Journal of Computer Applications
计算机应用,2020,40(6):1692－1697
ISSN 1001-9081
CODEN JYIIDU
2020-06-10
http://www.joca.cn

文章编号:1001-9081(2020)06-1692-06
DOI:10.11772/j.issn.1001-9081.2019111929

可穿戴设备的数值型流数据差分隐私均值发布

涂子璇,刘树波*,熊星星,赵　晶,蔡朝晖
(武汉大学 计算机学院,武汉 430072)
(*通信作者电子邮箱liu.shubo@whu.edu.cn)

摘　要:可穿戴设备实时产生的用户健康数据(如心率、血糖等)对健康监测及疾病诊断具有重大意义,然而健康数据属于用户的隐私信息。针对可穿戴设备的数值型流数据均值发布,为防止用户的隐私信息泄漏,提出一种基于自适应采样的可穿戴设备差分隐私均值发布方法。首先,引入适应可穿戴设备流数据均值波动小这一特点的全局敏感度;然后,采用基于卡尔曼滤波调整误差的自适应采样的方式分配隐私预算,提高发布数据的可用性。在发布两种健康数据的实验中,所提方法在隐私预算为0.1时,即高隐私保护强度下,在心率和血糖数据集上的平均相对误差(MRE)分别为0.01和0.08,相较于差分隐私时序监测的滤波和自适应采样(FAST)算法分别降低了36%和33%。所提的均值发布方法能够提高可穿戴设备均值流数据发布的可用性。

关键词:可穿戴设备;差分隐私;流数据发布;均值统计;自适应采样
中图分类号:TP309.2　　**文献标志码:**A

Differential private average publishing of numerical stream data for wearable devices

TU Zixuan, LIU Shubo*, XIONG Xingxing, ZHAO Jing, CAI Zhaohui
(*School of Computer Science, Wuhan University, Wuhan Hubei* 430072, *China*)

Abstract: User health data such as heart rate and blood glucose generated by wearable devices in real time is of great significance for health monitoring and disease diagnosis. However, health data is private information of users. In order to publish the average value of numerical stream data for wearable devices and prevent the leakage of users' privacy information, a new differential private average publishing method of wearable devices based on adaptive sampling was proposed. Firstly, the global sensitivity was introduced which was adaptive to the characteristic of small fluctuation of stream data average for wearable devices. Then, the privacy budget was allocated by the adaptive sampling based on Kalman filter error adjustment, so as to improve the availability of the published data. In the experiments of two kinds of health data publishing, while the privacy budget is 0.1, which means that the level of privacy protection is high, the Mean Relative Errors (MRE) of the proposed method on the heart rate dataset and blood glucose dataset are only 0.01 and 0.08, which are 36% and 33% lower than those of Filtering and Adaptive Sampling for differential private Time-series monitoring (FAST) algorithm. The proposed method can improve the usability of wearable devices' stream data publishing.

Key words: wearable device; differential privacy; stream data publishing; average statistics; adaptive sampling

0　引言

随着越来越多的可穿戴设备投入市场,可穿戴设备以其便携、可穿戴、易移动和长续航等特点正逐步为社会大众所接受和使用,满足了各类人群的健康管理要求。日常使用的可穿戴设备主要分为两类:一类是针对大众健康监测类设备,如智能手环、智能手表等,作为运动和健康管理的辅助产品;另一类是针对某种慢性疾病的设备,如血糖仪、血压计等设备,作为慢性疾病的临床参考。可穿戴设备实时采集和定期向服务器发送用户的健康数据,这些数据被服务提供商或医学机构等第三方通过数据挖掘等技术为人们提供更健康的生活方式,但同时也可能泄漏用户隐私[1],例如用户的健康状况、地理位置等。

可穿戴设备中除了用户基本信息如姓名、年龄等记录型数据,还含有大量的需要实时采集和定期发布的数据,如心跳、血压、血糖等。这些用户敏感数据大多是数值型数据,而且数据有一定的波动范围。人群特征值属性和医疗临床数值范围的统计,这些都会用到均值发布。例如健康机构统计用户群的每日平均步数反映用户整体运动情况,医疗机构定期收集糖尿病患者的平均血糖情况研究治疗方案的效果。这些数值型均值数据流本身的实时性、连续性和波动性等特点给实时分析和安全发布带来了很多挑战。

差分隐私[2]是当前比较先进的隐私保护方法。它通过使用随机噪声来确保整体的数据保持其统计特性,不会因为个体数据的变化而变化,从而抵御差分攻击保护用户个人隐私。

收稿日期:2019-11-13;**修回日期:**2020-03-05;**录用日期:**2020-03-09。　**基金项目:**武汉市应用基础研究计划项目(2017060201010162);湖北省技术创新重大专项(2018AAA046);国家自然科学基金资助项目(61872431)。

作者简介:涂子璇(1996—),女,湖北黄冈人,硕士研究生,主要研究方向:信息安全、差分隐私;　刘树波(1970—),男,黑龙江泰来人,教授,博士生导师,博士,主要研究方向:物联网、嵌入式系统;　熊星星(1989—),男,江西南昌人,博士研究生,主要研究方向:信息安全、差分隐私;　赵晶(1996—),男,内蒙古巴彦淖尔人,硕士研究生,主要研究方向:信息安全、差分隐私;　蔡朝晖(1968—),女,湖北黄冈人,副教授,博士,主要研究方向:分布式信息处理。

关于差分隐私流数据发布的研究有文献[3-6]等,这些研究都是基于对数据流进行计数统计的差分隐私发布,对均值数据流的连续性不能进行很好地度量。目前很少有差分隐私均值发布的研究,文献[7]研究本地化差分隐私数值型静态数据的均值估计,不适用于流式均值数据发布。

本文主要研究可穿戴设备中数值型流数据的均值统计。均值统计和计数统计有两点不同:一是全局敏感度不同(见第2章),即对数据集中增加或减少一条记录在进行统计操作后造成的影响不同,从而加入的随机噪声也不相同,因此适用于计数统计的隐私保护方法不一定适用于均值发布;二是数据波动范围不同,可穿戴设备中常见的数值型数据本身范围有限制,即数值有上界和下界。相较于计数统计,在经过均值统计之后,均值流数据的波动范围更小。根据均值流数据的数据波动小这一特点,对数据进行采样,达到节省隐私预算的目的。

基于上述情况,本文提出了一种基于卡尔曼滤波(Kalman Filter, KF)误差可调的自适应采样流数据差分隐私均值发布方法。本文的主要工作有以下两点:

1) 针对可穿戴设备流数据随时间在一定范围内波动不大的特点,研究流数据的差分隐私均值发布。引入差分隐私流数据均值发布的全局敏感度,通过结合卡尔曼滤波的采样间隔自适应的方法以降低隐私预算开销和减少预测误差,提高发布数据的可用性。

2) 改进均值流算法自适应采样中的反馈误差,解决已有针对计数统计的自适应采样算法对数据波动过于敏感的问题,进一步降低发布数据的误差,提高可用性。

1　相关工作

关于可穿戴设备安全方面的研究目前主要集中在访问控制[8]、无线通信安全[9]、信息加密[10]等。在隐私数据发布方向,传统的医疗数据的隐私保护方有 k-匿名、l-多样性[11]等,然而这些方法缺乏对自身隐私保护能力的量化标准和对攻击者能力的界定,且需要一定的背景知识。基于数学模型的差分隐私解决了这些问题[2],文献[12]研究适合健康体域网的差分隐私保护,主要针对心电图数据,不具有普适性;文献[13]提出一种可穿戴设备多维数值型数据的均值统计方案,但该方法无法满足流数据的发布。

基于差分隐私的流数据发布,已有一些学者展开了研究。文献[3]研究二进制流的持续计数发布,该方法对发布每个时间点出现“1”的个数进行扰动然后发布。该方法仅对用户的单点进行隐私保护。文献[4]建立一个时间序列的状态空间模型降低扰动误差的影响,结合比例积分微分(Proportion Integral Differential,PID)过程控制进行自适应采样;但是保护的是用户整个时间序列的计数统计值,并不能直接运用到均值统计发布中。文献[5]提出的方法保护滑动窗口内所有事件隐私,比较当前点的计数总和与上一个发布值的相似性来决定是否发布;但该方法在窗口内的隐私预算分配采用指数递减的方法存在一定的问题。文献[6]研究时空数据流的计数发布,在文献[5]的基础上加入动态分组和自适应隐私预算分配,得到了进一步优化;但该方法不适用于均值统计发布,不能解决本文研究的问题。

目前的差分隐私算法主要围绕计数统计展开[3-6]。计数统计的全局敏感度是1,而均值统计的全局敏感度[14](详细见第2章)受被统计的数据集上下界影响。对于均值统计的流数据差分隐私发布,一些计数统计的方法可以借鉴。最直观的方法是在每个时间戳的每个均值估计上添加拉普拉斯噪声[15],这个方法使每个点分到的隐私预算很小,从而导致非常高的扰动误差,而且忽略了数据流的连续性。文献[16-17]在拉普拉斯噪声基础上加入卡尔曼滤波保证了数据的时序关联以及降低误差,但仍存在隐私预算分配问题。文献[18]为了节省隐私预算,在卡尔曼滤波的基础上进行固定间隔采样,缺点是无法动态适应数据集自身的波动情况,可能造成较大的预测误差。

据以上分析可知:已有针对可穿戴设备数据差分隐私发布研究中缺少关于流数据发布方面的研究;且普适性差分隐私流数据发布研究主要针对计数统计,不适用于均值统计。而可穿戴设备中存在大量数值型流数据,因此,本文提出了一种基于卡尔曼滤波误差可调的自适应采样流数据均值发布方法。

2　定义和理论基础

差分隐私模型最初用于保护统计数据库中个体的隐私信息,对数据进行一定的函数操作后,数据集中单个记录的插入或者删除操作对函数操作结果影响极小。首先引入邻近数据集的概念。对于传统的静态数据,如果有两个数据集 $\boldsymbol{D}$ 和 $\boldsymbol{D}'$,只存在一条记录不同,即 $|\boldsymbol{D}\Delta\boldsymbol{D}'|=1$,则称 $\boldsymbol{D}$ 和 $\boldsymbol{D}'$ 为邻近数据集。对于动态数据流,邻近数据集的定义与之类似。给定两个数据流 $\boldsymbol{D}$ 和 $\boldsymbol{D}'$,只存在一个用户不同,则 $\boldsymbol{D}$ 和 $\boldsymbol{D}'$ 为邻近数据流。

定义1　ε-差分隐私[19]。给定两个邻近数据流 $\boldsymbol{D}$、$\boldsymbol{D}'$和一个随机算法 A,若算法 A 对这两个邻近数据流的所有可能输出 $\boldsymbol{O}$ 均满足不等式(1),则称算法 A 满足基于用户的 ε-差分隐私。

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

参数 ε 被称为隐私预算,表示隐私保护程度。参数 ε 越小,则算法 A 在邻近数据流输出相同结果的概率越接近,隐私保护程度越高。

定义2　全局敏感度[2]。对任意一个函数 $f:\boldsymbol{D}\rightarrow\mathbf{R}^d$,函数的全局敏感度定义为数据集增加或删除一条记录对函数结果的最大影响,即:

$$\Delta f=\max_{D,D'}\left\|f(\boldsymbol{D})-f(\boldsymbol{D}')\right\| \tag{2}$$

全局敏感度的大小和函数本身有关。特别地,对于计数统计函数 $\Delta f=1$。而对于数值型数据,假设数据含有 n 条记录,变化范围是[MIN,MAX],则该数据集的均值函数的全局敏感度 $\Delta f=(MAX-MIN)/n$。

定义3　Laplace 机制[3]。给定数据集 $\boldsymbol{D}$,对任意一个函数 $f:\boldsymbol{D}\rightarrow\mathbf{R}^d$,其全局敏感度为 Δf,若随机算法 A 满足式(3),则算法提供 ε-差分隐私。

$$A(\boldsymbol{D})=f(\boldsymbol{D})+\left(Lap\left(\Delta f/\varepsilon\right)\right)^d \tag{3}$$

式中:$Lap\left(\Delta f/\varepsilon\right)$ 表示服从尺度为 $\Delta f/\varepsilon$ 的 Laplace 分布。Laplace 分布记为 $x\sim Lap(\mu,\lambda)$,在 Laplace 机制中默认 $\mu=0$。结合定义1可知,当 ε 越小时,隐私保护效果越好;但加入的扰动噪声越大,数据的可用性越低。

性质1　序列组合性[3]。对于一个数据集 $\boldsymbol{D}$,给定 m 个随机算法 $A_i(1\leqslant i\leqslant m)$,若每个 A_i 满足 ε_i-差分隐私,且 $\sum_{i=1}^{m}\varepsilon_i=\varepsilon$,那么序列 $A_i(\boldsymbol{D})$ 满足 ε-差分隐私。

3 可穿戴设备数值型流数据均值发布方法

3.1 问题描述

下面针对可穿戴设备的数据流特点和均值发布的特点，为其设计合适的差分隐私数据发布方法。可穿戴设备如智能手环或血糖仪等实时采集用户的健康医疗数据，这些实时数据如表1所示，大多是一些数值型数据，如某次运动测得的心率、饭后血糖含量，且具有一定的范围。受信任的服务器收集了大量用户的健康医疗数据进行汇聚，并发布给第三方如健康服务厂商或者临床医疗机构以进行市场分析或医疗方案优化。若这些数据被不可信的第三方获取，则用户的健康信息等隐私安全无法得到保证。因此受信任的服务器在发布用户的可穿戴设备流数据时要对数据进行差分隐私保护。

表1 可穿戴设备数值型实时数据

Tab. 1 Numerical real-time data of wearable devices

数据属性	数据来源	取值范围	单位
心率	手环	60~160	bpm
睡眠时长	手环	0~24	h
血糖	血糖仪	70~220(正常)	mg/dl
血压	血压计	60~140(正常)	mmHg

本文研究数值型流数据的均值发布方法。基于可穿戴设备的实时数据均值发布场景，服务器在每一个时间戳为所有用户建立数据库 $\boldsymbol{D}_k$。假设单变量的离散数值型流数据集合为 $\boldsymbol{X}=\{x_k\}$，x_k 表示在时刻 k 下对原始数据库 $\boldsymbol{D}_k$ 进行统计得到的原始统计数据 x，其中 $0 \leqslant k < T$，T 是时间流的长度。特别地，本文的应用场景下 $\boldsymbol{X}$ 是一个均值序列，例如可穿戴设备收集的某个地区用户每天的睡眠时长，或某医疗机构心脏病患者的心率、糖尿病患者的血糖平均值。设定原始均值流数据 $\boldsymbol{X}$ 经过差分隐私发布算法之后数据为 $\boldsymbol{Release}=\{r_k\}$。

目前针对数值型数据的差分隐私统计发布方法主要是将数值型数据根据范围进行分类，从而将数值型数据转化为分类型数据进行计数统计，没有普适的针对数值型流数据均值发布算法。针对可穿戴设备数值型数据的均值发布流数据随时间在一定范围内波动不大的特点，采用采样的方法可以有效降低隐私预算开销。参考文献[4]，本文针对原始均值流数据，提出了一种基于卡尔曼滤波误差可调的自适应采样的流数据均值发布方法。

3.2 方法描述

本节介绍可穿戴设备流数据均值发布方法，如图1所示。现有一个含有长度为 T 的单变量数值型数据流 $\boldsymbol{D}=\{\boldsymbol{D}_k\}$，$\boldsymbol{D}_k$ 表示当前时刻 k 所有 n 个用户从可穿戴设备中采集到的原始数值型数据集。对每个时间戳的原始数据进行均值统计得到对应的 x_k。假设该数据型单变量的变化范围是 $[MIN, MAX]$，则均值函数的全局敏感度 $\Delta f = |MAX - MIN|/n$。

下面描述流数据均值发布方法的步骤：第一步，对于每一个时间点 k，根据自适应采样部分确定该点是不是采样点。如果该点是采样点，则对采样点用Laplace机制进行扰动。第一步得到的数据可能与原始数据误差较大，且相邻时间节点的连续性无法度量。因此第二步采用卡尔曼滤波模型增加时间的相关性，对含噪数据进行预测和修正得到先验估计和后验估计。第三步是把第二步中先验估计和后验估计的绝对误差作为自适应采样部分的反馈参数计算下一个采样间隔，即通过数据随时间的波动情况自适应调整采样间隔。第四步，采样点的发布值为第二步滤波中的后验估计，非采样点的发布值为滤波得到的先验估计，即上一个采样点的发布值。在该方案中，采样点的隐私预算是均匀分配的，当采样点用完，即隐私预算消耗完时方案会停止采样。

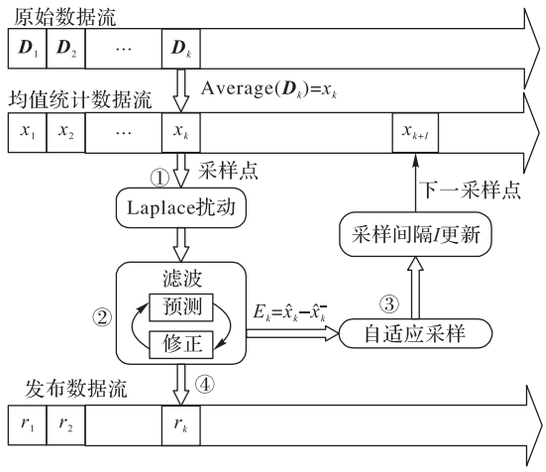


图1 流数据均值发布框架

Fig. 1 Framework of stream data average publishing

算法1描述了流数据均值发布方法的算法实现过程。已知现有的均值流序列为 $\boldsymbol{X}=\{x_k\}$，长度为 T，采样点个数为 M($M<T$)，则每个采样点分配到的隐私预算是 ε/M。均值统计的全局敏感度是 $\Delta f = |MAX - MIN|/n$，对采样点加入服从 $x\sim Lap\left(0, \Delta f \cdot M/\varepsilon\right)$ 分布的噪声。由于数值型均值流数据的波动范围小，文献[4]基于计数统计的反馈误差衡量方法调节的采样间隔对数据流的微小波动敏感，容易使隐私预算提前消耗完。本文方案对自适应采样中的反馈误差衡量方法进行了改进，减小了预测误差。后续结果表明本文的均值流数据发布(Average Stream Data Publishing, ASDP)算法的性能优于基于卡尔曼的差分隐私算法和原有针对计数统计的差分隐私时序监测的滤波和自适应采样(Filtering and Adaptive Sampling for differential private Time-series monitoring, FAST)算法。

算法1 ASDP算法。

输入 原始均值序列 $\boldsymbol{X}$，隐私预算 ε，最大采样点个数 M，全局敏感度 Δf；

输出 发布的序列 $\boldsymbol{Release}$。

```
for k=1:T do
通过卡尔曼预测算法得到先验估计 x̂_k^-
if k是采样点且采样点个数小于M
    z_k = x_k + Lap(0, Δf · M/ε)
    采样点个数加一
    通过卡尔曼修正算法得到后验估计 x̂_k
    r_k = x̂_k
    将 E_k = x̂_k - x̂_k^- 输入自适应采样算法调整采样间隔
else
    r_k = x̂_k^-
end
```

3.2.1 卡尔曼滤波

在流数据均值发布方法中，卡尔曼滤波能够提高数据之间的时序关联性和过滤部分噪声，同时为后续的自适应采样部分提供反馈误差。

卡尔曼滤波(KF)[20]是一种利用线性状态方程，通过系统输入输出观测数据，对系统状态进行最优估计的算法。该算法中有两个重要的常量参数 Q 和 R。从文献[4]中可知，Q 与原始均值序列的方差相关，R 与加入的噪声数据方差相关。因此，在本文的发布方法中：Q 取原始均值流数据的方差；R 为

Laplace噪声的方差,即$R = 2(\Delta f \cdot M/\varepsilon)^2$。基于上述取值,对均值流数据的后验估计效果可以达到最优。

针对均值序列的卡尔曼滤波差分隐私算法如算法2所示。首先,对原始均值流进行加噪。第3)~4)行是对数据的预测,即上一个时间点的发布数据作为当前时间点的先验估计。第5)~7)行对数据进行修正,得到当前时间点的后验估计,第8)行将后验估计作为发布值进行发布。在算法1中,若当前时间k为采样点,则对其进行预测和修正,后验估计作为发布值r_k;若该点为非采样点,则该点发布值r_k为该点的先验估计,即上一个采样点的发布值。用上一个采样点发布值替代非采样点的发布值造成的误差称为预测误差。合理地调整采样间隔可以减小预测误差。

算法2　卡尔曼滤波。

输入　原始均值序列$\boldsymbol{X}$,隐私预算ε;

输出　发布的序列$\boldsymbol{Release}$。

1)　对原始均值序列$\boldsymbol{X}$加入Laplace噪声,噪声服从$v \sim Lap(0, \Delta f/\varepsilon_k)$分布,得到扰动序列$\boldsymbol{Z}$

2)　for k=1:T do

3)　　$\hat{x}_k^- = \hat{x}_{k-1}$

4)　　$P_k^- = P_{k-1} + Q$

5)　　$K_k = P_k^-\left(P_k^- + R\right)^{-1}$

6)　　$\hat{x}_k = \hat{x}_k^- + K_k\left(z_k - \hat{x}_k^-\right)$

7)　　$P_k = (1 - K_k)P_k^-$

8)　　$r_k = \hat{x}_k$

9)　end

3.2.2　自适应采样

本节介绍可穿戴设备数值型流数据均值发布方法中采样自适应的目的以及具体针对均值发布的自适应采样改进方法。

对于流数据均值发布的隐私预算分配,若对每个时间点进行扰动,则每个点分到的隐私预算较小,噪声较大。随着时间的增长,添加的噪声也在不断累积,数据的可用性低,采样可以减少隐私预算分配。只在选定的采样点上加入扰动噪声进行发布,非采样点不进行扰动,即只在采样点分配隐私预算。均值流数据的数据波动不大,适合进行采样。常见的采样方法有固定间隔采样和自适应采样。固定间隔采样不能动态地根据数据流变化情况调整采样间隔,可能造成较大的预测误差。因此本方案对原始均值流进行自适应采样,使用PID控制器[21]捕捉数据流的动态变化,然后根据PID和当前采样间隔计算下一采样间隔,从而来确定下一个采样点。

PID控制器是最常见的反馈控制形式,用来度量采样效果随时间的变化。PID的输入为反馈误差。在过滤机制中,反馈误差用来衡量先验估计和后验估计的差距以表示数据的变化情况。当后验估计和先验估计相差较大时,表明数据变化较快,则应提高采样频率。在文献[4]中,PID的反馈误差为相对误差。但在均值流数据中,数据本身波动范围较小,且随时间波动变化相较于计数统计后得到的流数据也较小,若采用相对误差作为反馈误差,则算法对数据的波动过于敏感,采样频率过高从而过度采样,提前消耗完隐私预算。因此本方案采用式(4)所示的绝对误差作为反馈误差。

定义4　用$k_n(0 \leqslant k_n \leqslant T)$表示第$n$个采样点$(0 \leqslant n \leqslant M)$对应的时间戳,$\hat{x}_{k_n}$和$\hat{x}_{k_n}^-$分别为滤波机制得到的后验估计和先验估计,则本方案定义的反馈误差为:

$$E_{k_n} = \left|\hat{x}_{k_n} - \hat{x}_{k_n}^-\right| \tag{4}$$

自适应采样算法的过程如算法3所示。若当前时间戳为采样点,则从卡尔曼滤波算法中计算反馈误差,再根据反馈误差计算PID误差。再结合上一步采样间隔I自适应调整下一步采样间隔I'。

算法3　自适应采样。

输入　当前时间戳k,下一个采样点ns;

输出　新的采样间隔I',下一个采样点ns。

1)　if　k==ns

2)　　k是采样点

3)　　从算法2中根据式(4)获取E_{kn}

4)　　根据E_{kn}计算PID更新Δ

5)　　依据式(5)和Δ更新采样间隔I'

6)　　ns=ns+I',$I = I'$

7)　else

8)　　k不是采样点

9)　end

式(5)是更新采样间隔的方法:

$$I' = \max\left\{1, I + \theta(1 - \exp((\Delta - \xi)/\xi))\right\} \tag{5}$$

其中:θ表示采样间隔变化程度,ξ表示PID容忍度值,这两个值是经验参数。ξ值控制采样过程中的采样间隔,衡量PID误差的容忍程度。由式(5)知,当PID误差$\Delta > \xi$,表明相邻数据的波动超出预期,则下一次的采样间隔I'将小于当前采样间隔I;当$\Delta < \xi$时,表明数据波动较小,则下一次采样间隔将增大。本文将在实验部分研究这两个参数的变化对实验结果的影响。

3.3　可用性和隐私性分析

本文用平均相对误差(Mean Relative Error, MRE)衡量差分隐私发布的均值数据流的可用性,将多个用户数值型数据流在经过均值统计之后的原始均值数据流和经过差分隐私发布方法得到的均值数据流计算平均相对误差。所得误差越小,则隐私发布数据与原始数据的统计特性越接近,可用性越高。本文所提的方法与文献[4]方法相似,可用性理论分析和证明见文献[22](文献[4]的详细版本)。

接下来从理论上分析算法的隐私性。

性质2　ASDP算法满足ε-差分隐私。

证明　ASDP算法选取M个采样点,为每一个采样点分配$\varepsilon_k = \varepsilon/M$的隐私预算。在采样间隔参数$\theta$设置得偏小或者适中的情况下,采样点个数恰好为$M$,则根据性质1序列组合性可知算法整体消耗的隐私预算为ε,满足ε-差分隐私;若采样间隔θ设置不合理导致采样点个数G小于M(即下一个采样点时间戳超过了数据流长度T而采样点未用完),则消耗的隐私预算为$\varepsilon_k \cdot G = \varepsilon \cdot G/M < \varepsilon$,满足$\varepsilon$-差分隐私。　　证毕。

4　实验与结果分析

4.1　实验设置

实验采用模拟数据集和真实数据集对数值型流数据均值发布方法的可行性进行评估。模拟数据集模拟可穿戴设备手环常见的心率数据,选取5 760 000条记录,心率的数值范围是[60,160],模拟的心率数据服从正态分布。真实数据集为加州大学欧文分校的糖尿病数据集(http://archive. ics. uci. edu/ml/data-set//Diabetes),该数据集记录了自1989年到1991年糖尿病患者在早中晚餐及睡前的血糖含量数据,总共488个时间点,选取210 000条记录,血糖值的波动范围是[30,400]。将这两种数据集进行均值统计后的数据作为原始数据。

心率数据集和血糖数据集数据分布情况如表2所示。其中心率和血糖的单位同表1中描述单位。从标准差和最大

值、最小值可以看出，血糖数据集的数据波动比心率数据大，这个性质也与健康数据数值范围本身存在一定的背景知识限制有关。

表 2 两种数据集的数据分布特性

Tab. 2 Data distribution features of two datasets

数据集	心率数据/ bpm	血糖数据/（mg·dl^{-1}）
平均值	102	161.03
标准差	0.909	9.608
最小值	60	30
最大值	160	400

实验采用平均相对误差（MRE）来衡量该数据发布方案的可用性。给定原始数据流 $\boldsymbol{X}=\{x_k\}$ 和经过发布方法得到的数据流 $\boldsymbol{Release}=\{r_k\}$，在该场景下，其 MRE 为：

$$MRE=\frac{1}{T}\sum_{k=1}^{T}(|r_k-x_k|)/x_k \tag{6}$$

4.2 结果分析

在上述两个数据集上，改变采样误差容忍度 ξ 值，采样间隔控制 θ 值和采样点比例 M/T 以寻找合适的采样参数，然后选取基础拉普拉斯（LaPlAce，LPA）算法、KF 算法和针对计数统计的 FAST 算法[4]作对比来衡量本文设计的 ASDP 算法可用性。

4.2.1 不同 PID 容忍度 ξ 值下的误差分析

本组实验研究不同数据集对 ξ 值的最优值选择问题，将心率数据集和血糖数据集截取成相同的时间流长度 T=300，并设置 ε=1、M=90，其他采样参数均相同。ξ 的取值范围是从 0.2% 到 10% 以 0.002 为间隔，共 50 个取值。对于不同数据波动分布的数据集，ξ 取值的选取影响方案的可用性。

图 2 展示了 ξ 取值对两个不同数据集的影响。纵坐标是当前的 *MRE*/*MIN*，其含义是当前的 *MRE* 数值比上实验中最小的 *MRE* 值。原因是在 ξ 取值相同时，血糖数据集和心跳数据集的 *MRE* 相差较大，若直接比较 *MRE* 则呈现效果不佳。如图 2 所示，心率数据集的数据波动小，在 ξ=0.4% 时误差最小，在 0.4% 到 2% 误差呈指数上升，自 2% 到 10% 平稳上升；而血糖数据集的数据波动相对大一些，在 ξ=3% 时误差最小，在 3% 到 10% 接近直线上升。即对于数据波动小的数据集，ξ 的取值相较数据集波动大的数据集要小，因为该方案中引入 PID 是用于衡量相邻数据之间的相似性以确定采样间隔，对应 ξ 的选取应该与数据集的波动大小相适应。

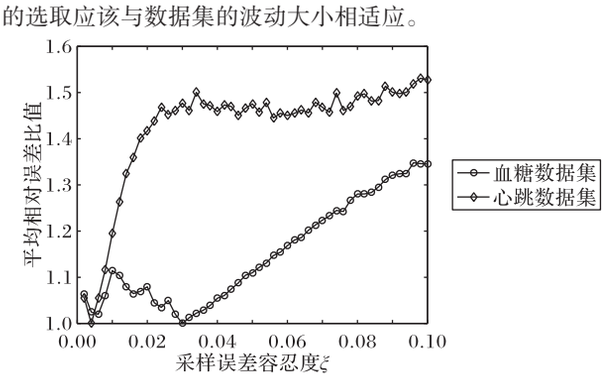


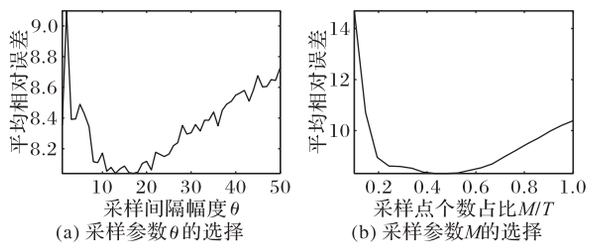
图 2 两种数据集不同 ξ 值下的平均相对误差比值比较

Fig. 2 Comparison of MRE ratio between two datasets under different ξ

4.2.2 采样参数 θ 值和采样点比例 M/T 的误差分析

本组实验研究采样参数 θ 值和采样点比例 M/T 对可用性的影响，在血糖数据集上进行实验，设置 ξ=3%，ε=1，T=300。θ 表示自适应采样间隔变化的幅度，M 表示采样点个数。

从图 3(a) 中可知，θ 在 10~20 时，该方案的效果比较好；当 θ=1 时，采样间隔过小导致采样频繁，过早用完采样点，在实际应用中不可取；当 θ>20 时，即随着采样间隔逐渐增大，相邻点的预测误差增大从而导致误差增大。



(a) 采样参数 θ 的选择 (b) 采样参数 M 的选择

图 3 自适应采样的参数选择

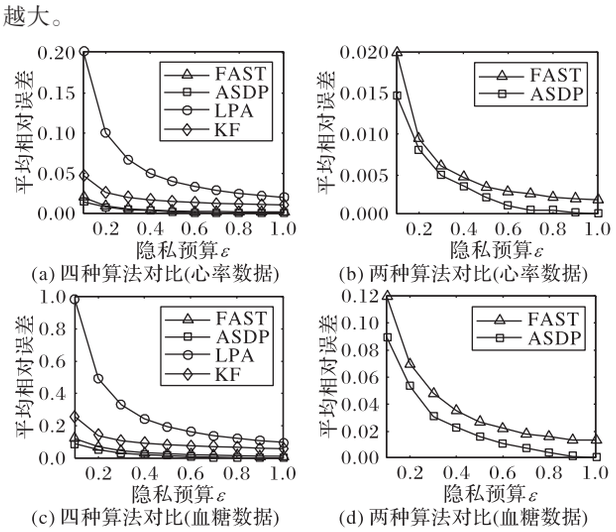
Fig. 3 Parameter selection of adaptive sampling

采样点的变化规律与采样间隔变化规律类似。每个采样点分配的隐私预算为 ε/M。当采样点比较少时，每个采样点加入的噪声偏小但相邻点间的预测误差较大；当采样点较多时，每个采样点加入的噪声偏大而相邻点间的预测误差小。如图 3(b) 所示，当 M/T 的比例在 0.2~0.6 时，该方案能够达到很好的效果；比例超过 0.6 后误差逐渐增大。

4.2.3 隐私预算对可用性的影响

实验设置隐私预算 ε 的范围是 0.1~1.0，间隔为 0.1。将 ASDP 算法与 LPA 算法、KF 算法和 FAST 算法分别在两种数据集中进行实验。对于 LPA 算法和 KF 算法，每个采样点分配到的隐私预算为 ε/T；而 FAST 算法和 ASDP 算法，每个点的隐私预算为 ε/M。对于心跳数据集取 PID 容忍度 ξ=0.4%，血糖数据集 ξ=3%。

如图 4 所示，在 PID 容忍度 ξ 取到适合于该数据集波动特性的情况下，ASDP 算法在心率和血糖两种数据集下取得了类似的效果。随着 ε 增大，四种发布算法的误差在所有的数据集中都在下降，这是因为当隐私预算越小所造成的噪声越大。



(a) 四种算法对比(心率数据) (b) 两种算法对比(心率数据)

(c) 四种算法对比(血糖数据) (d) 两种算法对比(血糖数据)

图 4 不同隐私预算下两种数据集的误差分析

Fig. 4 Error analysis of two datasets under different privacy budgets

将四种算法进行比较，由图 4(a) 和 (c) 可以看出，随着 ε 增大，ASDP 算法明显优于 LPA 算法和 KF 算法。在 ε=0.1 时该算法的效果明显优于 LPA 和 KF；在 ε>0.5 之后，该算法的效果与 KF 算法的效果近似，但均优于 KF 算法。这是因为该算法运用了采样机制，因此每一个采样点能分配到更多的隐

私预算;用PID机制衡量采样点之间的相似性保证了采样的可靠性。将FAST算法和ASDP算法单独进行对比,结果如图4(b)和图4(d)所示。由图4(b)和图4(d)可以看出,随着隐私预算的增大,ASDP算法的误差均明显小于FAST算法。这是由于ASDP算法更适合可穿戴设备流数据的波动特点,反馈误差的修改使采样点和采样间隔的选择较FAST更为合理,从而减小了发布数据的误差,提高了流数据均值发布数据的可用性。

5　结语

本文针对可穿戴设备数值型流数据的均值统计发布,提出了基于自适应采样的流数据差分隐私均值发布方法。该方法引入均值统计的全局敏感度;根据数值型数据进行均值统计之后数据波动不大的特点,采用基于卡尔曼滤波调整误差的自适应采样,既兼顾了流数据的实时性和连续性,又根据流数据的波动特点动态采样,节省隐私预算;进一步改进自适应采样的反馈误差,解决了对数据波动过于敏感而过度采样的问题。实验结果表明,本文的均值发布方法有效地减少了采样带来的预测误差,提高流数据均值发布数据的可用性。

本文提出的数值型流数据差分隐私均值发布方法仅考虑单维度流数据发布,未考虑多维数值型流数据发布。多维数值型流数据的差分隐私发布是接下来值得研究的方向。在每个采样点的隐私预算分配方面,仅采用简单的均匀分配,动态调整每个采样点的隐私分配策略是下一步需要考虑的方向。

参考文献 (References)

[1] 刘强,李桐,于洋,等. 面向可穿戴设备的数据安全隐私保护技术综述[J]. 计算机研究与发展,2018,55(1):14-29. (LIU Q, LI T, YU Y, et al. Data security and privacy preserving techniques for wearable devices: a survey [J]. Journal of Computer Research and Development, 2018, 55(1): 14-29.)

[2] DWORK C. Differential privacy [C]// Proceedings of the 2016 International Conference on Automata, Languages, and Programming, LNCS 4052. Berlin: Springer, 2006: 1-12.

[3] DWORK C, NAOR M, PITASSI T, et al. Differential privacy under continual observation [C]// Proceedings of the 42nd ACM Symposium on Theory of Computing. New York: ACM, 2010: 715-724.

[4] FAN L, XIONG L. An adaptive approach to real-time aggregate monitoring with differential privacy [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(9): 2094-2106.

[5] KELLARIS G, PAPADOPOULOS S, XIAO X, et al. Differentially private event sequences over infinite streams [J]. Proceedings of the VLDB Endowment, 2014, 7(12): 1155-1166.

[6] WANG Q, ZHANG Y, LU X, et al. Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy [J]. IEEE Transactions on Dependable and Secure Computing, 2018, 15(4): 591-606.

[7] NGUYÊN T T, XIAO X, YANG Y, et al. Collecting and analyzing data from smart device users with local differential privacy [EB/OL]. [2019-01-05]. https://arxiv.org/pdf/1606.05053.pdf.

[8] DIEZ F P, TOUCEDA D S, CÁMARA J M S, et al. Lightweight access control system for wearable devices [J]. IT Professional, 2019, 21(1): 50-58.

[9] GAI K, CHOO K K R, QIU M, et al. Privacy-preserving content-oriented wireless communication in Internet-of-things [J]. IEEE Internet of Things Journal, 2018, 5(4): 3059-3067.

[10] COELHO K, DAMIÃO D, NOUBIR G, et al. Cryptographic algorithms in wearable communications: an empirical analysis [J]. IEEE Communications Letters, 2019, 23(11): 1931-1934.

[11] YE H, CHEN E S. Attribute utility motivated k-anonymization of datasets to support the heterogeneous needs of biomedical researchers [C]// Proceedings of the 2011 American Medical Informatics Association Annual Symposium. Bethesda: American Medical Informatics Association, 2011: 1573-1582.

[12] LIN C, SONG Z, SONG H, et al. Differential privacy preserving in big data analytics for connected health [J]. Journal of Medical Systems, 2016, 40(4): Article No. 97.

[13] 马方方,刘树波,熊星星,等. 可穿戴设备数值型敏感数据本地差分隐私保护[J]. 计算机应用,2019, 39(7):1985-1990. (MA F F, LIU S B, XIONG X X, et al. Privacy protection based on local differential privacy for numerical sensitive data of wearable devices [J]. Journal of Computer Applications, 2019, 39(7): 1985-1990.)

[14] KARWA V, VADHAN S. Finite sample differentially private confidence intervals [EB/OL]. [2019-08-20]. https://salil.seas.harvard.edu/files/salil/files/karwa.pdf.

[15] MCSHERRY F D. Privacy integrated queries: an extensible platform for privacy-preserving data analysis [C]// Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2009: 19-30.

[16] ANDRE H, LE NY J. A differentially private ensemble Kalman filter for road traffic estimation [C]// Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2017: 6409-6413.

[17] WANG J, LUO J, LIU X, et al. Improved Kalman filter based differentially private streaming data release in cognitive computing [J]. Future Generation Computer Systems, 2019, 98: 541-549.

[18] 于东,康海燕. 面向时序数据发布的隐私保护方法研究[J]. 通信学报,2015,36(Z1):243-249. (YU D, KANG H Y. Privacy protection method on time-series data publication [J]. Journal on Communications, 2015, 36(Z1): 243-249.)

[19] LI H, XIONG L, JIANG X, et al. Differentially private histogram publication for dynamic datasets: an adaptive sampling approach [C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York: ACM, 2015: 1001-1010.

[20] KALMAN R E. A new approach to linear filtering and prediction problems [J]. Journal of Basic Engineering, 1960, 82(1): 35-45.

[21] KING M. Process Control: a Practical Approach [M]. New York: Wiley Publishing, 2011:36-45.

[22] FAN L, XIONG L. Adaptively sharing real-time aggregate with differential privacy [EB/OL]. [2019-01-05]. https://arxiv.org/pdf/1202.3461.pdf.

This work is partially supported by the Applied Basic Research Program of Wuhan (2017060201010162), the Major Technical Innovation Project of Hubei Province (2018AAA046), the National Natural Science Foundation of China (61872431).

TU Zixuan, born in 1996, M. S. candidate. Her research interests include information security, differential privacy.

LIU Shubo, born in 1970, Ph. D., professor. His research interests include Internet of things, embedded system.

XIONG Xingxing, born in 1989, Ph. D. candidate. His research interests include information security, differential privacy.

ZHAO Jing, born in 1996, M. S. candidate. His research interests include information security, differential privacy.

CAI Zhaohui, born in 1968, Ph. D., associate professor. Her research interests include distributed information processing.