

函数回归的差分隐私保护算法

钟可欣¹, 杨庚^{1,2}

(1. 南京邮电大学 计算机学院, 江苏 南京 210046;
2. 江苏省大数据安全与智能处理重点实验室, 江苏 南京 210023)

摘要:函数型数据回归是一种特殊的回归分析,其响应或协变量包含函数型数据,即样本元素为连续函数的数据。函数型数据在医疗保健、社交媒体、传感网络等诸多领域都有重要应用,通常包含一些敏感信息,在回归分析的过程中,不加保护会引起隐私的泄露。针对函数型数据回归分析中缺少隐私保护的问题,提出了一种基于拉普拉斯机制的函数回归的差分隐私保护算法。首先,对响应数据进行降维,将响应函数建模为相互正交的B样条基的张量积,建立函数回归的数学模型;其次,对回归模型的未知参数取值使用惩罚最小二乘法估计,并通过正交基函数的数量控制粗糙度;最后,对估计参数加入服从拉普拉斯分布的噪声扰动,得到最终的回归结果。理论分析和实验表明,函数回归的差分隐私保护算法满足拉普拉斯机制的差分隐私保护,并且随着隐私预算的减小,算法效率越高,在保证数据安全性的同时达到了良好的可用性。

关键词:函数型数据分析;差分隐私;函数回归;数据隐私保护;隐私预算分配

中图分类号:TP309

文献标识码:A

文章编号:1673-629X(2023)02-0132-06

doi:10.3969/j.issn.1673-629X.2023.02.020

Differential Privacy Preservation Algorithm in Functional Regression

ZHONG Ke-xin¹, YANG Geng^{1,2}

(1. School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210046, China;
2. Jiangsu Province Key Laboratory of Big Data Security and Intelligent Processing, Nanjing 210023, China)

Abstract: Functional data regression is a special kind of regression analysis whose responses or covariates contain functional data, that is, data whose sample elements are continuous functions. Functional data has important applications in many fields such as health care, social media, and sensor networks. It usually contains some sensitive information. In the process of regression analysis, if it is not protected, it will cause privacy leakage. Aiming at the lack of privacy protection in functional data regression analysis, a differential privacy protection algorithm based on Laplacian mechanism for functional regression is proposed. Firstly, reduce the dimension of the response data, model the response function as a tensor product of mutually orthogonal B-spline bases, and establish a mathematical model of function regression. Secondly, use the penalized least squares method for the unknown parameters of the regression model, and control the roughness through the number of orthogonal basis functions. Finally, add noise disturbance obeying Laplace distribution to the estimated parameters to obtain the final regression result. Theoretical analysis and experiments show that the differential privacy protection algorithm of functional regression satisfies the differential privacy protection of the Laplacian mechanism, and with the reduction of the privacy budget, the efficiency of the algorithm is higher, and it achieves excellent usability while ensuring data security.

Key words: functional data analysis; differential privacy; functional regression; data privacy preservation; privacy budget allocation

0 引言

函数型数据分析(Functional Data Analysis)是统计学中涉及对曲线、曲面或任何其他连续变化的信息分析的一个分支。对于函数型数据^[1],理想的观测单位是在某个连续域上定义的函数,观测数据由从某个总体中抽取的函数样本组成,每个函数在离散网格上

采样。随着信息科学技术的发展,函数数据在诸多领域中发挥了重要作用,例如医疗行业中的扫描成像数据、社交媒体的个人行为轨迹等。

然而,函数型数据的广泛应用也存在一些急需解决的问题。隐私泄漏的危机伴随数据分析与发布等应用的出现而加深,对隐私数据的保护问题与防止敏感

收稿日期:2022-03-27

修回日期:2022-07-28

基金项目:国家自然科学基金项目(61872197,61972209)

作者简介:钟可欣(1998-),女,硕士研究生,研究方向为差分隐私保护、数据回归;通讯作者:杨庚(1961-),男,教授,硕/博导,博士,研究方向为隐私保护、云计算与安全、访问控制。

信息泄露的需求因此而产生。

根据响应或协变量是函数还是标量,函数回归模型可以分为四种类型^[2]:(1)带有函数协变量的标量响应;(2)带有标量协变量的函数响应;(3)具有函数协变量的函数响应;(4)具有函数和标量协变量的标量或函数响应。目前,函数回归算法的研究主要集中在模型的优化和计算效率上,而基于函数回归的隐私保护研究还少有人涉足。Janet S. Kim 等人^[3]于2018年提出一种加性的函数对函数回归算法,Mark 等人^[4]针对该模型提出离散小波包变换的算法。针对高维的加性函数模型中 mFPCA 分数的估计误差问题,Wong 等人^[5]提出了一类部分线性泛函可加模型(PLFAM)。该文提出一种函数对函数回归的差分隐私保护算法,即计算函数回归,在回归的过程中加入满足差分隐私的拉普拉斯噪声,以达到隐私保护的作用。

主要贡献如下:

(1)结合函数回归和差分隐私保护的拉普拉斯机制,设计了一种满足 ϵ -差分隐私保护的函数回归算法,并通过理论分析和实验验证其可用性。

(2)使用 B 样条基对函数型数据进行降维和回归处理,允许观测数据含噪,在实现函数回归的基础之上,保证了一定的隐私保护功能。

(3)针对不同隐私预算进行实验,证明隐私预算 ϵ 与算法效率的关系,且添加噪声越小,算法效率越高。

1 相关工作

本节主要介绍差分隐私和函数回归相关的研究工作。

Dwork^[6]于2006年提出了差分隐私的概念,区别于传统的 k-匿名等隐私模型,差分隐私保护模型具有强大的数学模型和坚实的算法设计基础,它可以严格地定义和计算隐私的保护水平,有利于比较和研究在不同参数下的保护水平。目前,差分隐私的机制仍在逐步完善中^[7-9]。

Ramsay 和 Dalzell^[10]于1991年提出了一种函数对函数的回归线性模型,将函数预测器和函数响应回归设置中存在的问题结合在一起,Yao 等人^[11]设计了该模型基于函数主成分(Functional Principal Component, fPC)的方法,假设协变量和响应具有独立同分布的测量误差,并用 fPC 分解进行建模。Wu & Müller^[12]在估计回归系数时使用 WLS 来解释函数内的相关性。文献[13]将函数线性模型扩展到函数可加模型(Functional Addictive Model),该模型通过协变量的函数主成分得分的平滑函数之和对协变量的影响进行建模。Janet S. Kim 等人^[3]在加性函数对函数的回归中提出了一种当前响应与协变量的完整轨迹相关

联的非线性回归模型,可以更直接地捕获响应与完整协变量轨迹之间的复杂关系。

Mark 等人^[4]提出了一种使用离散小波包变换的函数对函数回归模型,适合无约束曲面,但是不适合建模滞后暴露的功能预测因子。Wong 等人^[5]改进了高维加性函数模型中 mFPCA 分数的估计误差,提出了一类部分线性泛函可加模型(PLFAM)。

迄今为止的大多数函数回归研究都假设存在独立同分布的测量误差,但是没有考虑到为观测对象进行隐私保护,也没有考虑实现满足差分隐私的加噪扰动。

2 理论基础

2.1 差分隐私保护技术

差分隐私保证受保护的数据集不会因为增加或删除一条记录而影响查询结果^[14]。其形式化的数学定义如下:

定义1(差分隐私)^[15]:给定邻近数据集(只相差一条记录) D 和 D' ,设有隐私算法 A , $\text{Range}(A)$ 为 A 所有可能的输出结果,若算法 A 在数据集 D 和 D' 上任意输出结果 O ($O \in \text{Range}(A)$) 满足下列不等式:

$$\Pr[A(D) = O] \leq e^\epsilon \times \Pr[A(D') = O] \quad (1)$$

则称算法 A 满足 ϵ -差分隐私, ϵ 的值称为隐私预算, ϵ 越小, $A(D) = O$ 和 $A(D') = O$ 的概率值越接近,算法 A 的隐私保护水平越高。

差分隐私算法满足以下组成属性。假设 $A_1(\cdot)$ 和 $A_2(\cdot)$ 是 ϵ_1 -和 ϵ_2 -差分隐私算法。

· 顺序合成:释放 $A_1(D)$ 和 $A_2(D)$ 的输出满足 $\epsilon_1 + \epsilon_2$ -差分隐私。

· 后处理:对于任何算法 $A_3(\cdot)$,释放 $A_3(A_1(D))$ 仍然满足 ϵ_1 -差分隐私。即对差分隐私算法的输出进行后处理不会导致任何其他隐私损失。

定义2(全局敏感度)^[6]:函数 $f: D \rightarrow R^n$ 的全局灵敏度(表示为 $\Delta(f)$)定义为来自任意两个相邻数据集 D_1 和 D_2 的输出的最大 L_1 距离:

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

其中, R 表示所映射的实数空间, d 表示函数 f 的查询维度。全局敏感度只与函数 f 有关,与数据集 D 无关。

差分隐私保护有两种常用的实现机制:Laplace 机制和指数机制。该文采用的是 Laplace 机制。Laplace 机制的实现方式是通过添加满足 Laplace 分布的随机噪声来达到 ϵ -差分隐私保护的效果。

定义3(Laplace 机制)^[15]:对于任意一个函数 $f: D \rightarrow R^d$,若算法 K 的输出结果满足等式(3),则 K 满足 ϵ -差分隐私:

$$K(D) = f(D) + \langle \text{Lap}_1(\Delta f/\epsilon), \dots, \text{Lap}_d(\Delta f/\epsilon) \rangle \quad (3)$$

其中, $\text{Lap}_i(\Delta f/\varepsilon)$ ($1 \leq i \leq d$) 是相互独立的拉普拉斯变量, 由上式可得: 噪声大小与 Δf 成正比, 与 ε 成反比。

2.2 函数型数据分析

函数型数据分析 (Functional Data Analysis) 是对曲线、曲面或任何其他连续变化的信息的一种统计分析方法, 其协变量或响应为函数型数据^[16]。函数型数据研究的对象是光滑曲线, 例如 $\{x_n(t): t \in [T_1, T_2]\}$, $1 \leq n \leq N$; 其中 $x_n(t) \in R$ 在每一点 $t \in [T_1, T_2]$ 都存在, 取观测点 $\{t_{j,n}: 1 \leq j \leq J_n\}$ 。如下为一个典型的函数型数据集:

$$\{x_n(t_{j,n}) \in R: t_{j,n} \in [T_1, T_2], 1 \leq n \leq N, 1 \leq j \leq J_n\}$$

如果每条曲线的观测数 J_n 都很小, 则称此函数型数据稀疏 (sparse); 例如血检得到的某蛋白浓度。如果每条曲线的观测数 J_n 都很大, 则称此函数型数据密集 (dense); 例如地磁仪记录的某地磁场强度, 高频交易的股票价格^[17]。

3 函数回归的差分隐私保护算法

本节包括函数回归的差分隐私保护算法的各部分概述及具体实现细节, 并给出算法实现差分隐私保护的证明。

3.1 场景描述

对于 $i = 1, 2, \dots, n$, 假设 $\{(X_{ik}, s_{ik}): k = 1, 2, \dots, m_i\}$, $\{(Y_{ij}, t_{ij}): j = 1, 2, \dots, m_{Y,i}\}$, 其中 X_{ik} 和 Y_{ij} 分别是在时间点 s_{ik} 和 t_{ij} 观察到的协变量和响应。对于所有 i 和 k , $s_{ik} \in \Gamma_x$, 以及所有 i 和 j , $t_{ij} \in \Gamma_y$, 其中 Γ_x 和 Γ_y 是紧凑的时间间隔。假设 $X_{ik} = X_i(s_{ik})$, 其中 $X_i(\cdot)$ 是定义在 Γ_x 上的平方可积、真平滑信号。同时假设 $Y_{ij} = Y_i(t_{ij})$, 其中 $Y_i(\cdot)$ 定义在 Γ_y 上。

考虑一个加性的函数对函数回归模型:

$$Y_i(t) = \int_{\Gamma_y} F\{X_i(s), s, t\} ds + \varepsilon_i(t) \quad (4)$$

其中, $F\{\cdot, \cdot, t\}$ 是定义在 $R \times \Gamma_x \times \Gamma_y$ 上的未知平滑三变量函数, $\varepsilon_i(\cdot)$ 是一个误差过程, 具有均值为零和未知的自协方差函数 $R(t, t')$, 并且与协变量 $X_i(s)$ 无关。函数 $F\{\cdot, \cdot, t\}$ 的定义量化了当前响应 $Y_i(t)$ 和完整的协变量轨迹 $X_i(\cdot)$ 之间的未知相关性, 而加性模型则允许对高维数据空间的响应和协变量之间的关系进行非参数建模。

如果 $F(x, s, t) = \beta(s, t)x$, 则模型 (4) 简化为标准函数线性模型。

3.2 数据预处理

由于实际观测的数据存在噪声或测量误差, 在数据预处理阶段, 需要对离散的响应和协变量进行平滑

处理, 使之从离散的多元观测变量变成内部存在关联的函数型数据。

对模型 (4) 中的 F 进行建模, 为了降低计算成本, 减少基函数的数量, 令 $\varphi(\cdot) \in L^2(\Gamma_y)$ 为一平滑函数, 则 Y_i 到 $\varphi(\cdot)$ 的投影为:

$$y_{i,\varphi} = \int_{\Gamma_y} Y_i(t) \varphi(t) dt$$

结合模型 (4) 可推出:

$$y_{i,\varphi} = \int_{\Gamma_x} \int_{\Gamma_y} F\{X_i(s), s, t\} \varphi(t) ds dt + e_{i,\varphi} = \int_{\Gamma_x} G_\varphi\{X_i(s), s\} ds + e_{i,\varphi}$$

其中, $G_\varphi\{X_i(s), s\} = \int_{\Gamma_y} F\{X_i(s), s, t\} \varphi(t) dt$, $e_{i,\varphi} = \int_{\Gamma_y} \varepsilon_i(t) \varphi(t) dt$, 假设积分存在。

令 $\{\varphi_k(\cdot)\}$ k 为一组正交基, 由正交性可得, $k = k'$ 时, 有 $L^2(\Gamma_y): \int_{\Gamma_y} \varphi_k(t) \varphi_{k'}(t) dt = 1$, 否则为 0。将函数 $F(x, s, t)$ 表示为:

$$F(x, s, t) = \sum_{k=1}^{\infty} G_k(x, s) \varphi_k(t)$$

其中, $G_k(x, s) = \int_{\Gamma_y} F(x, s, t) \varphi_k(t) dt$, $k = 1, 2, \dots$, 是在 x 和 s 上平滑变化的未知基系数。将 $G_k(\cdot, \cdot)$ 建模为样条基的张量积:

$$G_k(x, s) = \sum_{l=1}^{K_x} \sum_{l'=1}^{K_s} B_{x,l}(x) B_{s,l'}(s) \theta_{l,l',k}$$

其中, $\{B_{x,l}(x)\}_{l=1}^{K_x}$ 和 $\{B_{s,l'}(s)\}_{l'=1}^{K_s}$ 分别是维度 K_x 和 K_s 的正交 B 样条基。结合上述扩展, 三变量核函数 F 可以写为:

$$F(x, s, t) = \sum_{k=1}^{\infty} \sum_{l=1}^{K_x} \sum_{l'=1}^{K_s} B_{x,l}(x) B_{s,l'}(s) \varphi_k(t) \theta_{l,l',k} \quad (5)$$

其中, $\theta_{l,l',k}$ 是未知参数。因此, 模型 (4) 的三变量函数 F 可由 x 和 s 方向上的单变量 B 样条基函数和 $L^2(\Gamma_y)$ 正交基函数 $\varphi_k(\cdot)$ 的张量积获得, 由于只考虑两个样条基, 减少了所需的基函数和平滑参数, 降低了计算成本, 可以有效提高计算效率。

3.3 DP-in-FRA 算法思路

函数回归的差分隐私保护算法 (Differential Privacy Preservation Algorithm in Functional Regression) 简称 DP-in-FR。

令 Z_i 为 $\int_{\Gamma_y} B_{x,l}\{X_i(s)\} B_{s,l'}(s) ds$ 的 $K_x K_s$ -列向量, 令 Θ_k 为未知系数 $\theta_{l,l',k}$ 的 $K_x K_s$ -列向量, 其中 $l = 1, 2, \dots, K_x$, $l' = 1, 2, \dots, K_s$ 。那么, 模型 (4) 可以被近似为:

$$Y_i(t) \approx \sum_{k=1}^K Z_i^T \Theta_k \varphi_k(t) + \varepsilon_i(t) \quad (6)$$

未知参数 Θ_k 的取值使用惩罚最小二乘法估计,对方向 x 和 s 使用二次惩罚,并通过正交基函数的数量 K 控制 t 方向的粗糙度。由计算可得, x 的方向曲率为:

$$\begin{aligned} & \iiint \{ \partial^2 F(x, s, t) / \partial x^2 \} dx ds dt = \\ & \sum_{k=1}^K \iint \{ \partial^2 G_k(x, s) / \partial x^2 \}^2 dx ds = \\ & \sum_{k=1}^K \Theta_k^T (P_x \otimes I_{K_s}) \Theta_k \end{aligned}$$

其中, \otimes 是克罗内克积, I_K 是 K 维的单位矩阵, P_x 是 $K_x \times K_x$ 的惩罚矩阵, 其 (l, r) 项等于 $\int \{ \partial_{xx} B_{x,l}(x) \} \{ \partial_{xx} B_{x,r}(x) \} dx$, $l, r = 1, 2, \dots, K_x$ 。同理 s 的方向曲率为:

$$\begin{aligned} & \iiint \{ \partial^2 F(x, s, t) / \partial s^2 \} dx ds dt = \\ & \sum_{k=1}^K \iint \{ \partial^2 G_k(x, s) / \partial s^2 \}^2 dx ds = \\ & \sum_{k=1}^K \Theta_k^T (I_{K_x} \otimes P_s) \Theta_k \end{aligned}$$

其中, P_s 是 $K_s \times K_s$ 的惩罚矩阵, 其 (l, r) 项等于 $\int \{ \partial_{ss} B_{s,l}(s) \} \{ \partial_{ss} B_{s,r}(s) \} ds$, $l, r = 1, 2, \dots, K_s$ 。

则最小化的惩罚标准是:

$$\begin{aligned} & \sum_{i=1}^n \| Y_i(\cdot) - \sum_{k=1}^K Z_i^T \Theta_k(\cdot) \|^2 + \sum_{k=1}^K \Theta_k^T (\lambda_x P_x \otimes \\ & I_{K_s} + \lambda_s I_{K_x} \otimes P_s) \Theta_k = \\ & \sum_{k=1}^K [\sum_{i=1}^n \{ \xi_{ik} - Z_i^T \Theta_k \}^2 + \Theta_k^T (\lambda_x P_x \otimes \\ & I_{K_s} + \lambda_s I_{K_x} \otimes P_s) \Theta_k] \quad (7) \end{aligned}$$

其中, $\| \cdot \|^2$ 是对应于内积 $\langle f, g \rangle = \int fg$ 的 L2 范数, λ_x 和 λ_s 是控制函数 F 粗糙度和拟合优度之间权衡的平滑参数, ξ_{ik} 是响应的 FPC 分数, 由函数型数据主成分分析(FPCA)计算得出。

DP-in-FR 对回归模型的系数进行噪声扰动。具体地, 基于惩罚标准(7)中基础系数的估计 $\hat{\Theta}_k$, $k = 1, 2, \dots, K_x$, 随后给该估计加一个密度函数服从 $\text{Lap}(\Delta/\epsilon)$ 分布的噪声:

$$Y_i(t) \approx \sum_{k=1}^K Z_i^T (\Theta_k + \text{Lap}(\frac{\Delta}{\epsilon})) \varphi_k(t) + \varepsilon_i(t) \quad (8)$$

全局敏感度的推导与计算过程如下:

对于邻近数据集 D 和 D' , 以及它们的代价函数 f_D 和 $f_{D'}$:

$$\begin{aligned} f_D &= \sum_{i=1}^n \| Y_i(\cdot) - \sum_{k=1}^K Z_i^T \Theta_k \varphi_k(\cdot) \|^2 \\ f_{D'} &= \sum_{i=1}^{n'} \| Y_i(\cdot) - \sum_{k=1}^K Z_i^T \Theta_k \varphi_k(\cdot) \|^2 \end{aligned}$$

根据全局敏感度的定义(见定义 2)有:

$$\| f_D - f_{D'} \| \leq \max_i \sum_{k=1}^K \{ \xi_{ik} - Z_i^T \Theta_k \}^2$$

由此, 可以得到全局敏感度 Δ 为:

$$\Delta = \max_i \sum_{k=1}^K \{ \xi_{ik} - Z_i^T \Theta_k \}^2 \quad (9)$$

最后将加噪后的基系数的估计 $\hat{\Theta}_k^*$ 代入式(6), 截断点 K 是通过预先指定的解释方差百分比确定的, 该文取值为 95%。

将该算法记为算法 1, 其算法流程如下:

算法 1: DP-in-FR。

输入: 原始数据集 D , 隐私预算, 主成分预设值 p ;

输出: 函数回归模型系数的最优解 $\hat{\Theta}_k^*$ 。

1: 平滑每个 i 的数据重建响应 $Y_i(\cdot)$ 的平滑轨迹并对其去均值 $Y_i^c(\cdot) = Y_i(\cdot) - \mu Y(\cdot)$;

2: 使用函数数据主成分分析(FPCA)估计 $Y_i(\cdot)$ 的(边际)协方差的特征基 $\varphi_k(\cdot)$;

3: 计算 FPC 分数 $\tilde{\xi}_{ik} = \int_{\Gamma_i} Y_i^c(t) \varphi_k(t) dt$;

4: 最小化惩罚标准(8)并使用 $\tilde{\xi}_{ik}$ 替代 ξ_{ik} 获得基础系数的估计 $\hat{\Theta}_k$, $k = 1, 2, \dots, K$;

5: 计算全局敏感度 $\Delta = \max_i \sum_{k=1}^K \{ \xi_{ik} - Z_i^T \Theta_k \}^2$;

6: for $1 \leq k \leq K$ do

7: $\hat{\Theta}_k^* = \hat{\Theta}_k + \text{Lap}(\frac{\Delta}{\epsilon})$

8: end for

9: 返回 $\hat{\Theta}_k^*$

3.4 隐私性分析

定理 1: 算法 1 满足 ϵ -差分隐私保护机制。

证明: 由拉普拉斯噪声机制可知, $\hat{\Theta}_k^* = \hat{\Theta}_k + \text{Lap}(\frac{\Delta}{\epsilon})$ 的输出结果满足拉普拉斯机制的定义(3), 则满足 ϵ -差分隐私。又由差分隐私算法的组合性质可知, 对差分隐私算法的输出进行后处理不会导致任何其他隐私损失, 则 $Y_i(t) \approx \sum_{k=1}^K Z_i^T (\Theta_k + \text{Lap}(\frac{\Delta}{\epsilon})) \varphi_k(t) + \varepsilon_i(t)$ 同样满足 ϵ -差分隐私。

综上所述, 算法 1 满足 ϵ -差分隐私保护机制, 实现了对数据的隐私保护功能。

4 实验与分析

4.1 实验环境

实验环境为 AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz, 16G 内存, Win10 操作系统。算法均采用 R 语言实现, R 语言版本为 R-4.1.0, RTools 版本 4.0, 使用到的程序包有 MASS、Matrix、refund、mgcv、VGAM 等。其中 VGAM 版本为 1.1-5, 用于产生符合拉普拉斯分布的随机噪声。

数据集的具体信息如表 1 所示,分别为加拿大天气数据集、LipEMG 数据和扩散张量成像 (DTI2) 数据。以上数据集分别来自文献[18-19]。表 1 显示了

表 1 数据集信息

Dataset	n	x	$ S $	y	$ T $
Weather	35	Temperature	365	Precipitation	365
LipEMG	32	EMG	501	Acceleration	501
DTI2	340	FA-RCTS	55	FA-CCA	93

为了验证所设计算法的可行性,在这三个数据集上,依次使用文中算法进行训练,通过训练结果的精确度来判断其可用性。此外,为了检测隐私预算 ε 对模型准确性的影响,对每个数据集也以不同的隐私预算 ε 进行多次训练。由于噪声的影响,会进行多次实验取结果的均值。

4.2 实验结果及分析

回归分析有多种性能指标衡量其精确性,该文使用的性能指标是均方根预测误差 (RMSPE) 以及逐点

数据集的统计信息,其中 $|S|$ 和 $|T|$ 是相应域中的数据/时间点个数。

预测区间的平均覆盖概率 (ACP)。通过以下方式定义 RMSPE:

RMSPE =

$$\frac{1}{N} \sum_{r=1}^N \left[n^{-1} \sum_{i=1}^n m_{Y,i}^{-1} \sum_{j=1}^{m_{Y,i}} \{ Y_i^{(r)}(t_{ij}) - \hat{Y}_i^{(r)}(t_{ij}) \}^2 \right]^{1/2}$$

其中, $Y_i^{(r)}(t_{ij})$ 及其估计值 $\hat{Y}_i^{(r)}(t_{ij})$ 来自第 r 次 Bootstrap 采样。RMSPE 捕获预测错误,并且随着样本量的增加,期望 RMSPE 的值收敛到零。

实验结果如图 1 所示。

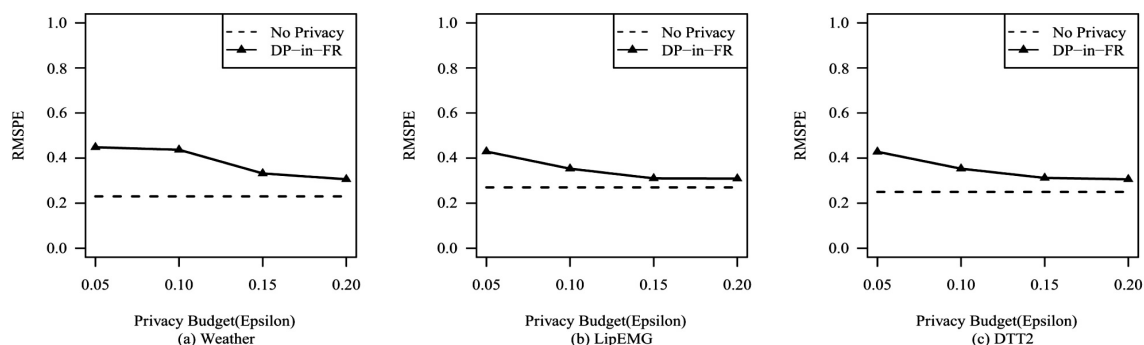


图 1 均方根预测误差

图 1(a)、(b)、(c) 分别是文中算法对三个数据集在不同隐私预算 ε 下训练结果的准确性的比较, ε 的取值范围为 $\{0.05, 0.1, 0.15, 0.2\}$ 。横坐标是隐私预算 ε 的取值,纵坐标是均方根预测误差 RMSPE。标签中, No Privacy 即不添加任何隐私保护机制的函数回归,它将作为算法精确性的比较基准。三个数据集的训练结果均遵循隐私预算越大,训练出的模型精确度越高的规律,并且当隐私预算足够大时,与无隐私保护的算法的精确度接近。

其次,对 $(1-\alpha)$ 水平点态预测区间进行近似,以观察名义水平上的覆盖概率。在 $(1-\alpha)$ 级别定义预测区间的 ACP 如下:

$$ACP_p(1-\alpha) = \sum_{r=1}^N \sum_{i=1}^{100} \sum_{j=1}^{101} I\{Y_{0,i}(t_j) \in P_{1-\alpha,i}^{(r)}(t_j)\} / (100 \cdot 101 \cdot N)$$

其中, $P_{1-\alpha,i}^{(r)}(t_j)$ 是第 r 次 Bootstrap 的逐点预测区间, $I(\cdot)$ 是指示函数。在此计算中,预测区间使用相同的固定测试数据集构建在 Bootstrap 样本上。

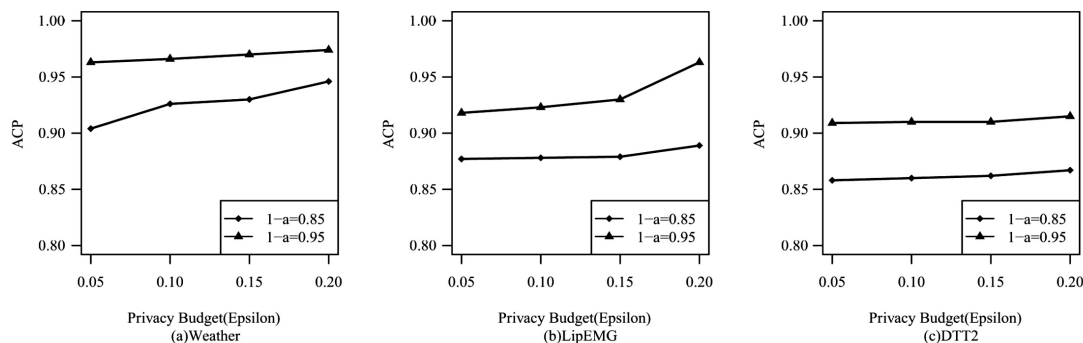


图 2 平均覆盖率

图2(a)、(b)、(c)分别为在 $1-\alpha=0.85$ 和 0.95 的名义显著性水平下,预测响应 $Y(t) \mid X(\cdot)$ 在三个数据集上的平均覆盖概率ACP得分。可以看见随着隐私预算 ε 增大,DP-in-FR算法预测平均覆盖率从整体上看有升高的趋势,这是因为随着 ε 增大,隐私保护程度变低,添加的噪声变小,所以可用性变高,因此预测准确率变高。

5 结束语

主要研究了差分隐私在函数回归中的应用,设计了一种基于差分隐私的函数回归方法。该方法允许观测数据含噪,对函数型数据进行降维和回归处理,在实现函数回归的基础之上,保证了一定的隐私保护功能。该文提出的函数回归算法对于输入数据降维并提取主成分,而隐私预算大小和保留主成分的个数是影响算法误差的因素,合理的加噪方式使得数据可用性更高。由于函数型数据回归的计算量大,计算成本高,所以更合理的隐私预算分配和加噪方式以提高计算效率是下一步的研究方向。

参考文献:

- [1] WANG J L, CHIOU J M, MÜLLER H G. Functional data analysis[J]. Annual Review of Statistics and Its Application, 2016, 3: 257-295.
- [2] MORRIS J S. Functional regression[J]. Annual Review of Statistics and Its Application, 2015, 2: 321-359.
- [3] KIM J S, STAIKU A M, MAITY A, et al. Additive function-on-function regression[J]. Journal of Computational and Graphical Statistics, 2018, 27(1): 234-244.
- [4] MEYER M J, MALLOY E J, COULL B A. Bayesian wavelet-packet historical functional linear models[J]. Statistics and Computing, 2020, 32(2): 2-14.
- [5] WONG R K W, LI Y, ZHU Z. Partially linear functional additive models for multivariate functional data[J]. Journal of the American Statistical Association, 2019, 114(525): 406-418.
- [6] DWORK C. Differential privacy[C]//International colloquium on automata, languages, and programming. Berlin: Springer, 2006: 1-12.
- [7] SORIA-COMAS J, DOMINGO-FERRER J, SÁNCHEZ D, et al. Individual differential privacy: a utility-preserving formulation of differential privacy guarantees[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(6): 1418-1429.
- [8] IMTIAZ H, SARWATE A D. Differentially private distributed principal component analysis[C]//2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). Calgary: IEEE, 2018: 2206-2210.
- [9] BI M, WANG Y, CAI Z, et al. A privacy-preserving mechanism based on local differential privacy in edge computing[J]. China Communications, 2020, 17(9): 50-65.
- [10] RAMSAY J O, DALZELL C J. Some tools for functional data analysis[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1991, 53(3): 539-561.
- [11] YAO F, MÜLLER H G, WANG J L. Functional linear regression analysis for longitudinal data[J]. The Annals of Statistics, 2005, 33(6): 2873-2903.
- [12] WU S, MÜLLER H G. Response-adaptive regression for longitudinal data[J]. Biometrics, 2011, 67(3): 852-860.
- [13] MÜLLER H G, YAO F. Functional additive models[J]. Journal of the American Statistical Association, 2008, 103(484): 1534-1544.
- [14] 谭作文, 张连福. 机器学习隐私保护研究综述[J]. 软件学报, 2020, 31(7): 2127-2156.
- [15] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[J]. Proceedings of the VLDB Endowment, 2006, 7(8): 637-648.
- [16] EARLS C, HOOKER G. Variational bayes for functional data registration, smoothing, and prediction[J]. Bayesian Analysis, 2017, 12(2): 557-582.
- [17] ZHANG J. Analysis of variance for functional data[J]. Monographs on Statistics and Applied Probability, 2014, 127: 127.
- [18] RAMSAY J O, DALZELL C J. Some tools for functional data analysis[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1991, 53(3): 539-561.
- [19] BASSER P J, PAJEVIC S, PIERPAOLI C, et al. In vivo fiber tractography using DT-MRI data[J]. Magnetic Resonance in Medicine, 2000, 44(4): 625-632.