多项式回归的差分隐私保护算法

谢雅琪[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:多项式回归是用来确定两种或两种以上变量间相互依赖的非线性定量关系的一种统计分析方法,在大数据分析中有广泛的应用。通常,挖掘的数据集包含一些敏感属性,在数据挖掘过程和数据发布中,如不加保护会引起隐私泄露。基于对代价函数添加噪声的方法,该文设计了一种满足差分隐私的多项式回归算法 FM-on-PR,并且针对现实应用中的需求,对该算法进行了优化,获得了两种分别对数据安全性和数据可用性进行加强的算法 DPC-on-PR 和 DPBA-on-PR。通过理论证明了它们满足差分隐私性质,并使用多个数据集进行实验仿真,测试算法性能,结果表明了这些方法具有有效性,并且经过对比,得出了其中拟合优度最高的 DPBA-on-PR 算法。

关键词:机器学习;差分隐私;多项式回归;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(2022)08-0103-07
doi:10.3969/j.issn.1673-629X.2022.08.017

Differential Privacy Preservation in Polynomial Regression Analysis

XIE Ya-qi[1], 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: Polynomial regression is used to find out the interdependent nonlinear quantitative relationships between multiple variables in mathematical statistics, which has a wide application in big data analysis. Usually, the dataset contains some sensitive attributes, which can cause privacy leakage without preservation in the data mining and data release. Based on the method of adding noise to the cost function, we design a polynomial regression algorithm FM-on-PR to satisfy the difference privacy. According to the requirements of practical applications, such algorithm is optimized, and two algorithms DPC-on-PR and DPBA-on-PR are obtained to enhance data security and data availability respectively. They are both proven to satisfy the differential privacy property through theory. In addition, simulation is performed with several datasets on these algorithms to test their performance. Results demonstrate the effectiveness of these methods and, after comparison, show DPBA-on-PR has the best goodness of fit.

Key words: machine learning; differential privacy; polynomial regression; data privacy preservation; privacy budget allocation

0 引 言

多项式回归是利用数理统计中的回归分析,来确定两种或两种以上变量间相互依赖的非线性定量关系的一种统计分析方法,一般应用于数据属性之间的关联呈非线性的情况,拥有比线性回归更加广泛的应用场合,在数据挖掘领域中通常使用它对数据进行预测。

挖掘的数据集包含一些敏感属性,在数据挖掘过程和数据发布中,如不加保护会引起隐私泄露[1]。在回归分析中,常用的隐私保护算法有几种类型:匿名算法、数据加密和数据扰动[2]等。匿名算法中,k-匿名算法使用较为广泛;数据加密算法的研究中,常用的有同态加密[3]、安全多方计算[4]等;此外还有数据泛化[5]。

差分隐私(Differential Privacy, DP)[6]属于数据扰动算法,它以数学理论为支撑,为数据的隐私保护提供了有力手段[7]。DP 通过对数据集或者查询函数添加噪声等方式实现隐私保护。DP 最为基础的两个机制为拉普拉斯机制和指数机制[8],分别应用于查询函数的输出是数值型和非数值型的情况,该文涉及的机制都是拉普拉斯机制。

收稿日期:2021-08-06 **修回日期**:2021-12-14
基金项目:国家自然科学基金资助项目(61872197,61972209)
作者简介:谢雅琪(1997-),女,硕士研究生,通信作者,研究方向为差分隐私保护;杨 庚,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。

在面向回归分析的差分隐私算法研究中,通常对数据集或查询函数加噪声。前者如 Lei[9] 提出的 DPME 算法,后者如函数扰动机制[10]。添加噪声的过程中,可以选择是否对代价函数进行敏感度分析。不做敏感度分析的情况下,可能会引入不必要的噪声;如果进行敏感度分析,相较于前者,能够减轻噪声添加过量影响可用性的问题[11],具有代表性的算法是 Zhang 等[12]提出的 FM(Function Mechanism),现在许多研究者都在研究基于 FM 的改进算法[13-16]。

1 相关工作

2006 年,Dwork 提出了差分隐私的定义,它通过对数据集或者模型添加噪声实现隐私保护。相较于传统的隐私保护算法,差分隐私不仅拥有数学理论作为支撑,而且能够无视攻击者所拥有的知识背景,提供了有力的隐私保护。目前,差分隐私的机制仍在逐步完善中[17-18]。在差分隐私保护算法被提出后,面向回归分析领域的差分隐私算法研究在噪声添加机制方面主要分为了两类:直接对数据集添加噪声和对查询函数添加噪声。

早期面向回归分析领域的差分隐私研究集中在对数据集添加噪声方面,如 Lei 的 DPME 算法[9],但这些算法不对数据集进行分析,直接对数据添加噪声,在数据集的维度较大时,会引入大量不必要的噪声影响数据集的可用性。

2011 年,Chaudhuri 等[10]提出了一种对回归模型的代价函数的系数添加噪声的差分隐私保护算法,并且对代价函数的敏感度进行了分析,进一步减少了噪声的添加量,提高了数据的可用性。在此基础上,Zhang 等[12]提出了函数机制 FM,该算法进一步提高了训练结果的精确性。Wang 等[15]基于 FM 算法的思路,提出了 DPC 算法,主要为了应对模型逆向攻击(Model Inversion Attack)[19],提高数据的安全性。

该文的主要贡献为:设计了三个面向多项式回归的差分隐私算法,通过多组实验衡量并比较它们在数据可用性方面的性能,并且优化了算法使其能用于高维度的数据集。

2 理论基础

2.1 多项式回归分析

多项式回归分析是根据自变量与因变量的依赖关系进行建模的统计分析方法,通常在自变量和因变量呈非线性关系时使用,如图 1,是一个简单一元二次多项式回归的拟合结果,图中点对应每条数据(x,y)。

设数据集 D 包含了 n 条记录:$t_1,t_2,\cdots,t_n$,并且拥有 d 个自变量 $X_1,X_2,\cdots,X_d$ 和 1 个因变量 Y。对于每条记录有 $t_i=(x_{i1},x_{i2},\cdots,x_{in},y)$,且假设 $\sqrt{\sum_{i=1}^{d}x_{id}^2}\leqslant 1$。设 $g(x)=x_1^{c_1}\cdot x_2^{c_2}\cdot\cdots\cdot x_d^{c_d}$,其中 $c_1,c_2,\cdots,c_d\in N$,$G_j=\{x_1^{k_1}\cdot x_2^{k_2}\cdot\cdots\cdot x_d^{k_d}\mid\sum_{i=1}^{d}k_i=j\}$ 中 $0\leqslant j\leqslant m$,设 λ_g 是 $g(x)$ 的系数,则 d 维 m 阶($d>m$)多项式回归模型就可以表示为式(1)。

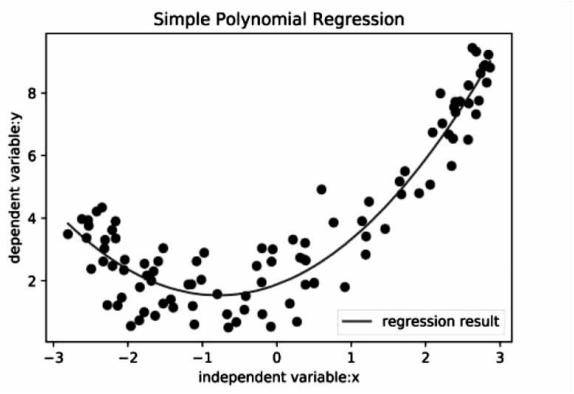


图 1 简单多项式回归的拟合结果

多项式回归的拟合可以通过变量替换,将其转换成多元线性回归处理,通过如公式(2)的映射,将每个 $g(x)$ 转化为新的自变量 z_i,就可以得到多元线性回归模型。

$$y=\sum_{j=1}^{m}\sum_{g\in G_j}\lambda_g g(x) \tag{1}$$

$$\begin{cases} z_1=x_1,z_2=x_2,\cdots,z_{C_d^1}=x_d \\ z_{C_d^1+1}=x_1^2,z_{C_d^1+2}=x_1x_2,\cdots,z_{C_d^1+C_d^2-1}=x_{d-1}x_d \\ z_{C_d^1+C_d^2}=x_d^2,z_{C_d^1+C_d^2+1}=x_1^3,\cdots,z_{\sum_{i=1}^{m}C_d^i}=x_d^m \end{cases} \tag{2}$$

经过映射(2),就可以将多项式回归转为关于 z_i 的多元线性回归,如式(3),$\vec{\omega}$ 是 z_i 系数向量,与从多项式回归中的 $g(x)$ 的系数 λ_g 相对应。

$$\vec{y}=\vec{Z}\cdot\vec{\omega} \tag{3}$$

设 $\sigma=\sum_{i=1}^{m}C_d^i$,其代价函数可以表示为如下形式:

$$f_D(\omega)=\sum_{i=1}^{n}f(t_i,\omega)=\sum_{i=1}^{n}(y_i-\vec{\omega}\vec{z_l})^2=\sum_{t_i\in D}(y_i)^2-\sum_{j=1}^{\sigma}(2\sum_{t_i\in D}y_iz_{ij})\omega_j+\sum_{1\leqslant j,l\leqslant\sigma}(\sum_{t_i\in D}Z_{ij}z_{il})\omega_j\omega_1 \tag{4}$$

同样,根据最小二乘法(OLS)求解出 $\vec{\omega}$:

$$\vec{\omega}=(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}\vec{y} \tag{5}$$

2.2 差分隐私保护技术

定义 1(差分隐私[6]):当且仅当算法 A 的输入为两个邻近数据集 D_1 和 D_2(最多只有一条记录有差别的两个数据集)且满足式(6)时,算法 A 满足 ε - 差分隐私。

$$\Pr[A(D_1) \in 0] \leqslant e^{\varepsilon} \cdot \Pr[A(D_2) \in 0] \quad (6)$$

其中,ε 是可以任意设置的隐私预算,$\Pr[A(D_1) \in 0]$ 代表输入为 D_1 时,算法 A 的输出属于集合 O 的概率。

定理 1　差分隐私的串行组合性(Sequence Composition)[30]:设有算法 $K_1, K_2, \cdots, K_n$ 和它们自己的隐私预算 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$,另有数据集 D,则算法 $K(K_1(D), K_2(D), \cdots, K_n(D))$ 满足 $\sum_{i=1}^{n} \varepsilon_i$ -差分隐私。

2.3　全局敏感度

定义 2(全局敏感度):假设有两个邻近数据集 D 和 D',那么对于一个函数 $f: D \to R^d$,它的全局敏感度定义为:

$$\Delta f = \max_{D,D'} \| f(D) - f(D') \| \quad (7)$$

3　面向多项式回归分析的差分隐私保护机制

3.1　多项式回归算法

为了防止多项式回归中常见的过拟合现象,该文采用了岭回归的思路,在代价函数中引入正则项来解决,在式(4)的基础上,加入 $\frac{\lambda}{2}\|\omega\|^2$ 项,代价函数就变成了如下形式:

$$f_D(\omega) = \sum_{i=1}^{n} f(t_i, \omega) = \sum_{i=1}^{n} (y_i - \vec{\omega}\vec{x_l})^2 + \frac{\lambda}{2}\|\omega\|^2 \quad (8)$$

再对式(8)利用最小二乘法求得 ω^*:

$$\omega^* = (X^{\mathrm{T}}X + \lambda E_{m+1})^{-1} X^{\mathrm{T}}\vec{y} \quad (9)$$

3.2　面向多项式回归分析的差分隐私保护算法

3.2.1　面向多项式回归的函数算法

面向多项式回归的函数算法(Functional Mechanism on Polynomial Regression)简称 FM-on-PR。

设 $\varphi(\omega)$ 是 $\omega_1, \omega_2, \cdots, \omega_d$ 组合的乘积,即 $\varphi(\omega) = \omega_1\omega_2\cdots\omega_d$,$c_1, c_2, \cdots, c_d \in N$。同时,定义 Φ_j 为满足 $\sum_{l=1}^{d} c_l = j$ 的所有 $\varphi(\omega)$ 的集合,即:

$$\Phi_j = \{\omega_1^{c_1}\omega_2^{c_2}\cdots\omega_d^{c_d} \mid \sum_{l=1}^{d} c_1 = j\} \quad (10)$$

FM-on-PR 对回归模型的代价函数使用多项式逼近后转化为多项式,如式(11):

$$f_D(\omega) = \sum_{i=1}^{n} f(t_i, \omega) = \sum_{j=1}^{J} \sum_{\varphi \in \Phi_j} \sum_{t_i \in D} \lambda_{\varphi_i} \varphi(\omega) \quad (11)$$

根据式(4)和式(11),可以将 $f_D(\omega)$ 简化为 $f_D(\omega) = a\omega^2 + b\omega + c$,对系数 a, b, c 各添加上服从 Lap($\Delta f/\epsilon$) 分布的噪声,如式(12)所示:

$$\overline{f_D}(\omega) = (a + \mathrm{Lap}(\frac{\Delta f}{\epsilon}))\omega^2 + (b + \mathrm{Lap}(\frac{\Delta f}{\epsilon}))\omega + c \quad (12)$$

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 $f_D(\omega)$ 和 $f_D(\omega)$:

$$f_D(\omega) = \sum_{j=1}^{J} \sum_{\varphi \in \Phi_j} \sum_{t_i \in D} \lambda_{\varphi_i} \varphi(\omega)$$

$$f_{D'}(\omega) = \sum_{j=1}^{J} \sum_{\varphi \in \Phi_j} \sum_{t_i \in D'} \lambda'_{\varphi_i} \varphi(\omega)$$

根据全局敏感度的定义(见定义 2)有:

$$\sum_{j=1}^{J} \sum_{\varphi \in \Phi_j} \sum_{t_i \in D} \lambda_{\varphi_i} - \sum_{t_i \in D'} \lambda'_{\varphi_i} \leqslant 2 \max_{t} \sum_{j=1}^{J} \sum_{\varphi \in \Phi_j} \lambda_{\varphi_i}$$

由此,可以得到全局敏感度 Δ 为:

$$\Delta = 2 \max_{t} \sum_{j=1}^{J} \sum_{\varphi \in \Phi_j} \lambda_{\varphi_i} \quad (13)$$

最后,只要对式(12) $\overline{f_D}(\omega)$ 进行最优化计算,求解出 $\bar{\omega}$,$\bar{\omega}$ 的求解如式(9),算法流程如下:

算法 1:FM-on-PR。

输入:数据集 D,隐私预算 ϵ,代价函数 $f_D(\omega)$;

输出:多项式回归模型系数的最优解 $\bar{\omega}$;

1:计算全局敏感度 $\Delta = 2 \max_{t} \sum_{j=1}^{J} \sum_{\varphi \in \Phi_j} \lambda_{\varphi_i}$

2:for each $1 \leqslant j \leqslant J$ do

3:　　for each $\varphi \in \Phi_j$ do

4:　　$\lambda_{\varphi} = \sum_{t_i \in D} \lambda_{\varphi_i} + \mathrm{Lap}(\frac{\Delta f}{\epsilon})$

5:　　end for

6:end for

7:令 $\overline{f_D}(\omega) = \sum_{j=1}^{J} \sum_{\varphi \in \Phi_j} \lambda_{\varphi} \varphi(\omega)$;

8:计算 $\bar{\omega} = \arg\min_{\omega} \overline{f_D}(\omega)$(算法如式(9));

9:返回 $\bar{\omega}$;

定理 2:算法 1 满足 ϵ -差分隐私保护机制。

证明:D 和 D' 是一对邻近数据集,且假设它们是在各自的最后一条记录 t_n 和 t'_n 上有所差别。

由 2.1 节可知,因为多项式回归可以通过变量替换转为多元线性回归进行训练,因此任意多项式回归的代价函数最终都可以表示为式(4)的多元线性回归形式。

此外,根据推导,全局敏感度的公式(13)表明其只与数据集的维度有关,与代价函数是否为线性并无关系。因此全局敏感度的计算公式同样适用于转换为多元线性回归之后的多项式回归。由此达成了使用 FM 算法的前提条件,后续的几个算法同理。

根据 FM 的算法流程,设代价函数的全局敏感度为 Δ,并且添加完噪声的代价函数为 $\overline{f_D}(\omega) = \sum_{j=1}^{J} \sum_{\varphi \in \Phi_j} \lambda_{\varphi} \varphi(\omega)$。

以下证明算法 1 满足 ϵ - 差分隐私保护机制。

$$\frac{\Pr\{\overline{f_D}(\omega) \mid D\}}{\Pr\{\overline{f_D}(\omega) \mid D'\}} =$$

$$\frac{\prod_{j=1}^{J}\prod_{\varphi\in\Phi_j}\exp\left(\frac{\epsilon\cdot\|\sum_{t_i\in D}\lambda_{\Phi_i}-\lambda_\varphi\|_1}{\Delta}\right)}{\prod_{j=1}^{J}\prod_{\varphi\in\Phi_j}\exp\left(\frac{\epsilon\cdot\|\sum_{t_i\in D'}\lambda_{\varphi_i}-\lambda_\varphi\|_1}{\Delta}\right)} \leqslant$$

$$\prod_{j=1}^{J}\prod_{\varphi\in\Phi_j}\exp\left(\frac{\epsilon}{\Delta}\cdot\|\sum_{t_i\in D}\lambda_{\varphi_i}-\lambda'_{\varphi_i}\|_1\right) =$$

$$\exp\left(\frac{\epsilon}{\Delta}\cdot\sum_{j=1}^{J}\sum_{\varphi\in\Phi_j}\|\lambda_{\varphi_i}-\lambda_{\Phi'_i}\|_1\right) \leqslant$$

$$\exp\left(\frac{\epsilon}{\Delta}\cdot 2\max_t\sum_{j=1}^{J}\sum_{\varphi\in\Phi_j}\|\lambda_{\varphi_i}\|_1\right) = \exp(\epsilon)$$

因此,算法 1 满足 ϵ - 差分隐私的定义(6)。

3.2.2 面向多项式回归的不同系数扰动函数算法

面向多项式回归的不同系数扰动函数算法(Functional Mechanism with Different Perturbation of Coefficients on Polynomial Regression)简称 DPC-on-PR。

定义 X_s 是包含了所有会泄露隐私的敏感属性 x_s 的集合。在 *FM* 算法的基础上(见 3.2.1),对 φ 也进行划分:只要 φ 中有包含了任何一个 X_s 中的元素对应的系数 ω_s,就将其纳入集合 Φ_s,否则纳入 Φ_n 中。对于 Φ_s 中的 φ 项,分给其相对较少隐私预算 ϵ_s, 而分给 Φ_n 的隐私预算 ϵ_n 则相对更多。设 $\epsilon_s = \gamma\epsilon_n$,其中 $0 < \gamma < 1$。给代价函数添加完噪声后, 只需对添加过噪声的代价函数进行最优化求解,算法流程如下:

算法 2:DPC-on-PR。

输入:数据集 D,隐私预算 ϵ,代价函数 $f_D(\omega)$,隐私预算比率 γ;

输出:多项式回归模型系数的最优解 $\bar{\omega}$;

1:设置 $\Phi_s = \{\}, \Phi_n = \{\}$;
2:for each $1 \leqslant j \leqslant J$ do
3: for each $\varphi \in \Phi_j$ do
4:if φ 不包括任意属于敏感集 X_s 中元素的 ω_s
5:then 将 φ 添加至集合 Φ_n;
6:else 将 φ 添加至集合 Φ_s;
7: end if
8: end for
9:end for
10:计算全局敏感度 Δ(算法同(13));
11:计算 $\Delta_1 = 2\max_t\sum_{j=1}^{J}\sum_{\varphi\in\Phi_n}\|\lambda_{\varphi_i}\|_1, \beta_1 = \Delta_1/\Delta$
12:计算 $\Delta_2 = 2\max_t\sum_{j=1}^{J}\sum_{\varphi\in\Phi_s}\lambda_{\varphi 1}, \beta_1 = \Delta_2/\Delta$;
13:计算 $\epsilon_n = \frac{1}{\beta_1+\gamma\beta_2}, \epsilon_n = \frac{\gamma}{\beta_1+\gamma\beta_2}\epsilon$;
14:for each $1 \leqslant j \leqslant J$ do
15: for each $\varphi \in \Phi_j$ do
16: if $\varphi \in \Phi_n$ then
17: 设置 $\lambda_\varphi = \sum_{t_i\in D}\lambda_{\Phi_i} + \mathrm{Lap}(\frac{\Delta f}{\epsilon_n})$;
18: else
19: 设置 $\lambda_\varphi = \sum_{t_i\in D}\lambda_{\Phi_i} + \mathrm{Lap}(\frac{\Delta f}{\epsilon_s})$;
20: end if
21: end for
22:end for
23:令 $\overline{f_D}(\omega) = \sum_{j=1}^{J}\sum_{\varphi\in\Phi_j}\lambda_\varphi\varphi(\omega)$;
24:计算 $\bar{\omega} = arg\min_\omega \overline{f_D}(\omega)$ (算法如式(9));
25:返回 $\bar{\omega}$;

定理 3:算法 2 满足 ϵ -差分隐私保护机制。

证明:假设 D 和 D' 是一对邻近数据集,且假设它们是在各自的最后一条记录 t_n 和 t'_n 上有所差别。与算法 1 一样,可以直接以线性回归形式的 $\overline{f_D}(\omega)$ 为条件(见定理 2)进行证明:

首先可以算出:

$$\Pr(\overline{f_D}(\omega \mid D))) = \prod_{\varphi\in\Phi_n}\exp\left(\frac{\epsilon_n\sum_{t_i\in D}\lambda_{\varphi_i}-\lambda_{\varphi 1}}{\Delta}\right)$$

$$\prod_{\varphi\in\Phi_s}\exp\left(\frac{\epsilon_s\sum_{t_i\in D}\lambda_{\varphi_i}-\lambda_{\varphi 1}}{\Delta}\right)$$

同理可得 $\Pr(\overline{f_D}(\omega \mid D'))$。

$$\frac{\Pr(\overline{f_D}(\omega \mid D)))}{\Pr(\overline{f_D}(\omega \mid D')))} \leqslant$$

$$\prod_{\varphi\in\Phi_n}\exp\left(\|\frac{\epsilon_n}{\Delta}\sum_{t_i\in D}\lambda_{\Phi_i}-\sum_{t_i\in D}\lambda_{\Phi_i}\|_1\right)$$

$$\prod_{\varphi\in\Phi_s}\exp\left(\|\frac{\epsilon_s}{\Delta}\sum_{t_i\in D}\lambda_{\varphi_i}-\sum_{t_i\in D'}\lambda_{\varphi_i}\|_1\right)$$

证毕。

3.2.3 面向多项式回归的差分隐私预算分配算法

面向多项式回归的差分隐私预算分配算法(Differentiated Privacy Budget Allocation on Polynomial Regression)简称 DPBA-on-PR。

根据式(13),可以得到代价函数 $f_D(\omega)$ 的全局敏感度为 $\Delta f_D(\omega) = 2(2d + d^2)$。

随后,将 $f_D(\omega)$ ($f_D(\omega)$ 的表达式同式(4)形式)拆解开为 2 个单项式,分别为:

$$g_D(\omega) = \sum_{1\leqslant j,l\leqslant d}(\sum_{t_i\in D}x_{ij}x_{il})\omega_j\omega_l$$

$$h_D(\omega) = \sum_{j=1}^{d}(2\sum_{t_i\in D}y_ix_{ij})\omega_j$$

根据式(13),同理可得 $g_D(\omega)$ 和 $h_D(\omega)$ 的全局敏感度分别为 $\Delta g_D(\omega) = 2d^2$, $\Delta h_D(\omega) = 4d$。

可以发现整个多项式的全局敏感度和两个单项式

的全局敏感度满足如下关系:

$$\Delta f_D(\omega) = 2(2d + d^2) = \Delta g_D(\omega) + \Delta h_D(\omega)$$

根据上式,就可以对 $g_D(\omega)$ 和 $h_D(\omega)$ 各自添加噪声得到 $\overline{g_D}(\omega)$ 和 $\overline{h_D}(\omega)$,再相加,就能得到 $\overline{f_D}(\omega)$:

$$\overline{f_D}(\omega) = \overline{g_D}(\omega) + \overline{h_D}(\omega)$$

随后只要对 $\overline{f_D}(\omega)$ 进行最优化计算即可。

算法 3:DPBA-on-PR。

输入:数据集 D,隐私预算 ϵ,代价函数 $f_D(\omega)$;

输出:多项式回归模型系数的最优解 $\bar{\omega}$;

1:计算 $g_D(\omega)$ 的全局敏感度 Δf_1, $h_D(\omega)$ 的全局敏感度 Δf_2;

2:设 $g_D(\omega)$ 的系数是 a, $h_D(\omega)$ 的系数是 b;

3:if $\Delta f_1 > \Delta f_2$ then

4:　设 $\alpha = \dfrac{\Delta f_1}{(\Delta f_1)^2 + \Delta f_2}\beta (0 \leqslant \beta \leqslant \dfrac{\Delta f_2}{\Delta f_1} + \Delta f_1)$;

5:else

6:　设 $\alpha = \dfrac{\Delta f_1}{(\Delta f_2)^2 + \Delta f_1}$;

7:end if

8:设 $\epsilon_1 = \alpha\epsilon$, $\epsilon_2 = \epsilon - \epsilon_1$;

9:计算 $\overline{g_D}(\omega) = (a + \mathrm{Lap}(\dfrac{\Delta g_D(\omega)}{\epsilon_g}))\omega^2$,

$\overline{h_D}(\omega) = (b + \mathrm{Lap}(\dfrac{\Delta h_D(\omega)}{\epsilon_h}))\omega$;

10:得到 $\overline{f_D}(\omega) = \overline{g_D}(\omega) + \overline{h_D}(\omega)$;

11:计算 $\bar{\omega} = \arg\min_{\omega} \overline{f_D}(\omega)$(算法如式(9));

12:返回 $\bar{\omega}$;

算法流程中的参数 β,是用来调节 $g_D(\omega)$ 和 $h_D(\omega)$ 的隐私预算分配比率的变量,其取值范围的端点(见上述算法流程中步骤 4)分别对应着将总的隐私预算 ϵ 全部分配给 $g_D(\omega)$ 和全部分配给 $h_D(\omega)$。显然,$g_D(\omega)$ 相对于 $h_D(\omega)$ 自然有较高的全局敏感度(式(13)),因此必须合理制定 β 的取值使得 $g_D(\omega)$ 添加噪声规模较小。

在高维度的数据集中,根据全局敏感度的计算方法(式(13)),$g_D(\omega)$ 的全局敏感度 Δf_1 更是远远高于 Δf_2 的,因此为了保证训练结果的准确性,$g_D(\omega)$ 应该被分配到更多的隐私预算使得对其添加的噪声规模降低,根据算法 3 流程中的步骤 4,β 的取值就要增大。

表 1 总结了 β 为各个取值时,$g_D(\omega)$ 和 $h_D(\omega)$ 被分配到的噪声服从的分布。

其中,d_0 为多项式经过变量替换转化为多元线性回归后的维度,即式(3)中 $\vec{Z}$ 的维度。并且以下对 FM-on-PR 和 DPBA-on-PR 隐私预算分配策略的比较均在它们的总隐私预算都为 ϵ 的条件下进行。

表 1　β 的取值以及 $g_D(\omega)$ 和 $h_D(\omega)$ 的噪声分布

β 取值	β 取值(d_0 表示)	$g_D(\omega)$ 的噪声分布	$h_D(\omega)$ 的噪声分布
Δf_1	$\dfrac{4d_0^4 + 4d_0 + 1}{2d_0^2 + 4d_0 + 1}$	$\mathrm{Lap}(\dfrac{\Delta f_1^2 + \Delta f_2}{\Delta f_1 \cdot \epsilon})$	$\mathrm{Lap}(\dfrac{\Delta f_1^2 + \Delta f_2}{\epsilon})$
$\dfrac{\Delta f_1^2 + \Delta f_2}{\Delta f_1 + \Delta f_2}$	$\dfrac{4d_0^4 + 2d_0^2}{2d_0^2 + 4d_0 + 1}$	$\mathrm{Lap}(\dfrac{\Delta f_1 + \Delta f_2}{\epsilon})$	$\mathrm{Lap}(\dfrac{\Delta f_1 + \Delta f_2}{\epsilon})$

首先说明,如果使用 FM-on-PR 中的分配隐私预策略,则 $g_D(\omega)$ 和 $h_D(\omega)$ 始终都会被分配到服从 $\mathrm{Lap}(\dfrac{\Delta f_1 + \Delta f_2}{\epsilon})$ 的噪声(算法 1 第 4 步)。

当 β 取值为 Δf_1,相比之下,DPBA-on-PR 的 $g_D(\omega)$ 确实被分配到了规模更少的噪声,但是 $h_D(\omega)$ 会被分配到规模更大的噪声。当 d_0 的维度很大时,Δf_1^2 也会很大,同时,相比较 FM-on-PR 算法,$h_D(\omega)$ 被分配到了更少的隐私预算,因此,此时的 $h_D(\omega)$ 是被分配了过量的噪声,所以,在 d_0 较大和隐私预算较小时,β 的取值不可以过大。

当 β 取值为 $\dfrac{\Delta f_1^2 + \Delta f_2}{\Delta f_1 + \Delta f_2}$ 时,由表 2 可见,$g_D(\omega)$ 和 $h_D(\omega)$ 分的噪声服从与 FM-on-PR 算法一样的分布,此时的 DPBA-on-PR 算法和 FM-on-PR 算法完全一致。因为 $g_D(\omega)$ 敏感度较高,是对拟合结果影响较大的子目标函数,所以,对 FM-on-PR 算法的优化即体现在降低 $g_D(\omega)$ 所分配的噪声上,因此,可以将 $\dfrac{\Delta f_1^2 + \Delta f_2}{\Delta f_1 + \Delta f_2}$ 设置为 β 取值的下限,只要 β 大于此值,$g_D(\omega)$ 就会被分配到相比 FM-on-PR 规模更小的噪声。

β 取值的上限,可以由基本不等式得出,由上文分析得,β 的取值须大于 $\dfrac{\Delta f_1^2 + \Delta f_2}{\Delta f_1 + \Delta f_2}$,小于 Δf_1,观察它们以 d_0 表示的形式(见表 2),可以得到如下结果:

$$\frac{4d_0^4 + 2d_0^2}{2d_0^2 + 4d_0 + 1} \leqslant \frac{(2d_0^2 + \sqrt{2}d_0)^2}{2d_0^2 + 4d_0 + 1} \leqslant \frac{4d_0^4 + 4d_0 + 1}{2d_0^2 + 4d_0 + 1} \tag{14}$$

因此,β 的取值上限可以定为 $\dfrac{(2d_0^2 + \sqrt{2}d_0)^2}{2d_0^2 + 4d_0 + 1}$。

最终 β 的取值范围为:

$\left[\frac{4d_0^4+2d_0^2}{2d_0^2+4d_0+1},\frac{(2d_0^2+\sqrt{2}d_0)^2}{2d_0^2+4d_0+1}\right]$。

定理 4:算法 3 满足 ϵ -差分隐私保护机制。

证明:如果给 $g_D(\omega)$，$h_D(\omega)$ 分别分配隐私预算 ϵ_g，ϵ_h，且设 $\epsilon=\epsilon_h+\epsilon_g$。根据 FM 机制中的相关证明(定理 2),这三个单独的代价函数又分别满足 ϵ_g - 差分隐私,ϵ_h -差分隐私。于是,由定理 1 差分隐私的串行组合性可知，$\Delta f_D(\omega)$ 的算法满足 ϵ -差分隐私。

4 实验与分析

4.1 实验环境

实验环境为 Intel(R) Core(TM) i7-9750H CPU 2.60 GHz,16G 内存。实验使用的数据集为:来自 UCI 的联合循环发电厂(CCPP)数据集、个人家庭用电(IHEPC)数据集和 Kaggle 上获取的 diamond 钻石价格预测数据集。属性如表 2 所示。

表 2 实验中使用的数据集

数据集名称	实例数	维数
CCPP	9 568	4
IHEPC	2 075 259	9
Diamond	53 940	10

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该算法进行训练,通过训练结果的精确度来判断它们的可用性。此外,为了检测隐私预算 ϵ 对模型准确性的影响,对每个数据集也将以不同的隐私预算 ϵ 进行多次训练。由于噪声的影响,也会进行多次实验取结果的均值。

4.2 实验结果及其分析

4.2.1 隐私预算对拟合结果准确度的影响

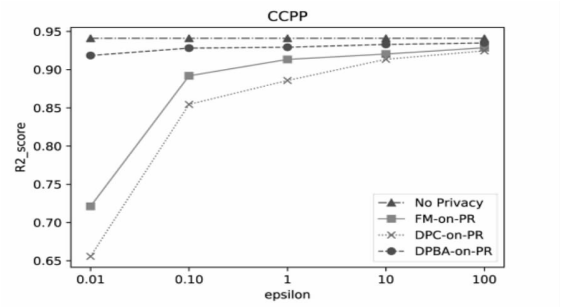
回归分析有多种性能指标衡量其精确性,该文使用拟合优度(R2)来衡量训练模型的准确性,R2 的取值不会超过 1,越接近 1 表示训练结果的准确度越高,特别地,当 R2<0 时,表示拟合结果尚不如取数据集的平均值的精度高。

在多项式回归的阶数设置方面,经测试,对于 CCPP 数据集,将它们的阶数设置为 3 最佳;对于 IHEPC 数据集和 Diamond 数据集,设置为 2 最佳。

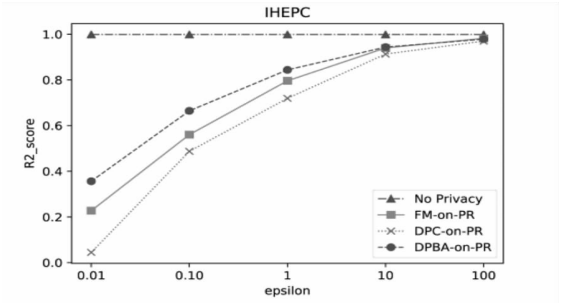
图 2 分别是三个算法对三个数据集在不同隐私预算 ϵ 下训练结果的准确性的比较,并且 ϵ 取值范围为 {0.01,0.1,1,10,100}。横坐标是隐私预算 ϵ 的取值,R2_score 是结果的拟合优度 R2。标签中,No privacy 即不含任何隐私保护机制的多项式回归,它将作为其他三个算法精确性的比较基准。

从图 2 可见,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,几个算法之间的精确度差距甚微,同时与无隐私保护的算法的精确度接近。

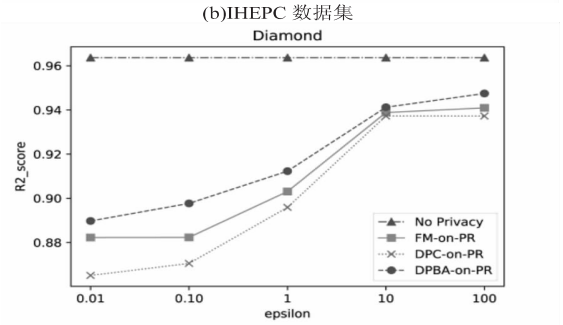
另一方面,无论基于哪个数据集的实验,结果都表明 DPBA-on-PR 算法拥有最高的精确度,其次是 FM-on-PR 算法,DPC-on-PR 算法最次。由于 DPC-on-PR 算法提出的目的是为了加强 FM-on-PR 算法的数据安全性,它只对会泄露隐私的特征变量添加更多噪声,并考虑每个特征变量与输出的关联性,因此可能会对高敏感度的特征变量添加很多不必要的噪声,从而大大降低训练结果的准确性,因此精确性方面不如 FM-on-PR 算法。



(a)CCPP 数据集



(b)IHEPC 数据集



(c)Diamond 数据集

图 2 在不同隐私预算下对三个数据集进行回归训练的结果

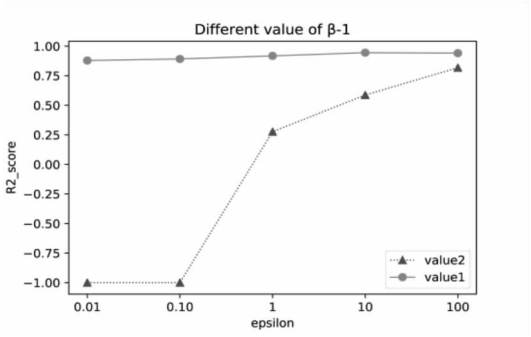
4.2.2 DPBA-on-PR 算法中 β 取值对训练精确度的影响

因为对该算法中隐私预算分配策略的优化是针对高维度的数据集,所以这部分实验中,选取维度较高的 diamond 数据集进行实验,经过数据集预处理后,它的原始维度 d 为 9,将其多项式形式的目标函数转换为多元线性回归之后的维度 d_0 达到了 54。同样,衡量准确性的标准仍然是拟合优度 R2。

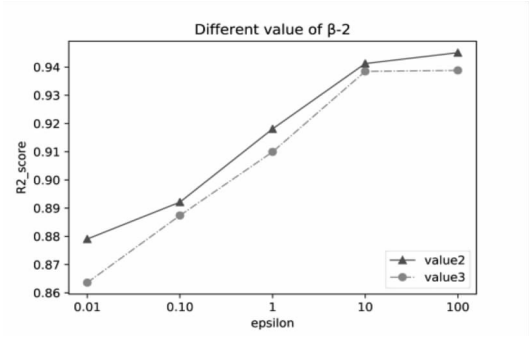
图 3 是 DPBA-on-PR 算法取不同 β 值并在不同

隐私预算 ϵ 下的训练结果,横坐标 epsilon 是隐私预算的取值,纵坐标是拟合优度 R2;标签为 value1 的曲线中 β 取值均为 value1 = Δf_1(即式(14)中的不等式上限),在本实验中具体取值为 5 832(计算公式参考表 2);标签为 value2 的曲线中的 β 取值为 value2 = $\frac{(2d_0^2+\sqrt{2}d_0)^2}{2d_0^2+4d_0+1}$(即式(14)中的不等式的中间项以及得出的$\beta$取值新上限),在本实验中具体取值为5 771;标签为 value3 的曲线中的 β 的取值为 value3 = $\frac{4d_0^4+2d_0^2}{2d_0^2+4d_0+1}$(即式(14)中的不等式下限),在本实验中具体取值为 5 621。

由于 value1 与 value2 的 R2_score(即纵坐标值)差值和 value2 与 value3 的差值相差过大不便于比较,因此将比较结果分在了两张图表中。



(a) β 取值为 value1 和 value2 时的训练结果比较



(b) β 取值为 value2 和 value3 时的训练结果比较

图 3　DPBA-on-PR 算法中 β 取值对准确性的影响

图 3(a)中,β 取值为 value1 时,在 ϵ = 0.01 和 0.1 的情况下,实验得到的 R2_score 均远远小于-1,而在此图中,value1 代表的曲线在这两个取值上均以-1 代替表示,仅代表在这两种情况下,训练结果不理想。结果看来,β 取值为 value2 的训练精确度比 value1 时高很多,尤其在隐私预算 ϵ 较小时更明显。

图 3(b)中,β 取值为 value3 时,如 3.2.3 小节所述,此时的 DPBA-on-PR 算法就与 FM-on-PR 算法一致,β 取值为 value2 的训练精确度比取值为 value3 时高,结合 4.2.1 的实验,这也证明了 DPBA-on-PR 算法的精确度性能是优于 FM-on-DPBA 的。

综上,在 3.2.3 节中 β 的取值策略是可行的。

5　结束语

该文研究并设计了三个面向多项式回归的差分隐私保护算法,并且理论证明了它们满足差分隐私性质;三个算法与无隐私保护的多项式算法的训练结果进行的比较,也证明了它们的可用性;此外经实验,得出了数据可用性最高的算法为 DPBA-on-PR 的结论。由于 DPC-on-PR 和 DPBA-on-PR 分别在数据安全性和数据可用性方面进行了改进,因此在实际应用中,可以根据需求来对两个算法进行选择。总体来说,面向多项式回归分析的差分隐私算法研究的重点是数据的安全性和可用性间平衡的问题,通常这方面研究的切入点是隐私预算的分配与全局敏感度的计算,目前,全局敏感度的计算也并没有达到相当精确的程度,因此,噪声的添加量仍有优化的余地,这也将是未来面向多项式回归的差分隐私保护算法研究的方向。

参考文献:

[1] 刘向宇,王　斌,杨晓春. 社会网络数据发布隐私保护技术综述[J]. 软件学报,2014,25(3):576-590.

[2] 叶青青,孟小峰,朱敏杰,等. 本地化差分隐私研究综述[J]. 软件学报,2018,29(7):1981-2005.

[3] KIKUCHI H,HAMANAGA C,YASUNAGA H,et al. Privacy-preserving multiple linear regression of vertically partitioned real medical datasets[C]//Proceedings of 2017 IEEE 31st international conference on advanced information networking and applications. Taipei:IEEE,2017.

[4] MOHASSEL P,ZHANG Y. SecureML:a system for scalable privacy-preserving machine learning[C]//Proceedings of 2017 IEEE symposium on security and privacy. San Jose,CA,USA:IEEE,2017:19-38.

[5] LI B,LIU Y,HAN X,et al. Cross-bucket generalization for information and privacy preservation[J]. IEEE Transactions on Knowledge and Data Engineering,2018,30(3):449-459.

[6] DWORK C, SMITH A. Differential privacy for statistics: what we know and what we want to learn[J]. Journal of Privacy and Confidentiality,2009,1(2):135-154.

[7] 熊　平,朱天清,王晓峰. 差分隐私保护及其应用[J]. 计算机学报,2014,37(1):101-122.

[8] MCSHERRY F,TALWAR K. Mechanism design via differential privacy[C]//Proceedings of 48th annual IEEE symposium on foundations of computer science. Providence,RI,USA:IEEE,2007:94-103.

[9] LEI J. Differentially private m-estimators[C]//Proceedings of the 24th international conference on neural information processing systems. Red Hook,NY,USA:Curran Associates

(下转第 128 页)