拓展差分隐私在其他新兴领域的应用,并尝试与机器学习和深度学习方法相结合,来提高数据可用性。

参考文献:

[1] LI J Y, GUO W Z, LI X Y, *et al.* Privacy-preserving real-time road conditions monitoring scheme based on intelligent traffic [J]. *Journal on Communications*, 2020, 41(7): 73-83.

[2] SWEENEY L. k-anonymity: a model for protecting privacy [J]. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 2002, 10(5): 557-570.

[3] MACHANAVAJJHALA A, KIFER D, GEHRKE J, *et al.* l-diversity: privacy beyond k-anonymity [J]. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2007, 1(1): 3-14.

[4] DWORK C. Differential privacy: a survey of results[C]//*International Conference on Theory and Applications of Models of Computation*. Heidelberg: Springer, 2008: 1-19.

[5] CHEN R, FUNG B C M, DESAI B C, *et al.* Differentially private transit data publication: a case study on the montreal transportation system[C]//*Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Beijing: ACM, 2012: 213-221.

[6] CHEN R, ACS G, CASTELLUCCIA C. Differentially private sequential data publication via variable-length n-grams[C]//*Proceedings of the 2012 ACM Conference on Computer and Communications Security*. Raleigh: ACM, 2012: 638-649.

[7] ZHAO X, DONG Y, PI D. Novel trajectory data publishing method under differential privacy [J]. *Expert Systems with Applications*, 2019, 138: 112791.

[8] HUA J, GAO Y, ZHONG S. Differentially private publication of general time-serial trajectory data[C]//*2015 IEEE Conference on Computer Communications (INFOCOM)*. Hong Kong: IEEE, 2015: 549-557.

- [9] LI M, ZHU L, ZHANG Z, *et al.* Achieving differential privacy of trajectory data publishing in participatory sensing [J]. *Information Sciences*, 2017, 400: 1-13.
- [10] GENG Q, VISWANATH P. The optimal noise-adding mechanism in differential privacy [J]. *IEEE Transactions on Information Theory*, 2015, 62(2): 925-951.
- [11] LI Y, YANG D, HU X. A differential privacy-based privacy-preserving data publishing algorithm for transit smart card data [J]. *Transportation Research Part C:Emerging Technologies*, 2020, 115: 102634.
- [12] GURSOY M E, LIU L, TRUEX S, *et al.* Differentially private and utility preserving publication of trajectory data [J]. *IEEE Transactions on Mobile Computing*, 2018, 18(10): 2315-2329.
- [13] NI T, QIAO M, CHEN Z, *et al.* Utility efficient differentially private K-means clustering based on cluster merging [J]. *Neurocomputing*, 2021, 424: 205-214.
- [14] LU Z, SHEN H. Differentially private K-means clustering with convergence guarantee [J]. *IEEE Transactions on Dependable and Secure Computing*, 2020, 18(4): 1541-1552.
- [15] LIU Q, YU J, HAN J, *et al.* Differentially private and utility-aware publication of trajectory data [J]. *Expert Systems with Applications*, 2021, 180: 115120.
- [16] HAMALAINEN J, KARKKAINEN T, ROSSI T. Scalable initialization methods for large-scale clustering[EB/OL]. arXiv preprint arXiv: 2007.11937 (2020-07-23)[2022-11-01].<https://doi.org/10.48550/arXiv.2007.11937>.
- [17] XU C, ZHU L, LIU Y, *et al.* DP-LTOD: differential privacy latent trajectory community discovering services over location-based social networks [J]. *IEEE Transactions on Services Computing*, 2018, 14(4): 1068-1083.
- [18] 陈思, 付安民, 苏铨, 等. 基于差分隐私的轨迹隐私保护方案[J]. *通信学报*, 2021, 42(9): 54-64.
CHEN S, FU A M, SU M, *et al.* Trajectory privacy protection scheme based on differential privacy [J]. *Journal on Communications*, 2021, 42(9): 54-64.
- [19] ZHAO X, PI D, CHEN J. Novel trajectory privacy-preserving method based on clustering using differential privacy [J]. *Expert Systems with Applications*, 2020, 149: 113241.
- [20] YUAN J, ZHENG Y, ZHANG C, *et al.* T-drive: driving directions based on taxi trajectories[C]// *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. California: ACM, 2010: 99-108.

(责任编辑: 赵少飞 英文审核: 费伦科)