基于自适应连续时间的群智感知轨迹隐私保护方案

蒋伟进[1,2,3],王海娟[1,3],周 为[3,4],陈艺琳[3,4],吴玉庭[3,4],韩裕清[3,4]
(1. 湖南工商大学前沿交叉学院,湖南长沙 410205;2. 武汉理工大学计算机与人工智能学院,湖北武汉 430070;
3. 湘江实验室,湖南长沙 410205;4. 湖南工商大学计算机学院,湖南长沙 410205)

摘 要: 针对轨迹差分隐私保护存在的预测精度差、隐私预算分配效用低的问题,本文提出自适应连续时间序列下的群智感知轨迹预测方案. 首先在任务分配阶段,为参与者分配轨迹路线;其次引入隐马尔可夫模型(Hidden Markov Model,HMM),对轨迹进行预测;然后使用预分配和自适应分配相结合的综合隐私预算分配方法,降低隐私预算;最后利用拉普拉斯机制,进行位置扰动. 实验结果表明,与相关工作相比,所提方法兼顾预测性和低预算性,对群智感知中参与者在轨迹隐私安全保护上具有良好的保护效果.

关键词: 轨迹预测;差分隐私;综合性隐私预算分配;自适应连续时间;群智感知
基金项目: 国家自然科学基金(No.61772196,No.72088101)
中图分类号: TP309 **文献标识码:** A **文章编号:** 0372-2112(2023)10-2894-08
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230359

Track Privacy Protection Scheme Based on Adaptive Continuous Time in Crowdsensing

JIANG Wei-jin[1,2,3], WANG Hai-juan[1,3], ZHOU Wei[3,4], CHEN Yi-lin[3,4], WU Yu-ting[3,4], HAN Yu-qing[3,4]
(1. *College of Frontier Intersection, Hunan University of Technology and Business, Changsha, Hunan* 410205, *China*;
2. *School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, Hubei* 430070, *China*;
3. *Xiangjiang Laboratory, Changsha, Hunan* 410205, *China*;
4. *School of Computer Science, Hunan University of Technology and Business, Changsha, Hunan* 410205, *China*)

Abstract: In response to the problems of poor prediction accuracy and low utility of privacy budget allocation in trajectory differential privacy protection, our paper proposes an adaptive trajectory prediction scheme for continuous time series in crowdsensing. Firstly, in the task assignment phase, trajectory routes are assigned to participants. Then, the HMM (Hidden Markov Model) is introduced to predict the trajectories. Next, a comprehensive privacy budget allocation method combining pre-allocation and adaptive allocation is used to reduce the privacy budget. Finally, the laplace mechanism is applied to perturb the locations. Experimental results show that compared with related work, the proposed method achieves a balance between prediction accuracy and low budget requirements, and provides good privacy protection for participants in trajectory privacy security in crowdsensing.

Key words: trajectory prediction; differential privacy; comprehensive privacy budget allocation; adaptive continuous time; crowdsensing
Foundation Item(s): National Natural Science Foundation of China (No.61772196, No.72088101)

1 引言

群智感知是一种实时收集、共享和分析物理世界中各种信息的感知方式[1,2],已广泛应用于环境监测[3]、城市规划[4]、智能交通[5]等领域. 但在感知活动过程中参与者通常需要移动位置来收集感知数据,这一过程往往存在较大的隐私泄露问题[6].

当参与者连续上传自身的位置数据时,攻击者可以通过推断重要位置点来获取他们的职业和人际关系等敏感信息,从而威胁到参与者的隐私安全. 为了保护连续实时状态下参与者的轨迹隐私并降低隐私预算,国内外学者们提出了多种算法框架和策略. 如 Huang 等人[7]考虑将每条轨迹视为高维空间中的一个向量,并

收稿日期:2023-04-21;修回日期:2023-08-25;责任编辑:覃怀银

设计了一种轨迹保护算法,在提交前对真实轨迹进行扰动,并使用差分隐私作为隐私模型,以估计给定隐私级别的噪声量. Wang等人[8]多关注于现有轨迹预测模型存在的轨迹缺乏多样性、准确性差和不稳定等局限性问题,提出了一种由生成器网络和能量网络组成的序列熵能量模型,实现轨迹一次性高效预测. Xie等人[9]多关注于现有轨迹预测模型存在的轨迹缺乏多样性、准确性差和不稳定等局限性问题,提出了一种由生成器网络和能量网络组成的序列熵能量模型,实现轨迹一次性高效预测.

目前针对轨迹隐私保护的方法如泛化、扰动和匿名化[10],但仍存在精度低开销大等问题. 且轨迹数据在空间和时间维度上具有相关性,相邻位置和时间戳的观察并非独立,而是连续且动态相关的,以往的研究常忽略这种相关性,导致轨迹预测不准确. 因此,本文综合考虑上述问题,提出了一种基于自适应连续时间的群智感知轨迹隐私保护方案,在尽可能降低隐私预算的前提下确保连续轨迹序列上参与者的隐私位置安全.

2　自适应连续时间下轨迹隐私保护方案

为了解决连续序列下的实时位置查询攻击,本文在对轨迹路径分配后引入隐马尔可夫模型(Hidden Markov Model, HMM)[11,12]进行轨迹预测,并在本地进行隐私预算方案的定制,最后对时序关联的轨迹数据添加噪声,从而实现位置数据发布时的隐私保护.

2.1　轨迹弹道位置模型建立

假设基于群智感知的智慧交通轨迹场景如图1所示:参与者起始位置位于P_1点,目的地为P_8点,从P_1点到P_8点路过的节点为$P_2,\cdots,P_7,P_5$和P_7为用户设置的敏感点或重要轨迹推断点. 敏感点由参与者根据自身隐私相关性进行确定,而重要轨迹推断点由基于HMM的轨迹预测推断得出. 再在图1的地理地图上建立物理轨迹弹道模型,如图2所示,将覆盖所有位置点的地理区域划分为大小相同的单元格,轨迹由相应的单元格序列表示.

图1　基于群智感知的智慧交通轨迹场景示意图

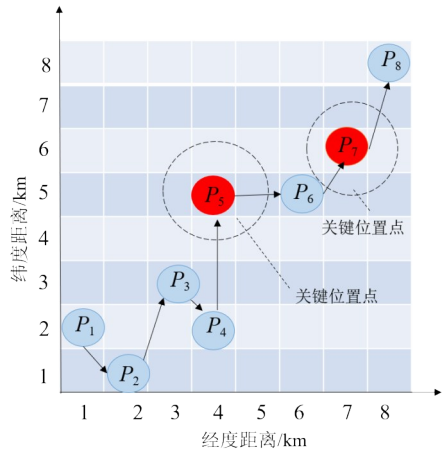


图2　位置数据弹道模型

2.2　轨迹分配规划

现有的分配机制没有考虑到在满足用户个性化和连续时间序列的情况下优化全局系统性,导致任务分配低效用,本节通过在任务分配阶段在轨迹弹道模型基础上刻画最优路径,高效分配节省任务预算和隐私预算,在选定参与者与路径后,再对该参与者后续的轨迹进行扰动.

假设目标为最小化总的移动距离,本节使用二分图的方法将参与者与任务进行匹配. 我们将任务集$K=\{k_1,k_2,\cdots,k_j\}$和参与者集$V=\{v_1,v_2,\cdots,v_i\}$分别作为二分图的左右顶点,每个任务对应一条轨迹Ω_j. 我们将结果矩阵$\boldsymbol{R}_{ji}$对角线之和即SDE值作为边的约束条件. 接着,计算该区域某任务与申请可执行该任务的用户的相对距离,并初始化带权二分图的顶标和边权值. 然后检查用户v_i与任务k_j是否存在可行边,判断能否为任务k_j找到一条增广路,若可以为该任务分配路径,若不行则更新二分图中的顶标值.

基于二分图的轨迹分配模型将轨迹路径提前到决策确定阶段,只需要确定边集约束条件SDE即可匹配到最短路径. SDE的计算具体为:首先分别计算任务位置矩阵$\boldsymbol{L}_{k_j}$和参与者的位置矩阵$\boldsymbol{L}_{v_i}$,位置矩阵由平台通过网格空间划分的区域得来. 接下来平台生成两个随机矩阵$\boldsymbol{D}_k$和$\boldsymbol{D}_v$,将随机矩阵与位置矩阵相乘得到任务和参与者的扰动矩阵$\boldsymbol{L}^*_{k_j}$和$\boldsymbol{L}^*_{v_i}$,然后将二者的扰动矩阵相乘则得到最终的结果矩阵$\boldsymbol{R}_{ji}=\boldsymbol{L}^*_{k_j}\times\boldsymbol{L}^*_{v_i}$. 本文利用结果矩阵$\boldsymbol{R}_{ji}$对角线之和SDE为衡量距离的标准,比较不同参与者与任务的距离. 参与者与任务的距离和SDE值成反比. 最后通过匈牙利算法寻找相等子图的完备匹配进行求解.

2.3 轨迹预测

本节我们将预测问题转换为隐马尔可夫决策过程. 通过计算单元格之间的转移概率和对应于特定位置点的概率确定参与者在参与活动过程中某时刻可能出现的位置.

定义1 隐马尔可夫模型. 是一种时序概率模型. 描述由隐藏的马尔可夫链随机生成不可预测的状态序列,再由各个状态生成观测序列的过程,模型参数μ由三元组$(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\pi})$决定. 其中,$\boldsymbol{A}$为状态转移矩阵,$a_{ij}$是$\boldsymbol{A}$中的元素;$\boldsymbol{B}$为发射概率矩阵,$b_i(k)$是$\boldsymbol{B}$中的元素;$\boldsymbol{\pi}$为初始状态概率向量,$\pi_i$是$\boldsymbol{\pi}$中的元素.

将轨迹序列视为HMM的可观察序列,单元格序列视为隐藏状态序列,移动物体的位置视为单元格产生的观测点.

定义2 轨迹. 是运动物体在运动过程中提交的一组位置点信息,表示为$\Omega=\{P_1,P_2,\cdots,P_T\}$,其中,$P_t=(x_t,y_t)$表示运动物体在时刻$t$的位置.

首先根据用户提交的历史轨迹信息,使用最大期望算法(Expectation Maximization,EM)对HMM进行训练,迭代计算步骤E、M得到概率模型参数μ的最大似然估计.

步骤E中,输入轨迹序列集,轨迹数据集为历史轨迹的集合. 定义隐藏状态序列为$\boldsymbol{s}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_T)$,在时间$t$的状态为$\boldsymbol{s}_t$,任意时刻下不可见,攻击者不能通过它来推断状态之间的转移概率等相关参数. 利用给定的观测序列$\boldsymbol{o}=(\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_T)$和HMM参数为$\mu$,计算位置点的概率$\xi_t(i,j)$,如式(1)所示:

$$\begin{aligned}\xi_t(i,j)&=Q(\boldsymbol{s}_t=\boldsymbol{s}_i,\boldsymbol{s}_{t+1}=\boldsymbol{s}_j|\boldsymbol{o},\mu)\\&=\frac{\alpha_t(i)a_{ij}\beta_{t+1}(j)b_j(\boldsymbol{o}_{t+1})}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)\beta_{t+1}(j)b_j(\boldsymbol{o}_{t+1})}\end{aligned}\tag{1}$$

其中,$\alpha_t(i)=b_i(\boldsymbol{o}_t)\sum_{j=1}^{N}\alpha_t(j)\times\alpha_{ij}$表示前向概率,$\beta_t(i)=\sum_{j=1}^{N}\beta_{t+1}(j)a_{ij}b_j(\boldsymbol{o}_{t+1})$表示后向概率.

根据式(1)得参与者在t时刻处于$\boldsymbol{s}_i$状态的概率为$\gamma_t(i)$:

$$\gamma_t(i)=Q(\boldsymbol{s}_t=\boldsymbol{s}_i|\boldsymbol{o},\mu)=\sum_{j=1}^{N}\xi_t(i,j)\tag{2}$$

在步骤M中,训练HMM的参数. 对获得的a_{ij}、$b_i(k)$和π_i不断迭代,直到收敛. 具体见式(3)~(5):

$$a_{ij}=Q(\boldsymbol{s}_t=\boldsymbol{s}_i|\boldsymbol{s}_{t+1}=\boldsymbol{s}_j)=\frac{\sum_{t=1}^{T}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}\tag{3}$$

$$\begin{aligned}b_i(k)&=Q(\boldsymbol{o}_t=\boldsymbol{o}_j|\boldsymbol{s}_t=\boldsymbol{s}_i)\\&=\frac{\sum_{t=1}^{T}\gamma_t(i)\times\delta(\boldsymbol{o}_t,\boldsymbol{o}_j)}{\sum_{t=1}^{T}\gamma_t(i)}\end{aligned}\tag{4}$$

$$\pi_i=Q(\boldsymbol{s}_1=\boldsymbol{s}_i|\boldsymbol{o},\mu)=\gamma_1(i)\tag{5}$$

其中,a_{ij}表示单元间位置转移的概率,$b_i(k)$表示每个单元对应的输出位置的概率,π_i表示每个单元中初始位置的概率.

但EM算法只能保证结果收敛到局部最优点,不能保证全局收敛最优. 因此,在初始化μ_0时,不使用随机化,而是根据历史轨迹信息计算$\boldsymbol{A}_0$和$\boldsymbol{B}_0$的值,将其设置为迭代的初始值,既增加了收敛到全局最优解的概率,又缩短了训练时间.

然后根据训练后的HMM模型和$\boldsymbol{o}=(\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_T)$,找到最可能产生这些位置点的隐藏状态序列.

定义$\delta_t(i)$表示在t时刻$\boldsymbol{s}_t$状态下所有隐藏序列的最大概率,表示如下:

$$\delta_t(i)=\max_{1\leqslant j\leqslant N}\{\delta_{t-1}(j)a_{ij}\}b_i(\boldsymbol{o}_i)\tag{6}$$

将隐藏状态序列在时刻t处隐藏状态具有的最大概率定义为$\psi_t(i)$表示如下:

$$\psi_t(i)=\operatorname*{argmax}_{1\leqslant j\leqslant N}\{\delta_{T-1}(j)a_{ij}\}\tag{7}$$

最后根据$\delta_t(i)$和$\psi_t(i)$,从初始时间递归到时间T,使用之前由$\psi_t(i)$记录的最可能状态节点进行回溯,直到找到最优隐藏状态序列,计算出这个隐藏状态对应的最有可能的结果,即预测结果$\boldsymbol{o}_T$.

算法1流程为:首先输入轨迹序列作为观测序列,进行模型训练,得到a_{ij}、$b_i(k)$和π_i,接着进行解码,寻找对应的单元格序列$\boldsymbol{s}^*=(\boldsymbol{s}_1^*,\boldsymbol{s}_2^*,\cdots,\boldsymbol{s}_{T-1}^*)$,预测当前位置所属的隐藏状态$\boldsymbol{s}_T^*$,并与发射概率矩阵$\boldsymbol{B}$相结合,最后找到隐藏状态对应的最可能的观测位置,得到预测结果$\boldsymbol{o}_T$. 具体详见算法1.

2.4 综合性隐私预算分配扰动算法

为了解决在滑动窗口在相似性计算中,存在的错误阈值会随着隐私预算的增加而误差增多导致采样不合理的现象和预算分配效率低的问题,提出了一种综合预分配和自适应分配的扰动算法,以改善分配效果,具体流程如下,具体算法见算法2.

2.4.1 隐私预分配

预分配的基本思想是利用预测结果来获取观测到的数据变化,并将数据采样点设置在特殊位置点所在的时间戳处,通过更早地分配隐私预算,改善分配效果.

算法1　轨迹预测算法

输入:轨迹数据集,待预测轨迹 $\Omega=\{P_1,P_2,\cdots,P_{T-1}\}$

输出:预测结果 $\boldsymbol{o}_T$

1 计算 $\boldsymbol{A}_0$ 和 $\boldsymbol{B}_0$ //设置初始值
2 $\mu_0\leftarrow$ 初始化 $\boldsymbol{A}_0,\boldsymbol{B}_0,\boldsymbol{\pi}_0$　　//进行HMM模型训练
3 n=0
4 WHILE μ_n 发散 DO
5　　计算 $\delta_t(i)$ 和 $\psi_t(i)$
6　　重新计算 $a_{ij},b_i(k),\pi_i$
7　　n=n+1
8 END WHILE
9 $\mu\leftarrow\mu_n$
10 初始化 $\delta_1(i)=\pi_i b_i(P_1),\psi_t(i)$　　//进行解码
11 FOR t=2 to $T-1$ DO:
12　　计算 $\delta_t(i)$ 和 $\psi_t(i)$
13 END FOR
14 $P^*\leftarrow\max(\delta_{T-1}),\boldsymbol{s}_T^*\leftarrow\operatorname{argmax}(\delta_{T-1})$
15 FOR $t=T-1$ to 1 DO
16　　$\boldsymbol{s}_t^*=\psi_{t+1}\boldsymbol{s}_{t+1}^*$
17 END FOR
18 $\boldsymbol{s}^*=(\boldsymbol{s}_1^*,\boldsymbol{s}_2^*,\cdots,\boldsymbol{s}_{T-1}^*)$
19 $\boldsymbol{s}_T^*\leftarrow\max(\boldsymbol{s}_{T-1}^*\boldsymbol{A})$
20 $\boldsymbol{o}_T\leftarrow\max(\boldsymbol{s}_T^*\boldsymbol{B})$
21 RETURN　$\boldsymbol{o}_T$

总隐私预算为 ε,预分配阶段隐私预算设置为 $\varepsilon_{r,1}=\frac{\varepsilon}{2}$. 在执行预分配操作之前,使用历史数据获得一阶差分均值 τ,来确定滑动窗口中的预分配数量 w. 一阶差分均值描述了每个数据点与这组数据的平均值之间的差距的平方的均值. 再将预分配数量设为 $c=\max\left(\frac{\varepsilon\times\tau}{2},1\right)$,预分配数组所需最大空间来缩短运算执行时间.

然后将历史数据输入算法2中,得到滑动隐私的预测值 X,再计算 X 的一阶差分序列,把 X 的一阶差分序列中 $\frac{w}{c}$ 最大值对应的时间戳设为 m. 在时间戳 m 处,预分配 $\varepsilon_{m,1}=\frac{\varepsilon}{2\times c}$,直接分配进行释放扰动. 下一次预分配决策的时间设置为 $d=i+\frac{w}{c}$. 若当前时间戳消耗预分配预算,则跳过下一个自适应分配.

2.4.2　自适应隐私预算分配

由于不同时间段的数据具有大小不同的相关性,同时当预测步数增加时,预测误差也随之增加,仅使用预分配不足以进行合理的预算分配,因此需要增加一个自适应分配过程来降低预算.

算法2　综合预算分配下的隐私扰动算法

输入:总隐私预算,实时轨迹数据集,预分配隐私数量 $c=\max\left(\frac{\varepsilon\times\tau}{2},1\right)$

输出: Z_t

1 $\chi=\frac{1}{1+d(P_t,\boldsymbol{o}_t)}$　　//计算位置点的可预测性
2 IF　$i=d$THEN　　//执行隐私预分配
3 计算 X 的一阶差分序列,获得当前时间戳 m
4 END
5 IF　$i=m$ THEN
6　　$\boldsymbol{o}_T\leftarrow\left(\mathrm{Lap}\left(\frac{2c}{\varepsilon}\right)\right)^N$
7　　执行预分配策略获取下一时刻时间戳
8　　$d=i+\frac{w}{c}$
9 ELSE
10　　计算扰动误差　　//执行自适应分配
11　　$\varepsilon_{m,2}\leftarrow\frac{\varepsilon_{r,2}}{2}-\sum_{m=i-w+1}^{i-1}\varepsilon_m$
12　　$\lambda_i=\frac{2}{\frac{\varepsilon}{2}-\varepsilon_{\text{used}}}$
13　　IF 扰动误差>λ_i THEN　　//相似性检验
14　　为当前时间戳的每个采样点分配隐私预算
15　　ELSE
16　　直接释放
17　　END
18　　$Z_t\leftarrow\left(\mathrm{Lap}(\lambda_i)\right)^N$
19 END

为了确定是否需要在每个时间戳进行采样和分配隐私预算,我们使用自适应采样方法进行采样. 即如果下一时间戳统计的预测值与上一次发布的统计值之间的误差小于扰动误差,表明数据波动大,近似释放误差大于微扰,无须进行采样或隐私预算分配过程;反之,如果误差较大,则为当前时间戳的每个采样点都分配隐私预算.

设置初始自适应阶段总隐私预算为 $\varepsilon_{r,2}=\frac{\varepsilon}{2}$. $i=\{i-w+1,\cdots,i-1\}$ 时,为其分配隐私预算 $\varepsilon_{\text{used}}=\sum_{m=i-w+1}^{i=1}\varepsilon_m$,剩余的隐私预算为 $\varepsilon_{m,2}=\frac{\varepsilon_{r,2}}{2}-\varepsilon_{\text{used}}$. 并使用指数衰减方法进行自适应分配预算,以灵活地调整预算分配以适应不同的查询需求和隐私风险,拉普拉斯噪声尺度设为 $\lambda_i=\frac{2}{\frac{\varepsilon}{2}-\varepsilon_{\text{used}}}$,通过比较扰动误差与拉普拉斯噪声尺度的值,来确定是否近似释放.

2.4.3　拉普拉斯位置扰动

我们将 ε-差分隐私应用于本文的地图数据中,将相

邻数据集转化为地理上接近的位置. 并使用拉普拉斯机制(Laplace)进行位置扰动,当两个相邻位置彼此接近时,所产生相同查询位置的概率相似,从而掩盖用户的真实位置信息.

定义3 拉普拉斯机制. 对于任意函数f,如果随机算法M的输出满足式(8),则算法M满足ε-差分隐私:

$$M(t)=f(t)+\mathrm{Lap}\left(\frac{\Delta f}{\varepsilon}\right) \tag{8}$$

其中,$\mathrm{Lap}\left(\frac{\Delta f}{\varepsilon}\right)$为添加的拉普拉斯噪声.

在获得当前位置的预测位置后,计算该数据点的可预测性,并调整相应的隐私预算. 将定位点的可预测性定义为式(9):

$$\chi=\frac{1}{1+d\left(P_t,\boldsymbol{o}_t\right)} \tag{9}$$

对平面Laplace机制进行加噪处理,所产生的任意近似位置,其概率函数满足式(10):

$$U_\varepsilon(P)(Z)=\frac{\varepsilon^2}{2\pi}\mathrm{e}^{-\varepsilon_d(P,Z)} \tag{10}$$

在中心点P假设一个敏感半径r,当实际位置为P时,满足ε地理不可区分性的干扰位置Z的极坐标半径的计算公式为式(11):

$$r=C_\varepsilon^{-1}(\rho) \tag{11}$$

再以点P为中心进行位置扰动. $P\left(x_0,y_0\right)$为实位置,给定参数r为敏感半径,根据等式计算扰动位置,即为原点加半径,在产生一组扰动位置中选择可预测性最低的位置作为替换位置. 对于生成的扰动位置,考虑扰动位置是否与真实位置在同一单元内,如果扰动位置与真实位置在同一单元内,则隐藏状态相同,并且扰动位置对后续位置可预测性的影响与真实位置对其的影响一致. 最后,再次递归调用本算法,实现下一时刻的位置扰动Z_t.

3 仿真实验与分析

本节对本文所提方法的隐私效率和预算效率进行实验,验证所提方法轨迹隐私保护效果. 实验环境为Intel(R) Xeon(R) 6240R CPU(3.5 GHz),RAM(8 GB),操作系统采用Windows10.

3.1 实验设置

实验选用微软亚洲研究院的GeoLife(地理轨迹数据)数据集. 选择北京范围内,时间间隔为10 min~1 h,查询点大于5个的轨迹作为样本轨迹. 对获得的4 567个轨迹进行采样,并使用58 084个位置进行实验. 为了计算方便,我们将用户的真实位置广义为50 m×50 m的区域,并用区域位置代替用户的真实位置. 同时,根据位置密度,将地图划分为2 240个单元格,作为轨迹预测的隐藏状态.

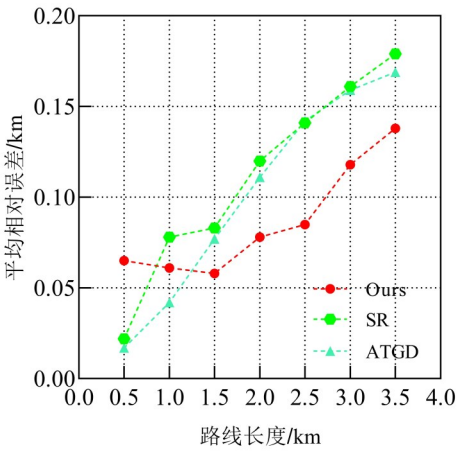
3.2 路径规划效果评估

为验证轨迹规划对于后续轨迹扰动效率的影响,我们设置一组实验,其中,扰动半径分别为$r=40$和$r=80$,并在数据集上,与空间众包中基于差异隐私的位置保护方案[13](Shift Route,SR)、自适应多重网格分解[14](Adaptive Three-level Grid Decomposition,ATGD)对比,比较不同路径长度下平均相对误差的变化情况,验证轨迹规划效果. 每组重复10次实验,取平均值作为最终结果,以表明通过合理的轨迹路径规划可以提高后续轨迹扰动效率.

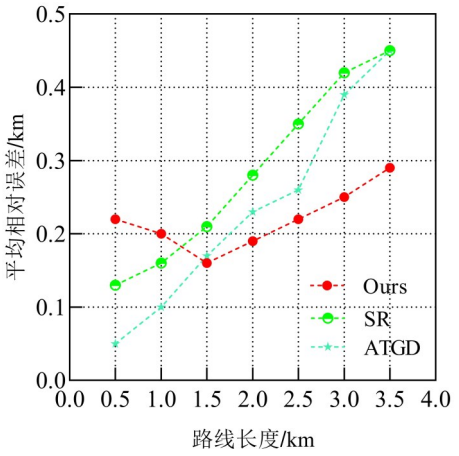
实验结果如图3所示,随着路径长度的增加,本文所提出的轨迹规划方案的平均相对误差明显低于SR和ATGD. 这表明通过合理的轨迹路径规划能显著提高后续轨迹扰动的效率,从而降低数据损失和误差.

3.3 隐私强度评估

我们在GeoLife数据集上,对比不同滑动下窗口和



(a) $r=40$



(b) $r=80$

图3 不同半径下路径长度对隐私效率的影响

隐私预算下的,基于位置的差分隐私(Privacy for Location-based,PL)[15]、基于移动轨迹的预测差分隐私机制(Predicted Mechanism,PM)[16]、基于移动人群感知中保护轨迹隐私机制(Trajectory Privacy Agent,TPA)[17]、图指数机制(Graph Exponential Mechanism,GEM)[18]、个性化轨迹隐私保护(Personalized Trajectory Privacy Protection,PTPP)[19]的累积分布函数(Cumulative Distribution Function,CDF),验证隐私保护水平.将隐私预算ε的取值范围设置为1~10,其中,隐私预算为总隐私预算.为了提高实验结果的可靠性,我们将实验划分为3组,每组滑动窗口数分别为4、6、8,每组实验重复10次,并取平均值作为最终结果.其他实验参数设置为:$\Delta\varepsilon=0.8\varepsilon,\beta_1=0.8,\beta_2=0.2$.

实验结果如表2所示.随着隐私预算的增多,CDF的值随之增大.在相同的隐私预算值下,我们的方案的CDF值高于其他方案,表明在达到相同的隐私保护效果时,本文所用算法所消耗隐私预算更小,隐私保护水平更高.

表2　不同滑动窗口数量下CDF比较

方案	$W=4$			$W=6$			$W=10$		
	$\varepsilon=2$	$\varepsilon=6$	$\varepsilon=10$	$\varepsilon=2$	$\varepsilon=6$	$\varepsilon=10$	$\varepsilon=2$	$\varepsilon=6$	$\varepsilon=10$
PL[15]	0.112	0.12	0.13	0.11	0.12	0.15	0.08	0.12	0.19
PM[16]	0.15	0.17	0.24	0.13	0.17	0.24	0.13	0.17	0.29
TPA[17]	0.25	0.27	0.36	0.2	0.27	0.36	0.2	0.27	0.36
GEM[18]	0.22	0.24	0.42	0.21	0.24	0.50	0.21	0.29	0.42
PTPP[19]	0.28	0.32	0.48	0.24	0.32	0.48	0.24	0.42	0.48
Ours	0.39	0.43	0.59	0.35	0.43	0.59	0.35	0.46	0.59

此外,我们还进行三组实验对比,比较不使用自适应预算分配方案(Without Adaptive Allocation,WAA)、不使用预分配方案(Without Pre-Allocation,WPA)和本文分配方案,在不同隐私预算下的平均绝对误差值(Mean Absolute Error,MAE).实验结果如表3所示,结果分析表明,本文方案可获得最为稳健的效果.WAA仅使用预分配方法,无法考虑最终发布数据,导致误差增加;WPA由于缺乏对时间窗口内整体数据波动的考虑,导致隐私预算分配不合理.本文将其组成一个有机体,可以有效提高数据的有用性,增加效率.并进一步比较不同隐私预算ε下WPA、WAA和Ours方案的均方根百分比误差(Root Mean Square Percentage,RMSP)的变化情况.RMSP表示扰动前后轨迹位置序列的均方根的可预测性,使用RMSP可以评估各方案抵预测攻击的能力.RMSP的定义如式(12):

$$\mathrm{RMSP}=\frac{1}{n}\sum_{i=1}^{n}\chi_i \tag{12}$$

实验结果如图4所示,随着隐私预算ε的增加,三种预算分配方案的RMSP在减小,但我们的RMSP最低,表明本文方案在隐私预算分配上具有良好的改善效果.

表3　不同方案下MAE对比　　单位:m

方案	$\varepsilon=0.8$	$\varepsilon=1.0$	$\varepsilon=1.2$	$\varepsilon=1.4$	$\varepsilon=1.8$
WPA	12.12	9.28	9.25	8.12	7.54
WAA	10.97	9.53	8.99	8.82	8.72
Ours	10.21	8.25	8.12	7.96	7.50

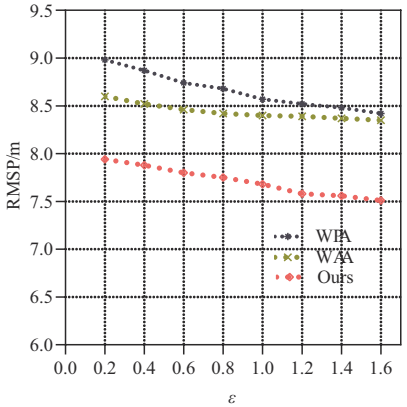


图4　隐私预算变化下隐私强度对比

3.4　轨迹模糊质量

当用户真实位置与扰动位置之间的误差越小,混淆质量就越高.在群智感知场景下,传感任务的数据质量会直接影响参与者服务质量和隐私效率.低质量的传感数据会增加分析误差,使中心节点难以进行有效的操控和数据挖掘与保护.采用均方根误差(Root Mean Square Error,RMSE)代表用户真实位置与扰动位置之间的误差,可以有效评估轨迹的模糊质量并降低数据的丢失率.RMSE定义如式(13):

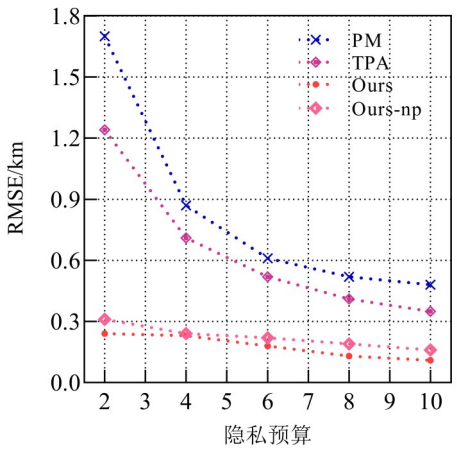
$$\mathrm{RMSE}=\frac{1}{n}\sum_{i=1}^{n}d\left(P_t,Z_t\right) \tag{13}$$

其中,d代表P_t和Z_t之间的曼哈顿距离.RMSE增大会导致隐私预算ε减小,隐私预算越大添加的扰动噪声越小,真实位置与扰动位置的平均误差也越小.

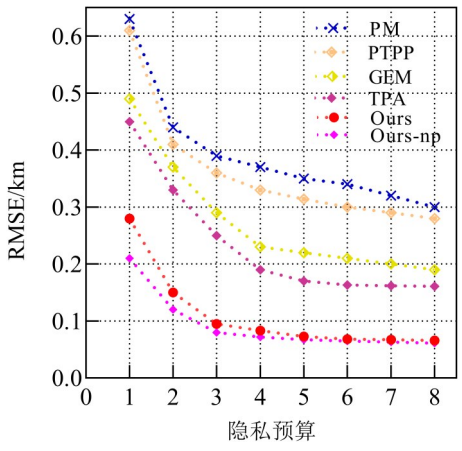
我们测试不同算法在$r=40$和$r=80$下,不同隐私预算变化的RMSE,实验结果如图5.

轨迹隐私的总预算取值为1到10之间,每个实验运行10次,取评估指标均值作为实验结果,其他参数设置为:$w=10,\Delta\varepsilon=0.8\varepsilon,\beta_1=0.8,\beta_2=0.2$.

结果分析如下:随着隐私预算ε的增加,各算法的RMSE减小,这意味着随着隐私预算越大,添加的扰动噪声就越小,真实位置和扰动位置的平均误差也越小.当隐私预算较小时,PM和TPA算法可以通过扰动位置提供一定的隐私保护,但当噪声增加,数据的可用性会降低.随着隐私预算的增加,隐私保护的需求会减少,数据可用性增加.在$r=80$的情况下和$r=40$的情况下



(a) $r=40$



(b) $r=80$

图5 不同隐私预算下轨迹模糊质量对比

图像基本相同. PM、PTPP和GEM算法的可用性高于PL和TPA算法,但本文算法仍然优于PL和TPA算法. 用$r=40$表示所提出的算法预测机制,显然,在加入预测机制后,该算法的可用性略有下降,因为预测机制虽然提高了数据的隐私保护程度,但以部分可用性为代价.

4 总结

针对现有轨迹隐私保护模型难以抵御强大攻击以及隐私预算较高的情况,本文提出了一种基于差分隐私的隐私轨迹分配方法,引入了基于自适应连续时间序列的实时群智感知轨迹预测方案,以实现参与者的轨迹隐私位置保护. 今后的研究可以将轨迹数据应用于更复杂性和可变性的场景,并进一步考虑不同用户对轨迹隐私要求的多样性.

参考文献

[1] DUTTA P, AOKI P M, KUMAR N, et al. Common Sense: Participatory urban sensing using a network of handheld air quality monitors[C]//Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems. New York: ACM, 2009: 349-350.

[2] MATHUR S, JIN T, KASTURIRANGAN N, et al. ParkNet: Drive-by sensing of road-side parking statistics [C]//Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services. New York: ACM, 2010: 123-136.

[3] THIAGARAJAN A, RAVINDRANATH L, LACURTS K, et al. VTrack: Accurate, energy-aware road traffic delay estimation using mobile phones[C]//Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems. New York: ACM, 2009: 85-98.

[4] 童咏昕, 袁野, 成雨蓉, 等. 时空众包数据管理技术研究综述[J]. 软件学报, 2017, 28(1): 35-58.
TONG Y X, YUAN Y, CHENG Y R, et al. Survey on spatiotemporal crowdsourced data management techniques[J]. Journal of Software, 2017, 28(1): 35-58. (in Chinese)

[5] YANG X, SHU L, CHEN J N, et al. A survey on smart agriculture: Development modes, technologies, and security and privacy challenges[J]. IEEE/CAA Journal of Automatica Sinica, 2020, 8(2): 273-302.

[6] WEI T K, LIU S C, DU X J, et al. Learning-based efficient sparse sensing and recovery for privacy-aware IoMT[J]. IEEE Internet of Things Journal, 2022, 9(12): 9948-9959.

[7] ZHANG Z M, XU X L, XIAO F. et al. LGAN-DP: A novel differential private publication mechanism of trajectory data[J]. Future Generation Computer Systems, 2023, 141: 692-703.

[8] WANG D F, LIU H B, WANG N Y, et al. SEEM: A sequence entropy energy-based model for pedestrian trajectory all-then-one prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 1070-1086.

[9] XIE Y, WANG Y, LI K, et al. Satisfaction-aware task assignment in spatial crowdsourcing[J]. Information Sciences, 2023, 622: 512-535.

[10] FENG J Y, WANG Y, WANG J L, et al. Blockchain-based data management and edge-assisted trusted cloaking area construction for location privacy protection in vehicular networks[J]. IEEE Internet of Things Journal, 2021, 8(4): 2087-2101.

[11] AL-GUMAEI A H, AZAM M, AMAYRI M, et al. ICA and IVA bounded multivariate generalized Gaussian mixture based hidden Markov models[J]. Engineering Appli-

cations of Artificial Intelligence, 2023, 123: 106345.
[12] ABBOOD A D, ATTEA B A, HASAN A A, et al. Community detection model for dynamic networks based on hidden Markov model and evolutionary algorithm[J]. Artificial Intelligence Review, 2023: 9665-9697.
[13] ZHANG P, HU C C, CHEN D, et al. ShiftRoute: Achieving location privacy for map services on smartphones[J]. IEEE Transactions on Vehicular Technology, 2018, 67 (5): 4527-4538.
[14] WEI J H, LIN Y P, YAO X, et al. Differential privacy-based location protection in spatial crowdsourcing[J]. IEEE Transactions on Services Computing, 2022, 15(1): 45-58.
[15] ANDRÉS M E, BORDENABE N E, CHATZIKOKOLAKIS K, et al. Geo-Indistinguishability: Differential privacy for location-based systems[C]//Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security - CCS' 13. New York: ACM Press, 2013: 901-914.
[16] CHATZIKOKOLAKIS K, PALAMIDESSI C, STRONATI M. A predictive differentially-private mechanism for mobility traces[M]//Privacy Enhancing Technologies. Cham: Springer International Publishing, 2014: 21-41.
[17] HUANG H Y, NIU X, CHEN C, et al. A differential private mechanism to protect trajectory privacy in mobile crowd-sensing[C]//2019 IEEE Wireless Communications and Networking Conference (WCNC). Piscataway: IEEE, 2019: 1-6.
[18] TAKAGI S, CAO Y, ASANO Y, et al. Geo-graph-indistinguishability: Protecting location privacy for LBS over road networks[M]//Data and Applications Security and Privacy XXXIII. Cham: Springer International Publishing, 2019: 143-163.
[19] LI J C, CHEN G Q. A personalized trajectory privacy protection method[J]. Computers & Security, 2021, 108: 102323.

作者简介

蒋伟进　男,1964年7月出生于湖南省益阳市. 现为湖南工商大学计算机学院二级教授、博士、博士生导师. 主要研究方向为联邦学习、群智感知、边缘计算、社会计算.
E-mail: jwjnudt@163.com

王海娟(通讯作者)　女,2000年8月出生于江西省九江市. 现为湖南工商大学前沿交叉学院硕士研究生. 主要研究方向为隐私安全、群智感知.
E-mail: 2502560411@qq.com

周　为　男,2000年5月出生于湖南省益阳市. 现为湖南工商大学计算机学院硕士研究生. 主要研究方向为隐私安全、群智感知.
E-mail: 1216330671@qq.com

陈艺琳　女,2000年9月出生于河南省许昌市. 现为湖南工商大学计算机学院硕士研究生. 主要研究方向为联邦学习、信息安全.
E-mail: 1986746095@qq.com

吴玉庭　女,1998年4月出生于湖南省益阳市. 现为湖南工商大学计算机学院硕士研究生. 主要研究方向为联邦学习、群智感知.
E-mail: 1321224262@qq.com

韩裕清　男,2000年11月出生于湖南省长沙市. 现为湖南工商大学计算机学院硕士研究生. 主要研究方向为联邦学习、边缘计算、群智感知.
E-mail: 897735614@qq.com