

基于主成分分析和函数机制的差分隐私线性回归算法

李可佳, 胡学先, 陈越, 杨鸿健, 徐阳, 刘扬

引用本文

李可佳, 胡学先, 陈越, 杨鸿健, 徐阳, 刘扬[基于主成分分析和函数机制的差分隐私线性回归算法](#)[J]. 计算机科学, 2023, 50(8): 342-351.

LI Kejia, HU Xuexian, CHEN Yue, YANG Hongjian, XU Yang, LIU Yang. [Differential Privacy Linear Regression Algorithm Based on Principal Component Analysis and Functional Mechanism](#) [J]. Computer Science, 2023, 50(8): 342-351.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[面向纵向图联邦学习的数据重构攻击方法](#)

Data Reconstruction Attack for Vertical Graph Federated Learning

计算机科学, 2023, 50(7): 332-338. <https://doi.org/10.11896/jsjcx.220900038>

[一种基于自适应加权的鲁棒联邦学习算法](#)

Robust Federated Learning Algorithm Based on Adaptive Weighting

计算机科学, 2023, 50(6A): 230200188-9. <https://doi.org/10.11896/jsjcx.230200188>

[差分隐私研究进展综述](#)

Review of Differential Privacy Research

计算机科学, 2023, 50(4): 265-276. <https://doi.org/10.11896/jsjcx.220500292>

[类脑心智计算的科学技术和工程应用的研究与思考](#)

Thoughts on Development and Research of Science, Technology and Engineering Application of Brain & Mind-inspired Computing

计算机科学, 2023, 50(2): 364-373. <https://doi.org/10.11896/jsjcx.220500023>

[RCP:本地差分隐私下的均值保护技术](#)

RCP:Mean Value Protection Technology Under Local Differential Privacy

计算机科学, 2023, 50(2): 333-345. <https://doi.org/10.11896/jsjcx.220700273>

基于主成分分析和函数机制的差分隐私线性回归算法

李可佳¹ 胡学先¹ 陈越¹ 杨鸿健¹ 徐阳¹ 刘扬^{1,2}

1 中国人民解放军战略支援部队信息工程大学 郑州 450001

2 郑州大学网络空间安全学院 郑州 450001

(kejia_li@outlook.com)

摘要 随着人工智能应用的不断落地以及隐私保护法律法规的持续出台,机器学习中的隐私保护问题已成为目前信息安全领域的一个研究热点。文中针对现有的差分隐私线性回归算法全局敏感度高、模型可用性较差的问题,基于高斯机制代替传统的Laplace机制,并通过在算法的两个主要阶段分别添加噪声的方法,提出了一种基于主成分分析和函数机制的差分隐私线性回归算法(PCAFM-DPLR)。首先,为了在降维的同时兼顾数据的隐私性,向原始数据集的协方差矩阵中注入高斯噪声,基于主成分分析得到具有差分隐私保护效果的低维数据集;其次,为防止模型训练过程中可能存在的隐私泄露,再向目标函数的展开多项式系数添加高斯噪声,并以扰动后的目标函数最小化为目标,求得最优模型参数。理论分析和实验结果表明,PCAFM-DPLR算法训练出的线性回归模型能够在有效保证隐私性的同时,具有良好的可用性。

关键词: 差分隐私;主成分分析;函数机制;线性回归;高斯噪声

中图法分类号 TP391

Differential Privacy Linear Regression Algorithm Based on Principal Component Analysis and Functional Mechanism

LI Kejia¹, HU Xuexian¹, CHEN Yue¹, YANG Hongjian¹, XU Yang¹ and LIU Yang^{1,2}

1 The PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China

2 School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450001, China

Abstract With the continuous development of artificial intelligence applications and the subsequent promulgation of privacy protection laws and regulations, the privacy protection issue in machine learning has become a research hotspot in the field of information security. To overcome the issues of high global sensitivity and poor model usability of the existing differential privacy linear regression algorithms, we present a differential privacy linear regression algorithm based on principal component analysis and functional mechanism (PCAFM-DPLR). In the PCAFM-DPLR algorithm, the traditional Laplace mechanism is replaced by the Gaussian mechanism, and the noise is added in the two major stages of the algorithm respectively. First, in order to take into account the privacy of the data while reducing the dimensionality, Gaussian noise is injected into the covariance matrix of the original data set, and a low-dimensional data set with differential privacy protection effect is obtained based on principal component analysis. Second, to prevent the possible privacy leakage during the model training, Gaussian noise is then added to the expansion polynomial coefficients of the objective function, and the minimization of the perturbed objective function is used as the objective to find the optimal model parameters. Theoretical analysis and experimental results show that the linear regression model trained by the PCAFM-DPLR algorithm can effectively guarantee privacy while having good utility.

Keywords Differential privacy, Principal component analysis, Functional mechanism, Linear regression, Gaussian noise

1 引言

随着云计算、大数据、人工智能等技术的快速发展和应用,海量数据中蕴含的巨大价值和发展潜力日益彰显。机器学习作为人工智能的一种实现方式,旨在从数据中学习到一个性能良好的预测模型,从而对未知的数据进行决策判断。

然而,已有研究表明,恶意的攻击者能够通过与机器学习模型之间的交互,逆向推理出训练数据中的敏感信息^[1],这使得机器学习过程中的隐私保护成为了亟待解决的问题。

差分隐私^[2]作为一种建立在坚实数学理论基础之上、量化的隐私保护技术,被证明能够有效保护个体的隐私数据^[3]。因此,其一经提出就受到了学术界和工业界的认可,

到稿日期:2022-08-29 返修日期:2023-03-23

基金项目:国家自然科学基金(62172433,62172434,61862011,61872449)

This work was supported by the National Natural Science Foundation of China(62172433,62172434,61862011,61872449).

通信作者:胡学先(xuexian_hu@hotmail.com)

例如基于差分隐私的直方图发布方法^[4]、划分发布方法^[5]、频繁模式挖掘^[6]、分类分析^[7]、聚类分析^[8]、回归分析^[9]等挖掘和学习方法,在数据发布、数据挖掘、机器学习等领域均发挥着重要的作用。

回归算法作为机器学习中最基本的模型之一,被广泛应用于药物剂量预测、投资风险分析、医疗产品管理等领域。然而,传统的回归算法也面临着隐私泄露的风险。Fredrikson等^[10]的研究表明,对基于线性回归算法设计的私人化用药系统进行模型反演攻击,不仅能够提取到病人的基因信息,还能影响药物的正确配置,威胁病人的生命安全。差分隐私能够在保留整个数据集基本统计信息的同时,防止攻击者从发布的结果中推断出单条记录的敏感信息^[11]。因此,诸多学者对基于差分隐私的回归算法展开了一系列的研究^[12-16],根据模型训练过程中添加噪声的位置不同,基于差分隐私的回归分析算法可分为输入扰动、目标扰动和输出扰动三大类。

在基于输出扰动的回归分析算法中,Dwork等^[9]采用直接扰动回归任务最优模型参数的方法来保护隐私,但该方法在计算模型参数的全局敏感度时受 argmin 函数的制约,并且在计算最优模型参数时要求其输出空间是有界的。但事实上,模型参数的输出空间通常是无界的,故此方案代价较大且通用性不强^[12]。因此,Smith^[12]选择在回归模型的输出结果上添加 Laplace 噪声。该方法能够有效防止攻击者从输出结果中推断出个体的敏感信息,但是会严重降低回归结果的可用性。

Lei^[13]基于输入扰动的思想,提出了一种回归分析算法 DPME。该算法在模型开始训练前,利用 Laplace 机制生成原始数据集的多维噪声直方图,并利用该噪声直方图生成与之相匹配的合成数据集,最后利用此合成数据集计算回归任务的模型参数。然而,该算法只能在低维数据集上使用^[14]。

Chaudhuri等^[15]基于目标扰动的思想,向训练数据集的目标函数进行 Laplace 扰动,并利用扰动后的目标函数求出最优模型参数。然而,该方法要求目标函数是凸函数且连续、可微,具有一定的局限性。鉴于此,Zhang等^[16]提出了函数机制(Functional Mechanism,FM)的思想,通过向目标函数的展开多项式系数中添加 Laplace 噪声,来达到扰动目标函数的效果,进而求得最优模型参数。然而,该方法存在着两个不足:1)全局敏感度的大小与训练数据集的维度正相关,当数据集的维度较高时,全局敏感度较大;2)对目标函数的多项式系数添加的是满足 ϵ -差分隐私保护的 Laplace 噪声,模型的可用性较差^[11]。

从已有的研究工作可以看出,多数基于差分隐私的回归分析算法都依赖于 Laplace 机制来实现差分隐私,大多存在全局敏感度高、模型可用性较差的问题。相比 Laplace 机制而言,高斯机制的分布函数会更紧密地集中在 0 附近,故在 ϵ 相同的情况下,高斯机制产生的噪声更少^[17]。文献[18]从理论分析和实验研究两个角度证明了高斯机制在 ϵ 很小时也能够产生较少的噪声。

因此,本文采用高斯机制代替传统的 Laplace 机制,降低注入到目标函数展开式系数中的噪声量。同时,利用主成分

分析对原始数据集进行降维处理,选用能够反映原始数据集绝大部分信息的低维数据集进行模型训练,从而将全局敏感度控制在一个较低的水平。进一步,通过分别向主成分分析降维过程中的协方差矩阵和回归分析过程中的目标多项式系数添加噪声的方法,提出了一种基于主成分分析和函数机制的差分隐私线性回归(PCAFM-DPLR)算法。

具体而言,本文的贡献主要有以下 3 个方面。

(1)提出了一种基于主成分分析的差分隐私数据降维方法。利用主成分分析技术,将原始高维数据转换为一组由线性无关变量表示的低维数据,并通过向原始高维数据的协方差矩阵添加高斯噪声的方式,来获得具有差分隐私保护效果的低维数据集,以供后续的回归任务使用。

(2)提出了一种基于主成分分析和函数机制的差分隐私线性回归算法,该算法能够在尽可能保留原始数据集统计特性的基础上,降低目标函数的全局敏感度。并通过添加高斯噪声让算法满足宽松的差分隐私定义,更好地平衡了模型的隐私性与可用性。

(3)通过严格的理论分析,证明了 PCAFM-DPLR 算法满足 (ϵ,δ) -差分隐私。通过实验仿真,验证了此算法能够在保证模型隐私性的前提下,具有良好的可用性。

2 预备知识

本节主要介绍线性回归、差分隐私、主成分分析和函数机制的基础概念。本文中的主要符号定义如表 1 所列。

表 1 主要符号定义
Table 1 Definition of main notations

符号	定义
$\mathbf{D}$	原始 d 维数据集,共有 n 条数据样本
$t_i=(\mathbf{x}_i,y_i)$	第 i 条数据样本
$\boldsymbol{\omega}$	线性回归模型的参数向量
$f_{\mathbf{D}}(\boldsymbol{\omega})$	线性回归模型的目标函数
$\boldsymbol{\omega}^*$	最优模型参数
$\varphi(\boldsymbol{\omega})$	任意多个 $\boldsymbol{\omega}$ 中的元素的乘积,如 $\omega_1 \cdot \omega_2$
Φ_S	$\varphi(\boldsymbol{\omega})$ 中次数为 S 的单项式的集合
$\lambda_{\phi t_i}$	目标函数 $f_{\mathbf{D}}(\boldsymbol{\omega})$ 的展开多项式系数
$\mathbf{A}$	协方差矩阵
θ	保留原始信息的百分比阈值
$\tilde{\mathbf{A}}$	加噪后的协方差矩阵
$\tilde{\mathbf{D}}_L$	降维后的 k 维数据集
$f_{\tilde{\mathbf{D}}_L}(\boldsymbol{\omega})$	降维后的目标函数

2.1 线性回归

回归分析是机器学习中常用的分析方法之一,其中线性回归(Linear Regression,LR)是对两种或两种以上变量之间相互依赖的定量关系进行建模的一种线性方法。其具体描述如下。

给定数据集 $\mathbf{D}=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\cdots,(\mathbf{x}_n,y_n)\}$,共包含 n 条数据样本,其中每条数据样本包含 d 个特征属性和 1 个标签值,记为 $x_1,x_2,\cdots,x_d$ 和 y 。第 i 条数据样本表示为 $t_i=(\mathbf{x}_i,y_i)=(x_{i1},x_{i2},\cdots,x_{id},y_i)$ 。为了方便起见,本文将数据集 $\mathbf{D}$ 表示为属性集 $\mathbf{X}$ 和标签集 $\mathbf{y}$ 的组合,即 $\mathbf{D}=\{\mathbf{X},\mathbf{y}\}=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\cdots,(\mathbf{x}_n,y_n)\}$,其中 $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$ 为特征属性组成的矩阵,每一个行向量对应一个 d 维

特征向量 $\mathbf{x}_i^T=(x_{i1},x_{i2},\cdots,x_{id})$,且满足 $\sqrt{\sum_{j=1}^dx_{ij}^2}\leqslant 1,\mathbf{y}=(y_1,y_2,\cdots,y_n)^T$ 是与 n 个 d 维特征向量一一对应的标签向量,满足 $y_i\in[-1,1]$ 。

定义 1(线性回归) 给定数据集 $\mathbf{D}=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\cdots,(\mathbf{x}_n,y_n)\}$,其中 $\sqrt{\sum_{j=1}^dx_{ij}^2}\leqslant 1,y_i\in[-1,1]$,线性回归的预测函数为 $\rho(\mathbf{x}_i)=\mathbf{x}_i^T\cdot\boldsymbol{\omega}^*$,其中 $\boldsymbol{\omega}^*$ 是使目标函数 $f_D(\boldsymbol{\omega})=\sum_{i=1}^nf(t_i,\boldsymbol{\omega})=\sum_{i=1}^n(y_i-\mathbf{x}_i^T\boldsymbol{\omega})^2$ 最小化的参数向量,即:

$$\boldsymbol{\omega}^*=\arg\min_{\boldsymbol{\omega}}\sum_{i=1}^n(y_i-\mathbf{x}_i^T\boldsymbol{\omega})^2\tag{1}$$

2.2 差分隐私

差分隐私(Differential Privacy,DP)是一种建立在严格数学定义之上的隐私保护技术,最早是由 Dwork 于 2006 年提出的^[2]。满足差分隐私定义的算法的输出结果不会因增加、删除或修改数据集中的任何一条记录而发生显著变化,确保攻击者无法根据输出结果的差异推测出个体的敏感信息。其具体定义如下。

定义 2(ϵ -差分隐私^[2]) 给定一种随机算法 M ,其定义域为 $\mathcal{D}$,所有可能输出的集合为 $\mathcal{R}$ 。若对于任意的相邻数据集 $\mathbf{D},\mathbf{D}'\in\mathcal{D}$ 和任意的输出结果 $S\subseteq\mathcal{R}$,均满足:

$$Pr[M(\mathbf{D})\in S]\leqslant e^\epsilon Pr[M(\mathbf{D}')\in S]\tag{2}$$

则称算法 M 满足 ϵ -差分隐私。其中 ϵ 为隐私预算,用于控制隐私保护水平。一般而言, ϵ 越小,隐私保护程度就越高。

ϵ -差分隐私要求算法在最大背景知识攻击的假设下仍能提供隐私保护,但在实际应用中这种假设往往过于保守,若一味严格遵循这种保守的假设来设计隐私算法,则往往会破坏数据隐私与模型可用性之间的平衡^[19]。因此,为了扩展差分隐私的应用范围,使其能够被广泛应用于更多不同的场景,研究者们对差分隐私进行了多种宽松的定义。 (ϵ,δ) -差分隐私^[17]是最早提出的一种差分隐私的宽松定义,其形式化定义如下。

定义 3((ϵ,δ) -差分隐私^[17]) 给定一个随机算法 M ,其定义域为 $\mathcal{D}$,所有可能输出的集合为 $\mathcal{R}$,若对任意的相邻数据集 $\mathbf{D},\mathbf{D}'\in\mathcal{D}$ 和任意的输出结果 $S\subseteq\mathcal{R}$,均满足:

$$Pr[M(\mathbf{D})\in S]\leqslant e^\epsilon Pr[M(\mathbf{D}')\in S]+\delta\tag{3}$$

则称算法 M 满足 (ϵ,δ) -差分隐私。其中 ϵ 为隐私预算,用于控制隐私保护水平; δ 是一个松弛项,表示可以容忍隐私预算超出 ϵ 的概率,通常将其设置为一个很小的值。

差分隐私的本质是通过添加随机噪声来提供保护,并且添加的噪声量与相应的函数 f 的全局敏感度密切相关。噪声添加机制是实现差分隐私保护的主要技术,针对数值型数据,常用的噪声添加机制主要有 Laplace 机制^[9]和高斯机制^[20]。针对非数值型数据,常采用指数机制^[21]。

定义 4(全局敏感度^[22]) 给定任意函数 $f:\mathcal{D}\rightarrow\mathcal{R}$,对于任意的相邻数据集 $\mathbf{D},\mathbf{D}'\in\mathcal{D}$, f 的全局敏感度为:

$$\Delta f=\max_{\mathbf{D},\mathbf{D}'\in\mathcal{D}}\|f(\mathbf{D})-f(\mathbf{D}')\|_q\tag{4}$$

其中, q 的取值可以为 1 或 2。当 q 为 1 时, $\|f(\mathbf{D})-f(\mathbf{D}')\|_1$ 为函数 f 在相邻数据集 $\mathbf{D}$ 和 $\mathbf{D}'$ 之间的一阶范数距离,对应的 Δf 为函数 f 的 l_1 -敏感度。当 q 为 2 时, $\|f(\mathbf{D})-f(\mathbf{D}')\|_2$

为函数 f 在相邻数据集 $\mathbf{D}$ 和 $\mathbf{D}'$ 之间的二阶范数距离,对应的 Δf 为函数 f 的 l_2 -敏感度,也记为 $\Delta_2 f$ 。

Laplace 机制通过向数值型的查询结果中加入服从 Laplace 分布的随机噪声来实现 ϵ -差分隐私保护。高斯机制通过向确切的查询结果中加入服从高斯分布的随机噪声来实现 (ϵ,δ) -差分隐私保护。

定义 5(Laplace 机制^[9]) 给定任意的函数 $f:\mathcal{D}\rightarrow\mathbb{R}^d$,其中 $\mathbb{R}^d$ 为 d 维实空间, f 的敏感度为 Δf ,则对任意的数据集 $\mathbf{D}\in\mathcal{D}$,随机算法 $M(\mathbf{D})=f(\mathbf{D})+(Y_1,Y_2,\cdots,Y_d)$ 满足 ϵ -差分隐私。其中, $Y_i\sim Lap(0,\Delta f/\epsilon)$,即 $\{Y_i,1\leqslant i\leqslant d\}$ 独立同分布且服从于位置参数为 0,尺度参数为 $\Delta f/\epsilon$ 的 Laplace 分布。

定义 6(高斯机制^[20]) 给定任意的函数 $f:\mathcal{D}\rightarrow\mathbb{R}^d$,其中 $\mathbb{R}^d$ 为 d 维实空间, f 的敏感度为 $\Delta_2 f$,则对于任意的数据集 $\mathbf{D}\in\mathcal{D}$,随机算法 $M(\mathbf{D})=f(\mathbf{D})+(Y_1,Y_2,\cdots,Y_d)$ 满足 (ϵ,δ) -差分隐私。其中,对于任意的 $\epsilon\in(0,1),\sigma\geqslant\sqrt{2\ln(1.25/\delta)}\Delta_2 f/\epsilon$,随机噪声 $Y_i\sim N(0,\sigma^2)$,即 $\{Y_i,1\leqslant i\leqslant d\}$ 独立同分布且服从于数学期望为 0、方差为 σ^2 的高斯分布。

差分隐私具有两个重要的组合性质,分别是序列组合性和并行组合性。在合理分配隐私预算 ϵ 的过程中,可以充分利用这两个重要性质^[22]。

性质 1(序列组合性^[23]) 设有随机算法 $M_i:\mathcal{D}\rightarrow\mathcal{R}_i,i\in[n]$ 满足 ϵ_i -差分隐私,那么当它们独立作用于同一数据集 $\mathbf{D}\in\mathcal{D}$ 时,由这些算法构成的组合算法 $M_{[n]}(\mathbf{D})=(M_1(\mathbf{D}),M_2(\mathbf{D}),\cdots,M_n(\mathbf{D}))$ 满足 $(\sum_{i=1}^n\epsilon_i)$ -差分隐私。

性质 2(并行组合性^[23]) 设有随机算法 $M_i:\mathcal{D}\rightarrow\mathcal{R}_i,i\in[n]$ 满足 ϵ_i -差分隐私,那么当它们分别作用于 n 个不相交的数据集 $\mathbf{D}_1,\mathbf{D}_2,\cdots,\mathbf{D}_n$ 时,由这些算法构成的组合算法 $M_{[n]}(\mathbf{D})=(M_1(\mathbf{D}_1),M_2(\mathbf{D}_2),\cdots,M_n(\mathbf{D}_n))$ 满足 $(\max_{i\in[n]}\epsilon_i)$ -差分隐私。

2.3 主成分分析

主成分分析(Principal Component Analysis,PCA)最早是由 Pearson^[24]提出的一种多元统计方法,该方法通过正交变换将原始数据转换成一组由线性无关变量表示的数据,以达到数据降维的目的。其具体描述如下。

给定标准化后的数据集 $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n]^T\in\mathbb{R}^{n\times d}$,其中 n 为样本总数, d 为每条数据样本的属性个数。则原始数据集的协方差矩阵为:

$$\mathbf{A}=\frac{1}{n}\mathbf{X}^T\mathbf{X}=\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^T=\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T\in\mathbb{R}^{d\times d}\tag{5}$$

对 $\mathbf{A}$ 进行特征分解后得到一组降序排列的特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_d>0$ 及相对应的单位特征向量 $\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_d$,且 $\boldsymbol{\Sigma}=\text{diag}(\lambda_1,\lambda_2,\cdots,\lambda_d)$ 。其中,特征值越大,其相对应的主成分越重要。通过指定欲保留的主成分比重阈值 $\theta\in(0,1]$ 来确定目标维度 k 的大小。具体来说, k 可以由式(6)得到:

$$\frac{\sum_{i=1}^k\lambda_i}{\sum_{i=1}^d\lambda_i}\geqslant\theta\tag{6}$$

取前 k 个特征值对应的特征向量组成主成分空间

$\mathbf{V}_{d \times k} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k)$, 其中 $\mathbf{v}_i$ 和 $\mathbf{v}_j$ 是相互正交的单位向量。将原始数据 $\mathbf{X}_{n \times d}$ 投影到由前 k 个特征向量构成的主成分空间 $\mathbf{V}_{d \times k}$ 中, 得到降维后的 k 维数据集。

$$\mathbf{Z}_{n \times k} = \mathbf{X}_{n \times d} \mathbf{V}_{d \times k} \quad (7)$$

2.4 函数机制

函数机制是一种由 Laplace 机制扩展而来的隐私保护技术, 最早由 Zhang 等^[16]提出。不同于 Laplace 机制直接向回归分析结果中注入噪声的方式, FM 向回归分析算法的目标函数 $f_D(\omega)$ 的展开多项式系数中添加 Laplace 噪声, 得到扰动后的目标函数 $\tilde{f}_D(\omega)$, 进而计算出带有隐私保护效果的最优模型参数 $\tilde{\omega}^*$ 。

给定训练数据集 $\mathbf{D} = \{\mathbf{X}, \mathbf{y}\} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$, 其中 $\mathbf{t}_i = (\mathbf{x}_i, y_i) = (x_{i1}, x_{i2}, \dots, x_{id}, y_i)$ 。由于线性回归分析算法的目标函数 $f_D(\omega) = \sum_{i=1}^n f(\mathbf{t}_i, \omega) = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \omega)^2$ 是模型参数 $\omega = (\omega_1, \omega_2, \dots, \omega_d)$ 的二次多项式, 因此可以将其表示成形如式(8)所示的代数表达式。

$$f_D(\omega) = \sum_{s=0}^2 \sum_{\phi(\omega) \in \Phi_s} \sum_{t_i \in \mathbf{D}} \lambda_{\phi t_i} \phi(\omega) \quad (8)$$

其中, $\Phi_s = \{\omega_1^{c_1} \times \omega_2^{c_2} \times \dots \times \omega_d^{c_d} \mid \sum_{l=1}^d c_l = s\}$ 表示 ω 的所有 S 次单项式组成集合, 如 $\Phi_0 = \{1\}$, $\Phi_1 = \{\omega_1, \dots, \omega_d\}$, $\Phi_2 = \{\omega_i \times \omega_j \mid 1 \leq i, j \leq d\}$, 实数 $\lambda_{\phi t_i} \in \mathbb{R}$ 表示对应于单项式 $\phi(\omega) = \omega_1^{c_1} \times \omega_2^{c_2} \times \dots \times \omega_d^{c_d} \in \Phi_s$ 的系数。此时, 函数 $f(\mathbf{t}_i, \omega)$ 可以表示为:

$$f(\mathbf{t}_i, \omega) = \sum_{s=0}^2 \sum_{\phi(\omega) \in \Phi_s} \lambda_{\phi t_i} \phi(\omega)$$

在上述表示下, 由文献[16]可知, 目标函数 $f_D(\omega)$ 的

全局敏感度为:

$$\Delta f = \max_{\mathbf{D}, \mathbf{D}'} \|f_D(\omega) - f_{D'}(\omega)\|_1 \leq 2 \max_{t, s=1}^2 \sum_{\phi(\omega) \in \Phi_s} \|\lambda_{\phi t}\|_1 \quad (9)$$

对该目标函数 $f_D(\omega)$ 的多项式系数 $\lambda_{\phi t_i}$ 注入满足 $Lap(\Delta f/\epsilon)$ 分布的 Laplace 噪声, 则可以达到直接扰动 $f_D(\omega)$ 的效果。令扰动后的目标函数 $\tilde{f}_D(\omega)$ 最小, 可计算出模型参数向量 $\tilde{\omega}^*$, 进一步得到具有隐私保护效果的线性回归模型的预测函数 $\rho(\mathbf{x}_i) = \mathbf{x}_i^T \tilde{\omega}^*$ 。

3 基于 PCA 和 FM 的差分隐私线性回归算法

本节主要介绍 PCAFM-DPLR 算法的整体框架以及算法的具体流程, 其中包括基于主成分分析的差分隐私数据降维算法 DP-PCA 以及扰动目标函数的差分隐私线性回归算法 FM-DPLR。

3.1 算法整体框架

本文在设计 PCAFM-DPLR 算法时主要考虑两个原则: (1) 在数据降维阶段, 尽可能兼顾数据的隐私性与可用性; (2) 在模型训练阶段, 尽可能保证模型的预测准确性。

针对原则(1), 考虑向原始数据集 $\mathbf{D} = \{\mathbf{X}, \mathbf{y}\}$ 的协方差矩阵中注入服从高斯分布的随机噪声, 得到扰动后的主成分空间 $\tilde{\mathbf{V}}_{d \times k}$, 进而计算出扰动后的低维数据集 $\tilde{\mathbf{D}}_L = \{\tilde{\mathbf{Z}}, \mathbf{y}\}$ 。

针对原则(2), 本文通过向线性回归目标函数 $f_{\tilde{\mathbf{D}}_L}(\omega)$ 的展开多项式系数中注入高斯噪声, 得到扰动后的目标函数 $\tilde{f}_{\tilde{\mathbf{D}}_L}(\omega)$, 并通过 $\tilde{f}_{\tilde{\mathbf{D}}_L}(\omega)$ 计算出最优模型参数 $\tilde{\omega}^*$ 。

PCAFM-DPLR 算法的整体框架如图 1 所示。

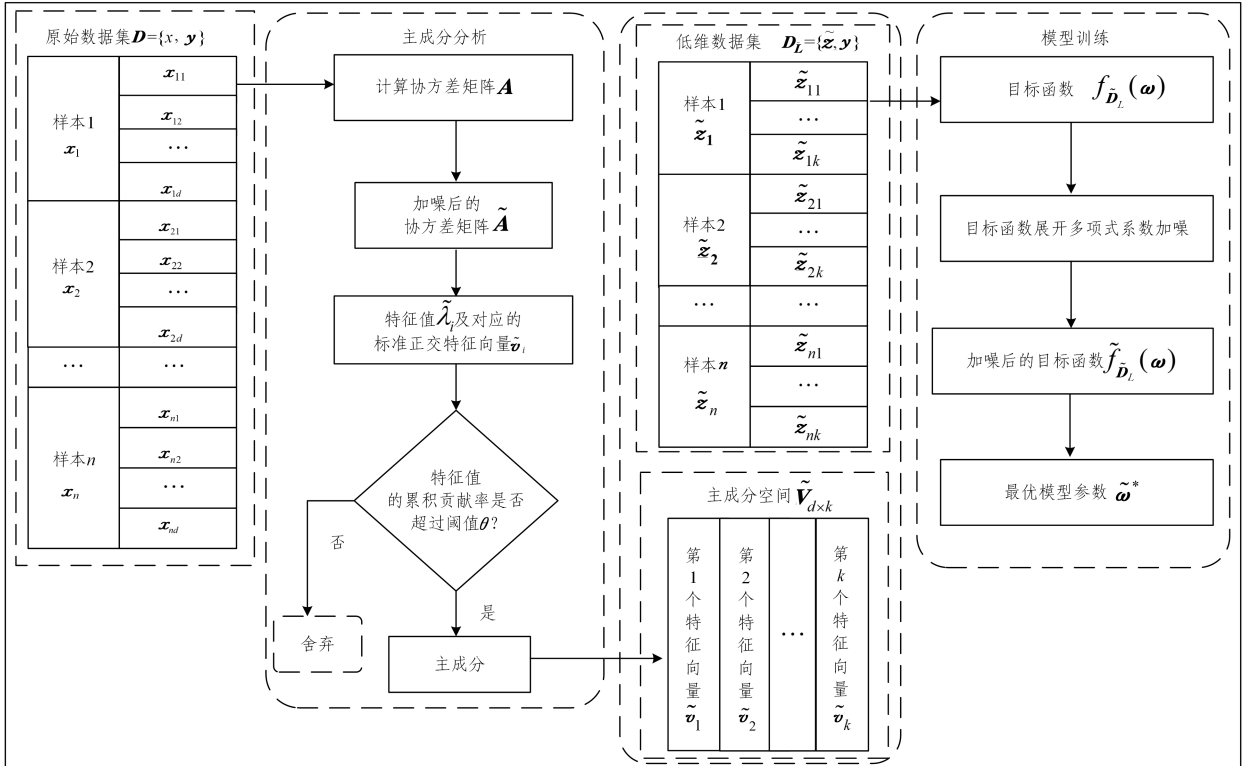


图 1 PCAFM-DPLR 算法框架的示意图

Fig. 1 Illustration of PCAFM-DPLR algorithm framework

3.2 PCAFM-DPLR 算法的具体步骤

基于 PCA 和函数机制的差分隐私线性回归算法具体分为两步:1)利用 PCA 和差分隐私对原始数据进行降维处理;2)利用 FM 对线性回归目标函数进行扰动,并通过扰动后的目标函数求解最优化模型参数,得到满足 (ϵ, δ) -差分隐私的最优模型参数。

首先,对原始 d 维数据集 $\mathbf{D}$ 进行标准化处理,使得对于任意的 $\mathbf{x}_i \in X$, 均满足 $\|\mathbf{x}_i\|_2 \leq 1, y_i \in [-1, 1]$ 。接着,通过 DP-PCA 算法提取数据的主成分。在此阶段,向协方差矩阵 $\mathbf{A} = \frac{1}{n} \mathbf{X}^T \mathbf{X}$ 的每一个元素注入服从高斯分布的随机噪声。然后,对扰动后的协方差矩阵 $\tilde{\mathbf{A}}$ 进行特征值分解,得到特征值 $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \dots \geq \tilde{\lambda}_d$ 和特征向量 $\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \dots, \tilde{\mathbf{v}}_d$ 。随后,通过阈值 θ 及各维度特征值占比,确定目标维度 k , 从而得到扰动后的主成分空间 $\tilde{\mathbf{V}}_{d \times k}$ 。最后,将原始属性集投影到主成分空间 $\tilde{\mathbf{V}}_{d \times k}$ 中,得到降维后的属性集 $\tilde{\mathbf{Z}}_{n \times k}$, 并构成低维数据集 $\tilde{\mathbf{D}}_L$ 。其详细流程如算法 1 所示。

算法 1 DP-PCA 算法

输入:原始数据集 $\mathbf{D} = \{\mathbf{X}, \mathbf{y}\}$, 隐私参数 $\epsilon_1, \delta_1 > 0$, 目标阈值 θ

输出:低维数据集 $\tilde{\mathbf{D}}_L$

1. 计算协方差矩阵 $\mathbf{A} = \frac{1}{n} \mathbf{X}^T \mathbf{X} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$;
2. 生成对称高斯噪声矩阵 $\mathbf{E} = [e_{ij}] \in \mathbb{R}^{d \times d}$, 其中 $e_{ij} \sim N(0, \sigma_1^2), i = 1, \dots, d, i \geq j, e_{ji} = e_{ij}, i < j, \sigma_1 = \sqrt{2 \ln(1.25/\delta_1)} / n \epsilon_1$;
3. 计算扰动后的协方差矩阵 $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ 的特征值 $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \dots \geq \tilde{\lambda}_d$ 和特征向量 $\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \dots, \tilde{\mathbf{v}}_d$;
4. 通过累积方差贡献率 $\sum_{i=1}^k \tilde{\lambda}_i / \sum_{i=1}^d \tilde{\lambda}_i \geq \theta$ 来确定目标维度 k 的值;
5. 选取前 k 个特征值对应的 k 个特征向量组成主成分空间 $\tilde{\mathbf{V}}_{d \times k} = \{\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \dots, \tilde{\mathbf{v}}_k\}$ 。
6. 计算降维后的 k 维属性集 $\tilde{\mathbf{Z}}_{n \times k} = \mathbf{X}_{n \times d} \tilde{\mathbf{V}}_{d \times k}$;
7. 得到降维后的 k 维数据集 $\tilde{\mathbf{D}}_L = \{(\tilde{\mathbf{z}}_1, y_1), (\tilde{\mathbf{z}}_2, y_2), \dots, (\tilde{\mathbf{z}}_n, y_n)\}$ 。

随后,利用 DP-PCA 算法降维处理后的低维数据集 $\tilde{\mathbf{D}}_L$ 训练线性回归模型,其中 $\mathbf{r}_i = (\tilde{\mathbf{z}}_i, y_i) = (\tilde{z}_{i1}, \tilde{z}_{i2}, \dots, \tilde{z}_{ik}, y_i)$ 。首先,对 k 维数据集 $\tilde{\mathbf{D}}_L$ 进行标准化处理,使得对于任意的 $\tilde{\mathbf{z}}_i \in \tilde{\mathbf{Z}}$, 均满足 $\|\tilde{\mathbf{z}}_i\|_2 \leq 1, y_i \in [-1, 1]$ 。接着,计算目标函数 $f_{\tilde{\mathbf{D}}_L}(\omega)$ 以及其展开多项式系数的全局敏感度 $\Delta_2 f$ 。然后,向展开式系数中注入服从高斯分布的随机噪声,得到扰动后的目标函数 $\tilde{f}_{\tilde{\mathbf{D}}_L}(\omega)$ 。最后,通过 $\tilde{f}_{\tilde{\mathbf{D}}_L}(\omega)$ 最小化计算出最优模型参数 $\tilde{\omega}^*$, $\tilde{\omega}^*$ 是一个 k 维的实数向量。具体的训练过程如算法 2 所示。

算法 2 FM-DPLR 算法

输入:数据集 $\tilde{\mathbf{D}}_L = \{\tilde{\mathbf{Z}}, \mathbf{y}\}$, 隐私参数 $\epsilon_2, \delta_2 > 0$

输出:最优模型参数 $\tilde{\omega}^* = (\tilde{\omega}_1^*, \tilde{\omega}_2^*, \dots, \tilde{\omega}_k^*)$

1. 计算目标函数

$$f_{\tilde{\mathbf{D}}_L}(\omega) = \sum_{i=1}^n f(\mathbf{r}_i, \omega) = \sum_{i=1}^n (y_i - \tilde{\mathbf{z}}_i^T \omega)^2;$$

2. 计算敏感度

$$\Delta_2 f = 2 \max_{\mathbf{r}} \sum_{s=1}^S \sum_{\phi \in \Phi_s} \|\lambda_{\phi} \mathbf{r}\|_2;$$

3. 计算扰动目标函数的展开式系数

$$\lambda_{\phi} = \sum_{i=1}^n \lambda_{\phi \mathbf{r}_i} + g_{\phi}, \text{ 其中 } g_{\phi} \sim N(0, \sigma_2^2), \phi \in \Phi_s, \sigma_2 = \sqrt{2 \ln(1.25/\delta_2)} \cdot \Delta_2 f / \epsilon_2;$$

4. 计算扰动后的目标函数

$$\tilde{f}_{\tilde{\mathbf{D}}_L}(\omega) = \sum_{s=1}^S \sum_{\phi \in \Phi_s} \lambda_{\phi} \phi(\omega);$$

5. 计算最优模型参数 $\tilde{\omega}^* = \arg \min_{\omega} \tilde{f}_{\tilde{\mathbf{D}}_L}(\omega)$ 。

基于上述分析,本文提出的 PCAFM-DPLR 算法的完整执行过程如算法 3 所示。

算法 3 PCAFM-DPLR 算法

输入:数据集 $\mathbf{D} = \{\mathbf{X}, \mathbf{y}\}$, 隐私参数 $\epsilon, \delta > 0$, 目标阈值 θ

输出:最优模型参数 $\tilde{\omega}^*$

1. 设置 $(\epsilon_1, \delta_1), (\epsilon_2, \delta_2)$, 使其满足 $\epsilon = \epsilon_1 + \epsilon_2, \delta = \delta_1 + \delta_2$;
2. DP-PCA($\mathbf{D}, \epsilon_1, \delta_1, \theta$) $\rightarrow \tilde{\mathbf{D}}_L$; /* 调用 DP-PCA 算法, 得到降维后的数据集 $\tilde{\mathbf{D}}_L$ */
3. FM-DPLR($\tilde{\mathbf{D}}_L, \epsilon_2, \delta_2$) $\rightarrow \tilde{\omega}^*$. /* 调用 FM-DPLR 算法, 得到最优模型参数 $\tilde{\omega}^*$ */

4 隐私性分析

本节将基于下述定理证明 PCAFM-DPLR 算法满足 (ϵ, δ) -差分隐私。

引理 1^[25] 设 $\mathbf{X}$ 与 $\mathbf{X}'$ 为相邻数据集, 只有一条数据样本不同, 设 $f(\mathbf{X})$ 和 $f(\mathbf{X}')$ 是 $\mathbf{X}$ 与 $\mathbf{X}'$ 的协方差矩阵, 则协方差矩阵 $f(\mathbf{X}) = \frac{1}{n} \mathbf{X}^T \mathbf{X}$ 的敏感度为:

$$\Delta_2 f = \max_{\mathbf{X}, \mathbf{X}'} \left\| \frac{1}{n} \mathbf{X}^T \mathbf{X} - \frac{1}{n} \mathbf{X}'^T \mathbf{X}' \right\|_2 \leq \frac{1}{n}$$

定理 1 算法 1 中添加高斯噪声扰动后的协方差矩阵 $\tilde{\mathbf{A}}$ 满足 (ϵ_1, δ_1) -差分隐私。

证明: 假设相邻数据集 $\mathbf{X}, \mathbf{X}' \in \mathbf{D}$, 只有最后一条数据样本 $\mathbf{x}_n$ 与 $\mathbf{x}_n'$ 不同, $\mathbf{E} = [e_{ij}] \in \mathbb{R}^{d \times d}$ 为高斯噪声矩阵, 其中所有 e_{ij} 独立同服从正态分布 $N(0, \sigma_1^2)$ 。记添加扰动后的协方差矩阵 $\tilde{\mathbf{A}} = M(\mathbf{X}, f, \sigma_1) = f(\mathbf{X}) + \mathbf{E}$, 输出 $\tilde{\mathbf{A}}$ 恰好为 $\mathbf{O} = [o_{ij}] \in \mathbb{R}^{d \times d}$ 时的隐私损失定义为:

$$\mathcal{L}(\mathbf{O}; M, \mathbf{X}, \mathbf{X}') = \ln \frac{\Pr[M(\mathbf{X}, f, \sigma_1) = \mathbf{O}]}{\Pr[M(\mathbf{X}', f, \sigma_1) = \mathbf{O}]}$$

又记 $\mathbf{V} = [v_{ij}] = f(\mathbf{X}) - f(\mathbf{X}') \in \mathbb{R}^{d \times d}$, 则有:

$$\begin{aligned} & |\mathcal{L}(\mathbf{O}; M, \mathbf{X}, \mathbf{X}')| \\ &= \left| \ln \frac{\Pr[f(\mathbf{X}) + \mathbf{E} = \mathbf{O}]}{\Pr[f(\mathbf{X}') + \mathbf{E}' = \mathbf{O}]} \right| \\ &= \left| \ln \frac{\Pr[\mathbf{O} - f(\mathbf{X}) = \mathbf{E}]}{\Pr[\mathbf{O} - f(\mathbf{X}') = \mathbf{E}']} \right| \\ &= \left| \ln \frac{\prod_{1 \leq i, j \leq d} \exp\left(-\frac{1}{2\sigma_1^2} (o_{ij} - f(\mathbf{X})_{ij})^2\right)}{\prod_{1 \leq i, j \leq d} \exp\left(-\frac{1}{2\sigma_1^2} (o_{ij} - f(\mathbf{X})_{ij} + v_{ij})^2\right)} \right| \\ &= \frac{1}{2\sigma_1^2} \left| \sum_{1 \leq i, j \leq d} ((o_{ij} - f(\mathbf{X})_{ij})^2 - (o_{ij} - f(\mathbf{X})_{ij} + v_{ij})^2) \right| \end{aligned}$$

由于 $e_{ij} \sim N(0, \sigma_1^2)$, 因此 $o_{ij} - f(\mathbf{X})_{ij} \sim N(0, \sigma_1^2)$ 。将所有 $o_{ij} - f(\mathbf{X})_{ij}, 1 \leq i, j \leq d$ 按照角标 i, j 的字典序排列成为一个 d^2 维度的向量 $\mathbf{z} = (o_{11} - f(\mathbf{X})_{11}, o_{12} - f(\mathbf{X})_{12}, \dots, o_{dd} - f(\mathbf{X})_{dd})$, 其分布恰为各方向球形对称的 d^2 维正态

分布。记 $\mathbf{v}=(v_{11}, v_{12}, \dots, v_{dd})$ 为角标按照同样方式排列得到的 d^2 维实数向量。可注意到 d^2 维空间中各方向球形对称的正态分布在任意的标准正交坐标系下的分布函数都是一样的,因此不妨选用一组特别的标准正交基 $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_{d^2}$, 使得其中的基向量 $\mathbf{b}_1$ 与向量 $\mathbf{v}$ 是同方向的。在上述正交基下,向量 $\mathbf{z}$ 可以表示为 $\mathbf{z}=\sum_{i=1}^{d^2} \mathbf{z}_i=\sum_{i=1}^{d^2} \lambda_i \cdot \mathbf{b}_i$, 其中 $\mathbf{z}_i=\lambda_i \cdot \mathbf{b}_i$ 表示向量 $\mathbf{z}$ 在基向量 $\mathbf{b}_i$ 上的分量, $|\lambda_i|$ 为分量的模长且 $\lambda_i \sim N(0, \sigma_1^2)$ 。此时有:

$$\begin{aligned} \sum_{1 \leq i, j \leq d} (o_{ij} - f(\mathbf{X})_{ij})^2 &= \|\mathbf{z}\|^2 = \sum_{i=1}^{d^2} \|\mathbf{z}_i\|^2 \\ \sum_{1 \leq i, j \leq d} (o_{ij} - f(\mathbf{X})_{ij} + v_{ij})^2 &= \|\mathbf{z} + \mathbf{v}\|^2 \\ &= \|\mathbf{z}_1 + \mathbf{v}\|^2 + \sum_{i=2}^{d^2} \|\mathbf{z}_i\|^2 \end{aligned}$$

注意到 $\|\mathbf{z}_1 + \mathbf{v}\|^2 = (\lambda_1 + \|\mathbf{v}\|)^2$, 进一步可推导出:

$$\begin{aligned} |\mathcal{L}(\mathbf{O}; \mathbf{M}, \mathbf{X}, \mathbf{X}')| &= \frac{1}{2\sigma_1^2} \left| \sum_{i=1}^{d^2} \|\mathbf{z}_i\|^2 - (\|\mathbf{z}_1 + \mathbf{v}\|^2 + \sum_{i=2}^{d^2} \|\mathbf{z}_i\|^2) \right| \\ &= \frac{1}{2\sigma_1^2} |2\|\mathbf{z}_1\| \cdot \|\mathbf{v}\| + \|\mathbf{v}\|^2| \\ &= \frac{1}{2\sigma_1^2} |2\lambda_1 \cdot \|\mathbf{v}\| + \|\mathbf{v}\|^2| \end{aligned}$$

显然, 当 $|\lambda_1| < \sigma_1^2 \epsilon / \Delta_2 f - \Delta_2 f / 2$ 时, $|\mathcal{L}(\mathbf{O}; \mathbf{M}, \mathbf{X}, \mathbf{X}')| \leq \epsilon_1$ 。因此, 为了保证 $\tilde{\mathbf{A}}$ 满足 (ϵ_1, δ_1) -差分隐私, 只需使得:

$$Pr[|\lambda_1| < \sigma_1^2 \epsilon / \Delta_2 f - \Delta_2 f / 2] > 1 - \delta_1$$

成立即可, 这等价于要求:

$$Pr[\lambda_1 \geq \sigma_1^2 \epsilon / \Delta_2 f - \Delta_2 f / 2] < \frac{\delta_1}{2}$$

由文献[20]可知, 正态分布 $\xi \sim N(0, \sigma_1^2)$ 的尾部概率满足不等式:

$$Pr[\xi > t] \leq \frac{\sigma_1}{\sqrt{2\pi} \cdot t} e^{-\frac{t^2}{2\sigma_1^2}}$$

$$\text{故当 } \ln(t/\sigma_1) + t^2/2\sigma_1^2 > \ln\left(\sqrt{\frac{2}{\pi}} \cdot \frac{1}{\delta_1}\right) \text{ 时,}$$

$$Pr[\xi > t] \leq \frac{\sigma_1}{\sqrt{2\pi} \cdot t} e^{-\frac{t^2}{2\sigma_1^2}} < \frac{\delta_1}{2}$$

由于 $\lambda_1 \sim N(0, \sigma_1^2)$ 服从正态分布, 因此可知当式(11)成立时, 不等式(10)成立。另又根据文献[25]可知, 当 $\sigma_1 = \sqrt{2\ln(1.25/\delta_1)}/n\epsilon_1$ 时, 上述不等式(11)成立, 因此得证 $\tilde{\mathbf{A}}$ 满足 (ϵ_1, δ_1) -差分隐私, 即定理 1 得证。

$$\ln\left(\frac{\sigma_1 \epsilon_1}{\Delta_2 f} - \frac{\Delta_2 f}{2\sigma_1}\right) + \left(\frac{\sigma_1 \epsilon_1}{\Delta_2 f} - \frac{\Delta_2 f}{2\sigma_1}\right)^2 / 2 > \ln\left(\sqrt{\frac{2}{\pi}} \cdot \frac{1}{\delta_1}\right) \quad (11)$$

引理 2 设 $\tilde{\mathbf{D}}_L$ 与 $\tilde{\mathbf{D}}_L'$ 为相邻数据集, 只有一条数据样本不同, 设 $f_{\tilde{\mathbf{D}}_L}(\boldsymbol{\omega})$ 和 $f_{\tilde{\mathbf{D}}_L'}(\boldsymbol{\omega})$ 是 $\tilde{\mathbf{D}}_L$ 与 $\tilde{\mathbf{D}}_L'$ 的目标函数, 则目标函数展开多项式系数的敏感度 $\Delta_2 f = 2\sqrt{k^2 + 4k + 1}$ 。

证明: 假设相邻数据集 $\tilde{\mathbf{D}}_L$ 与 $\tilde{\mathbf{D}}_L'$, 只有最后一条数据样本 $\mathbf{r}_n$ 与 $\mathbf{r}_n'$ 不同。由于数据集 $\tilde{\mathbf{D}}_L$ 的目标函数的展开多项式形式为:

$$\begin{aligned} f_{\tilde{\mathbf{D}}_L}(\boldsymbol{\omega}) &= \sum_{\mathbf{r}_i \in \tilde{\mathbf{D}}_L} (y_i - \tilde{\mathbf{z}}_i^T \boldsymbol{\omega})^2 \\ &= \sum_{1 \leq j, l \leq k} \sum_{\mathbf{r}_i \in \tilde{\mathbf{D}}_L} (\tilde{z}_{ij} \tilde{z}_{il}) \omega_j \omega_l - \sum_{j=1}^k (2 \sum_{\mathbf{r}_i \in \tilde{\mathbf{D}}_L} y_i \tilde{z}_{ij}) \omega_j + \sum_{\mathbf{r}_i \in \tilde{\mathbf{D}}_L} (y_i)^2 \end{aligned}$$

因此, 线性回归的目标函数的展开多项式是由 3 个单项式 $\Phi_0 = \{1\}$, $\Phi_1 = \{\omega_j | 1 \leq j \leq k\}$, $\Phi_2 = \{\omega_j \omega_l | 1 \leq j, l \leq k\}$ 组成的。且 ω_j 的系数是 $-2y_i \tilde{z}_{ij}$, $\omega_j \omega_l$ 的系数是 $\tilde{z}_{ij} \tilde{z}_{il}$ 。由于 $\tilde{\mathbf{D}}_L = \{(\tilde{\mathbf{z}}_1, y_1), (\tilde{\mathbf{z}}_2, y_2), \dots, (\tilde{\mathbf{z}}_n, y_n)\}$, 其中 $\tilde{\mathbf{Z}} \in \mathbb{R}^{n \times k}$, $\|\tilde{\mathbf{z}}_i\|_2 \leq 1, y_i \in [-1, 1]$, 因此:

$$\begin{aligned} \Delta_2 f &= 2 \max_{\mathbf{r}=(\tilde{\mathbf{z}}, y)} \sum_{s=1}^S \sum_{\phi \in \Phi_s} \|\lambda_{\phi r}\|_2 \\ &\leq 2 \max_{\mathbf{r}=(\tilde{\mathbf{z}}, y)} \sqrt{\sum_{1 \leq j, l \leq k} (\tilde{z}_j \tilde{z}_l)^2 + \sum_{j=1}^k (-2y \tilde{z}_j)^2 + y^2} \\ &\leq 2\sqrt{k^2 + 4k + 1} \end{aligned}$$

定理 2 算法 2 中扰动后的目标函数 $\tilde{f}_{\tilde{\mathbf{D}}_L}(\boldsymbol{\omega})$ 满足 (ϵ_2, δ_2) -差分隐私。

证明: 设相邻数据集 $\tilde{\mathbf{D}}_L, \tilde{\mathbf{D}}_L' \in \mathcal{D}$, 只有最后一条数据样本 $\mathbf{r}_h$ 与 $\mathbf{r}_h'$ 不同, 则:

$$\begin{aligned} &\left| \ln \frac{Pr[\tilde{f}_{\tilde{\mathbf{D}}_L}(\boldsymbol{\omega})]}{Pr[\tilde{f}_{\tilde{\mathbf{D}}_L'}(\boldsymbol{\omega})]} \right| \\ &= \left| \ln \frac{\prod_{s=1}^S \prod_{\phi \in \Phi_s} \exp\left(-\frac{1}{2\sigma_2^2} (\sum_{i=1}^n \lambda_{\phi r_i} - \lambda_{\phi})^2\right)}{\prod_{s=1}^S \prod_{\phi \in \Phi_s} \exp\left(-\frac{1}{2\sigma_2^2} (\sum_{i=1}^n \lambda_{\phi r_i'} - \lambda_{\phi})^2\right)} \right| \\ &= \frac{1}{2\sigma_2^2} \left| \sum_{s=1}^S \sum_{\phi \in \Phi_s} ((\sum_{i=1}^n \lambda_{\phi r_i} - \lambda_{\phi})^2 - (\sum_{i=1}^n \lambda_{\phi r_i'} - \lambda_{\phi})^2) \right| \\ &= \frac{1}{2\sigma_2^2} |\|\mathbf{g}\|^2 - \|\mathbf{g} + \mathbf{u}\|^2| \end{aligned}$$

其中, $\mathbf{g} = \{\sum_{i=1}^n \lambda_{\phi r_i} - \lambda_{\phi}\}_{\phi \in \bigcup_{s=1}^S \Phi_s}$, $\mathbf{u} = \{\sum_{i=1}^n \lambda_{\phi r_i'} - \sum_{i=1}^n \lambda_{\phi r_i} - \lambda_{\phi}\}_{\phi \in \bigcup_{s=1}^S \Phi_s}$ 。

由于球形对称正态分布不依赖于自变量空间上的正交基, 因此我们选用一组特别的标准正交基 $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{|\Phi_s|}$, 使得其中的基向量 $\mathbf{c}_1$ 与向量 $\mathbf{u}$ 是同方向的。在上述正交基下, 向量 $\mathbf{g}$ 可以表示为 $\mathbf{g} = \sum_{\phi \in \Phi_s} g_{\phi} = \sum_{\phi \in \Phi_s} \mu_{\phi} \mathbf{c}_{\phi}$, 其中 $g_{\phi} = \mu_{\phi} \mathbf{c}_{\phi}$ 表示向量 $\mathbf{g}$ 在基向量 $\mathbf{c}_{\phi}$ 上的分量, $|\mu_{\phi}|$ 为分量的模长且满足 $\mu_{\phi} \sim N(0, \sigma_2^2)$ 。此时有: $\|\mathbf{g}\|^2 - \|\mathbf{g} + \mathbf{u}\|^2 = \|\sum_{\phi \in \Phi_s} g_{\phi}\|^2 - (\|\mathbf{g}_1 + \mathbf{u}\|^2 + \sum_{s=2}^S 2g_{\phi})$ 。又由于 $\|\mathbf{g}_1 + \mathbf{u}\|^2 = (\mu_1 + \|\mathbf{u}\|)^2$, 进一步可以推导出:

$$\begin{aligned} &\left| \ln \frac{Pr[\tilde{f}_{\tilde{\mathbf{D}}_L}(\boldsymbol{\omega})]}{Pr[\tilde{f}_{\tilde{\mathbf{D}}_L'}(\boldsymbol{\omega})]} \right| = \frac{1}{2\sigma_2^2} |\|\mathbf{g}\|^2 - \|\mathbf{g} + \mathbf{u}\|^2| \\ &= \frac{1}{2\sigma_2^2} |\|\sum_{\phi \in \Phi_s} g_{\phi}\|^2 - (\|\mathbf{g}_1 + \mathbf{u}\|^2 + \sum_{s=2}^S 2g_{\phi})| \\ &= \frac{1}{2\sigma_2^2} |2\mu_1 \|\mathbf{u}\| + \|\mathbf{u}\|^2| \end{aligned}$$

显然, 当 $|\mu_1| < \sigma_2^2 \epsilon_2 / \Delta_2 f - \Delta_2 f / 2$ 时,

$\left| \ln \frac{Pr[\tilde{f}_{\tilde{\mathbf{D}}_L}(\boldsymbol{\omega})]}{Pr[\tilde{f}_{\tilde{\mathbf{D}}_L'}(\boldsymbol{\omega})]} \right| \leq \epsilon_2$ 。因此, 为了保证扰动后的展开式系数 $\tilde{\lambda}_{\phi r}$ 满足 (ϵ_2, δ_2) -差分隐私, 只需使得

$Pr[|\mu_1| < \sigma_2^2 \epsilon_2 / \Delta_2 f - \Delta_2 f / 2] > 1 - \delta_2$
成立即可, 这等价于要求:

$$Pr[\mu_1 \geq \sigma_2^2 \epsilon_2 / \Delta_2 f - \Delta_2 f / 2] < \frac{\delta_2}{2} \tag{12}$$

同理, 由文献[20]可知, 正态分布 $\xi \sim N(0, \sigma_2^2)$ 的尾部概率满足不等式:

$$Pr[\xi > t] \leq \frac{\sigma_2}{\sqrt{2\pi} \cdot t} e^{\frac{-t^2}{2\sigma_2^2}}$$

故当 $\ln(t/\sigma_2) + t^2/2\sigma_2^2 > \ln\left(\sqrt{\frac{2}{\pi}} \cdot \frac{1}{\delta_2}\right)$ 时, 有:

$$Pr[\xi > t] \leq \frac{\sigma_2}{\sqrt{2\pi} \cdot t} e^{\frac{-t^2}{2\sigma_2^2}}$$

由于 $\mu_1 \sim N(0, \sigma_2^2)$ 服从正态分布, 因此可知当式(13)成立时, 不等式(12)成立。将 $\Delta_2 f = 2 \max_{r=1}^S \sum_{\phi \in \Phi_r} \|\lambda_{\phi r}\|_2 \cdot \sigma_2 = \sqrt{2\ln(1.25/\delta_2)} \cdot \Delta_2 f / \epsilon_2$ 代入可得上述不等式(13)成立, 因此得证扰动后的展开式系数 $\tilde{\lambda}_{\phi r}$ 满足 (ϵ_2, δ_2) -差分隐私。

$$\ln\left(\frac{\sigma_2 \epsilon_2}{\Delta_2 f} - \frac{\Delta_2 f}{2\sigma_2}\right) + \left(\frac{\sigma_2 \epsilon_2}{\Delta_2 f} - \frac{\Delta_2 f}{2\sigma_2}\right)^2 / 2 > \ln\left(\sqrt{\frac{2}{\pi}} \cdot \frac{1}{\delta_2}\right) \tag{13}$$

又因为 $\tilde{f}_{\tilde{D}_L}(\omega) = \sum_{s=1}^2 \sum_{\phi(\omega) \in \Phi_s} \tilde{\lambda}_{\phi r} \phi(\omega)$, 所以扰动后的目标函数 $\tilde{f}_{\tilde{D}_L}(\omega)$ 满足 (ϵ_2, δ_2) -差分隐私, 即定理 2 得证。

定理 3 PCAFM-DPLR 算法满足 (ϵ, δ) -差分隐私。

证明: 算法 1 和算法 2 分别满足 (ϵ_1, δ_1) -差分隐私和 (ϵ_2, δ_2) -差分隐私, 且在算法 3 中设置 $(\epsilon_1, \delta_1), (\epsilon_2, \delta_2)$ 使其满足 $\epsilon = \epsilon_1 + \epsilon_2, \delta = \delta_1 + \delta_2$ 。由差分隐私的序列组合性可知, 算法 3 最终输出的最优模型参数 $\tilde{\omega}^*$ 满足 $(\sum_{i=1}^2 \epsilon_i, \sum_{i=1}^2 \delta_i)$ -差分隐私, 即 PCAFM-DPLR 算法满足 (ϵ, δ) -差分隐私。

5 实验设计及分析

5.1 实验数据与环境

本文仿真实验的环境为 Intel(R) Core(TM) i7-4770 CPU 3.40 GHz, 16 GB 内存, Windows10 64 位操作系统。选用 Matlab 2018b 模拟实现基于主成分分析和函数机制的差分隐私线性回归算法。

为了体现实验的可靠性和科学性, 本文采用 FM 算法^[16]中选用的数据集 US 和 Brazil, 以及 Kaggle 数据库中用于训练线性回归模型的公开数据集 Boom Bikes 来进行仿真分析。数据集的具体信息如表 2 所列。其中 US 和 Brazil 数据集分别包括美国和巴西 370 000 条和 190 000 条的人口普查记录, 用于预测一个美国人和一个巴西人的年收入。Boom Bikes 数据集的目的是预测 Covid-19 结束后人们对共享单车的需求。

表 2 数据集信息

Table 2 Dataset information

数据集	样本数/ <i>n</i>	特征数/ <i>d</i>
US	370 000	14
Brazil	190 000	14
Boom Bikes	730	31

本文首先对数据集中的每个特征属性值进行归一化处理, 使其取值范围为 $[-1, 1]$ 。然后把归一化后的数据集随机分成两部分, 其中随机选取总样本数的 80% 作为训练集, 将剩余的 20% 作为测试集。最后分别对 3 个数据集进行线性回归分析。

为了更好地评估 PCAFM-DPLR 算法的有效性, 本文与 FM 算法^[16]、NoPrivacy 两种方案进行对比。

- (1) FM: 向线性回归模型目标函数的展开多项式系数中添加满足 ϵ -差分隐私的拉普拉斯噪声, 进而求得最优模型参数。
- (2) NoPrivacy: 不添加任何隐私保护机制, 直接最小化线性回归模型的目标函数, 求得最优模型参数。

考虑到噪声的随机性, 本文将每组实验进行 500 次, 取其平均值作为实验的仿真结果。本文使用均方误差 (MSE) 作为模型可用性的评价指标, MSE 越小, 模型的可用性就越高。

5.2 实验结果分析

5.2.1 主成分分析

图 2 为本文实验所用数据集经特征分析后得到的碎石图。其中, 横轴表示特征值由大到小排列后各个主成分的序号, 纵轴表示特征值的大小。

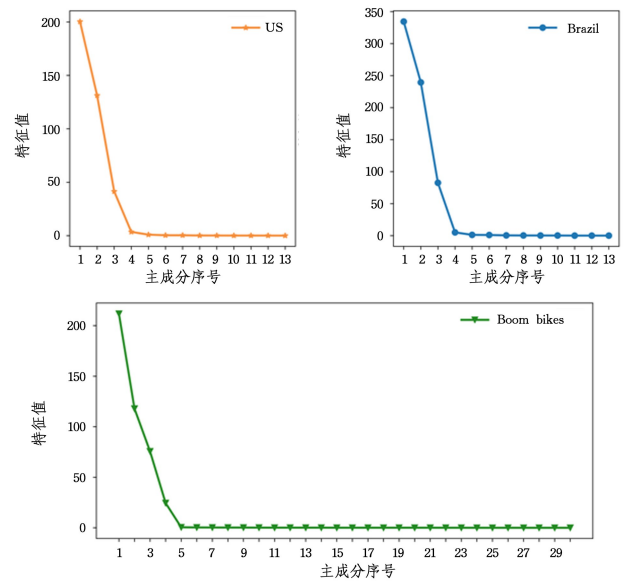


图 2 数据集的碎石图
Fig. 2 Gravel plot of datasets

通过图 2 所示的数据可知, 3 个数据集的碎石图形状均类似山崖, 其中 US 和 Brazil 数据集的曲线是从第 1 个主成分开始迅速下降, 在第 4 个主成分处曲线变化转至平缓, 随后曲线变为一条趋近于 0 的直线。Boom Bikes 数据集的曲线则是在第 5 个主成分处变化转至平缓, 随后变为一条趋近于 0 的直线。通过分析碎石图的变化趋势, 本文选择保留曲线变化趋势开始转至平缓的点之前的所有主成分。

为具体显示主成分的贡献率, 本文在表 3 中列出了 3 个数据集经 PCA 算法处理后的前 8 个主成分的特征值占比。由表 3 所列的数据可知, 第 1 主成分几乎都承载了原始数据 50% 的信息。具体而言, US 和 Brazil 数据集的前 3 个主成分分别承载了原始数据集 98.43% 和 98.76% 的信息, Boom Bikes 数据集的前 3 个主成分承载了其原始数据集 93.62%

的信息,前4个主成分承载了其原始数据集99.28%的信息。实验结果表明,对于US和Brazil数据集,想要保留原始数据集95%的信息,仅需要3个主成分。对于Boom Bikes数据集

而言,仅需要4个主成分,便能将原始的31维数据集中95%以上的信息保留下来。由此可见,这3个数据集均能在有效降维的同时承载原始数据集大量的信息。

表 3 数据集的主成分分析特征值及累积贡献率
Table 3 Eigenvalues and cumulative contribution of PCA of datasets

主成分	US			Brazil			Boom Bikes		
	特征值	贡献率/%	累积贡献率/%	特征值	贡献率/%	累积贡献率/%	特征值	贡献率/%	累积贡献率/%
1	200.3571	52.87	52.87	334.4171	50.32	50.32	211.7594	48.92	48.92
2	131.2324	34.63	87.50	239.4167	36.02	86.34	117.8674	27.23	76.15
3	41.4349	10.93	98.43	82.5491	12.42	98.76	75.6123	17.47	93.62
4	3.5483	0.25	98.68	5.0723	0.76	99.52	24.5010	5.66	99.28
5	0.9546	0.10	98.78	1.1716	0.18	99.70	0.4982	0.12	99.40
6	0.3776	0.09	98.87	0.9660	0.15	99.85	0.3443	0.08	99.48
7	0.3540	0.04	98.93	0.3669	0.06	99.91	0.2934	0.07	99.55
8	0.1637	0.04	98.97	0.2401	0.04	99.95	0.2491	0.06	99.61

5.2.2 不同隐私预算下模型的可用性

由于隐私预算 ϵ 决定了差分隐私的保护程度,因此本文主要衡量 ϵ 的取值范围在 $(0,1)$ 时,模型的可用性问题。即在保证模型隐私性的基础上,分析模型的可用性问题。在采用PCAFM-DPLR算法训练模型时,本文固定 δ 的值为 10^{-5} ,合理分配隐私预算 ϵ_1, ϵ_2 ,使 $\epsilon = \epsilon_1 + \epsilon_2$,并且使用PCA技术保留原始数据集95%以上的信息,分别在3个数据集上进行实验。

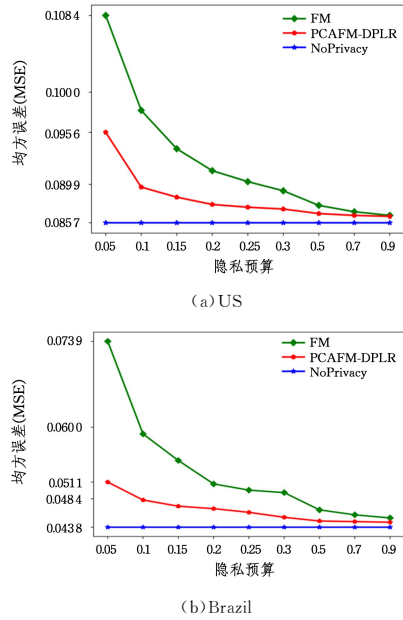


图 3 不同隐私预算下 US 和 Brazil 数据集的 MSE 对比
Fig. 3 MSE comparison of US and Brazil datasets with different privacy budgets

为具体对比不同隐私预算对线性回归模型预测准确性的影响,本文将 ϵ 分别设置为0.05,0.1,0.15,0.2,0.25,0.3,0.5,0.7和0.9,计算 ϵ 取不同值时MSE的大小。

针对US和Brazil数据集,不同隐私预算下FM算法和PCAFM-DPLR算法对实验结果产生的影响如图3所示。可以看出,当 ϵ 相同时,采用PCAFM-DPLR算法训练出的线性回归模型的MSE低于FM算法,模型的可用性更高。并且,随着隐私预算 ϵ 的不断增大,PCAFM-DPLR算法的MSE不断接近无任何隐私保护机制的NoPrivacy方案所产生的

误差。因此,在保证模型隐私性的同时,能够通过适当增加隐私预算的方式来提高模型的可用性,从而更好地实现模型的隐私性与可用性的平衡。

针对Boom Bikes数据集,由于FM算法和PCAFM-DPLR算法的实验结果相差较大,因此本文在表4中列出了不同隐私预算下3种方案在Boom Bikes数据集上的MSE对比。由表4所列的数据可知,在相同的隐私预算下,PCAFM-DPLR算法的预测准确性远远高于FM算法,并且随着隐私预算的增大,PCAFM-DPLR算法的均方误差逐渐减小。

表 4 不同隐私预算下 3 种算法在 Boom Bikes 数据集上的 MSE 对比
Table 4 MSE comparison of three algorithms on Boom Bikes dataset with different privacy budgets

ϵ	MSE		
	FM	PCAFM-DPLR	NoPrivacy
0.05	935.0438	0.2024	0.0339
0.10	881.2532	0.1964	0.0339
0.15	555.9116	0.1932	0.0339
0.20	468.7536	0.1897	0.0339
0.25	364.3762	0.1861	0.0339
0.30	273.5945	0.1814	0.0339
0.50	140.8523	0.1651	0.0339
0.70	80.4989	0.1509	0.0339
0.90	52.4557	0.1474	0.0339

为了直观地显示其差异,本文借鉴文献[26]的做法,对折线图数值进行了对数化处理。图4中的虚线代表模型的均方误差为1时的情况,可以看出,FM算法已远远超出虚线,而PCAFM-DPLR算法的均方误差均在虚线之下。

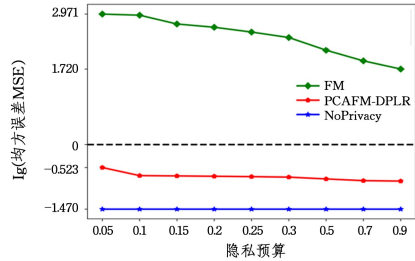


图 4 不同隐私预算下 3 种算法在 Boom Bikes 数据集上的 lg(MSE) 对比
Fig. 4 Comparison of lg(MSE) of three algorithms on Boom Bikes dataset with different privacy budgets

仿真结果表明,由 FM 算法训练出的线性回归模型的均方误差过大,模型可用性较差。而由 PCAFM-DPLR 算法训练出的模型,虽然模型预测精度有所损失,但是相比 FM 算法,仍然具有较好的可用性。其主要原因在于:PCAFM-DPLR 算法首先利用 PCA 技术对原始数据集进行降维,仅前 4 个主成分便能够承载原始数据集 95% 以上的信息。因此,在对目标函数的展开多项式系数加噪时,相比原始的 31 维,PCAFM-DPLR 算法能够显著降低其全局敏感度,从而减少噪声的注入。

具体来说,FM 算法为目标函数的展开多项式系数添加的是满足 ϵ -差分隐私保护的 Laplace 噪声,此时目标函数展开式的全局敏感度为 $2(d^2 + 2d + 1)$ 。而 PCAFM-DPLR 算法为目标函数的展开多项式系数添加的是满足 (ϵ, δ) -差分隐私保护的高斯噪声,并且将隐私预算分解为两部分,一部分用于 PCA 降维过程,对协方差矩阵加噪,此时全局敏感度为 $1/n$,另一部分用于对目标函数的展开多项式系数加噪,此时全局敏感度为 $2\sqrt{k^2 + 4k + 1}$ 。针对 PCAFM-DPLR 算法,本文在保证总隐私预算不变的情况下,给全局敏感度较大的展开多项式系数分配较大的隐私预算,给全局敏感度较小的协方差矩阵分配较小的隐私预算,从而使得总体添加的噪声量更少,模型的预测准确性更高。

结束语 本文针对现有的差分隐私线性回归算法存在的全局敏感度大及模型可用性较差的问题,提出了一种基于主成分分析和函数机制的差分隐私线性回归算法。该算法向原始数据集的协方差矩阵及目标函数的展开多项式系数中注入高斯噪声,在保证模型隐私性的同时,提高了模型的可用性。并从差分隐私定义的角度严格证明了文中所提的 PCAFM-DPLR 算法满足 (ϵ, δ) -差分隐私。实验结果表明,该算法能够有效降低全局敏感度、提高模型的可用性。未来工作将考虑如何根据数据集属性的重要程度来自适应分配隐私预算,以获得更高的预测准确性。

参 考 文 献

[1] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]// Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2015:1322-1333.

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

[3] DWORK C. Differential privacy: A survey of results[C]// Proceedings of the 5th International Conference on Theory and Applications of Models of Computation. Berlin: Springer, 2008: 1-19.

[4] ZHANG Y X, WEI J H, LI J, et al. Graph degree histogram publication method with node-differential privacy[J]. Journal of Computer Research and Development, 2019, 56(3): 508-520.

[5] CORMODE G, PROCOPIUC M, SHEN E, et al. Differentially private spatial decompositions[C]// Proceedings of the IEEE 28th International Conference on Data Engineering. Piscataway, NJ: IEEE, 2012: 20-31.

[6] BHASKAR R, LAXMAN S, SIMTH A, et al. Discovering frequent patterns in sensitive data[C]// Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2010: 503-512.

[7] MOHAMMED N, CHEN R, FUNG B, et al. Differentially private data release for data mining[C]// Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2011: 493-501.

[8] NISSIM K, RASKHODNIKOVA S, SMITH A. Smooth sensitivity and sampling in private data analysis[C]// Proceedings of the 39th Annual ACM Symposium on Theory of Computing. New York: ACM, 2007: 11-13.

[9] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]// Proceedings of the 3rd Theory of Cryptography Conference. Berlin: Springer, 2006: 265-284.

[10] FREDRIKSON M, LANTZ E, JHA S, et al. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing[C]// Proceedings of the USENIX Security Symposium. Berkeley, CA: USENIX, 2014: 17-32.

[11] WU X, FREDRIKSON M, WU W T, et al. Revisiting differentially private regression: Lessons from learning theory and their consequences[J]. arXiv:1512.06388, 2015.

[12] SMITH A. Privacy-preserving statistical estimation with optimal convergence rate[C]// Proceedings of the 43th Annual ACM Symposium on Theory of Computing. New York: ACM, 2011: 813-822.

[13] LEI J. Differentially private m-estimators[C]// Proceedings of the 23rd Annual Conference on Neural Information Processing Systems. New York: ACM, 2011: 361-369.

[14] GONG M G, PAN K, XIE Y. Differential privacy preservation in regression analysis based on relevance[J]. Knowledge-Based Systems, 2019, 173(6): 140-149.

[15] CHAUDHURI K, MONTELEONI C, SARWATE D. Differentially private empirical risk minimization[J]. The Journal of Machine Learning Research, 2011, 12(2): 1069-1109.

[16] ZHANG J, ZHANG Z J, XIAO X K, et al. Functional mechanism: regression analysis under differential privacy[J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1364-1375.

[17] DWORK C, KENTHAPADI K, MCSHERRY F, et al. Our data, ourselves: privacy via distributed noise generation[C]// Proceedings of the 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques. Berlin: Springer, 2006: 486-503.

[18] JAYARAMAN B, EVANS D. Evaluating differentially private machine learning in practice[C]// Proceedings of the 28th USE-

NIX Security Symposium. Berkeley, CA: USENIX, 2019: 1895-1912.

[19] LIU J X, MENG X F. Survey on Privacy-Preserving Machine Learning[J]. Journal of Computer Research and Development, 2020, 57(2): 346-362.

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

[21] MCSHERRY F, TALWAR K. Mechanism design via differential privacy[C]//Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science. Piscataway, NJ: IEEE, 2007: 94-103.

[22] XIONG P, ZHU T Q, WANG X F. A survey on differential privacy and applications[J]. Chinese Journal of Computers, 2014, 37(1): 101-122.

[23] MCSHERRY F. Privacy integrated queries: an extensible platform for privacy-preserving data analysis[J]. Communications of the ACM, 2010, 53(9): 89-97.

[24] PEARSON K. On lines and planes of closest fit to systems of points in space[J]. Philosophical Magazine, 1901, 2(11): 559-572.

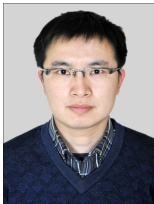
[25] DWORK C, TALWAR K, THAKURTA A, et al. Analyze

gauss: optimal bounds for privacy-preserving principal component analysis[C]//Proceedings of the 46th Annual ACM Symposium on Theory of Computing. New York: ACM, 2014: 11-20.

[26] PENG C G, ZHAO Y Y, FAN M M. A differential private data publishing algorithm via principal component analysis based on maximum information coefficient[J]. Netinfo Security, 2020, 20(2): 37-48.



LI Kejia, born in 1997, postgraduate. Her main research interests include federated learning, differential privacy, and blockchain.



HU Xuexian, born in 1982, Ph.D, associate professor, Ph.D supervisor. His main research interests include big data security, applied cryptography, and network security.

(责任编辑:喻黎)