第 46 卷 第 2 期 计 算 机 工 程 2020 年 2 月
Vol. 46 No. 2 Computer Engineering February 2020

·网络空间安全· 文章编号：1000-3428(2020)02-0134-07 文献标志码：A 中图分类号：TP391

一种基于 ExtraTrees 的差分隐私保护算法

李 杨，陈子彬，谢光强

（广东工业大学 计算机学院，广州 510006）

摘 要： 为在同等隐私保护级别下提高模型的预测准确率并降低误差，提出一种基于 ExtraTrees 的差分隐私保护算法 DiffPETs。在决策树生成过程中，根据不同的准则计算出各特征的结果值，利用指数机制选择得分最高的特征，通过拉普拉斯机制在叶子节点上进行加噪，使算法能够提供 ε-差分隐私保护。将 DiffPETs 算法应用于决策树分类和回归分析中，对于分类树，选择基尼指数作为指数机制的可用性函数并给出基尼指数的敏感度，在回归树上，将方差作为指数机制的可用性函数并给出方差的敏感度。实验结果表明，与决策树差分隐私分类和回归算法相比，DiffPETs 算法能有效降低预测误差。

关键词： 差分隐私；ExtraTrees 算法；分类；回归分析；决策树

开放科学（资源服务）标志码（OSID）：

中文引用格式： 李杨，陈子彬，谢光强. 一种基于 ExtraTrees 的差分隐私保护算法[J]. 计算机工程，2020，46(2)：134-140.

英文引用格式： LI Yang，CHEN Zibin，XIE Guangqiang. A differential privacy protection algorithm based on ExtraTrees[J]. Computer Engineering，2020，46(2)：134-140.

A Differential Privacy Protection Algorithm Based on ExtraTrees

LI Yang，CHEN Zibin，XIE Guangqiang

（School of Computers，Guangdong University of Technology，Guangzhou 510006，China）

【Abstract】 To improve the prediction accuracy and reduce prediction error under the same level of privacy protection，this paper proposes a differential privacy protection algorithm DiffPETs based on ExtraTrees. During the decision tree generation process，the result value of each feature is calculated according to different criteria，the feature with the highest score is selected by the exponential mechanism and noise is added on the leaf nodes through Laplace mechanism，enabling the algorithm to provide the-differential privacy protection. Then，this paper applies DiffPETs algorithm to the classification and regression analysis of decision tree. For classification tree，Gini index is selected as the availability function of index mechanism and the sensitivity of Gini index is given. For the regression tree，variance is taken as the availability function of index mechanism and the sensitivity of variance is given. Experimental results show that compared with decision tree differential privacy classification and regression algorithm，the DiffPETs algorithm can effectively reduce prediction error.

【Key words】 differential privacy；ExtraTrees algorithm；classification；regression analysis；decision tree

DOI： 10.19678/j.issn.1000-3428.0053824

0 概述

随着计算机存储数据成本的不断降低以及云计算、大数据、移动互联网等技术的广泛应用和普及，企业、政府组织的信息系统存储了海量数据，如社交网络公司的社交数据、各大电商平台的顾客购买数据以及医院的病人医疗数据等，这些隐私数据常被用于分析、统计和挖掘。从这些数据中发现潜在的商业价值和规律，同时保证不泄露用户敏感信息，引起了数据挖掘和隐私保护领域相关学者的广泛关注。

在早期的相关研究中，数据隐私保护的方法主要是 k-anonymity[1] 及其扩展模型，如 l-diversity[2]、t-closeness[3] 和 (α,k)-anonymity[4] 等。尽管这些模

基金项目： 国家自然科学基金(61472089)；广东省科技计划项目(2014B010103005，2016A040403078)；NSFC-广东联合基金项目(U1501254)。

作者简介： 李 杨(1980—)，女，副教授、博士，主研方向为隐私保护、数据可视化、机器学习；陈子彬，硕士研究生；谢光强(通信作者)，副教授、博士。

收稿日期： 2019-01-28 **修回日期：** 2019-03-04 **E-mail：** superxgq@163.com

型能够在一定程度上保护数据中用户的敏感信息,但是它们的共同缺点是需要假设攻击者拥有的背景知识和特定的攻击模型,导致其应用场景受到很多制约。在这种背景下,DWORK C 等学者[5-6]在2006 年提出具有严格理论证明的差分隐私保护模型,该模型为保护数据隐私带来了一种新的解决思路。在差分隐私保护的定义下,即便攻击者拥有除目标用户以外的全部其他信息,用户仍然能够防御攻击并且不会泄露敏感信息。差分隐私保护通过在针对某个数据集上的查询结果添加噪声来提供保护,且基于差分隐私保护的算法在输出结果上添加的噪声独立于数据集的规模,添加少量的噪声能够保证隐私数据的安全性以及输出结果的可用性。

目前,差分隐私保护的研究方向集中在差分隐私数据发布[7-9](延伸至社交网络[10]、轨迹数据[11])、社交网络研究[12-13]、查询处理[14-15]以及和数据挖掘相结合(如聚类[16]、分类[17-19]、集成学习[20])等方面,而在回归分析上的研究则相对较少。进行差分隐私保护回归分析研究的难点在于求解回归目标函数的敏感度很大,难以保证回归准确度。

ExtraTrees 作为一种使用频率较高的集成学习模型,能够应用于决策树分类和回归分析中,本文提出一种基于 ExtraTrees 的差分隐私保护算法。在决策树分类上,利用 ExtraTrees 算法 2 次随机选择的特性,提高指数机制在同等隐私预算下的使用效率,通过将基尼指数作为指数机制在分类树特征选择上的可用性函数来降低敏感度,从而减少所添加的噪声量。在回归树上,选择最佳分裂特征和分裂点时将方差作为指数机制的可用性函数,标记叶子节点时对其均值添加拉普拉斯噪声。

1　相关工作

1.1　差分隐私的定义

定义 1(ε-差分隐私)　设有 2 个数据集 D_1 和 D_2,两者的属性结构相同,不同的记录数量为 $|D_1\Delta D_2|$。当 $|D_1\Delta D_2|=1$ 时,设一个随机化算法为 M,$R_{\text{Range}}(M)$ 代表算法 M 的所有输出构成的集合体,S 是 $R_{\text{Range}}(M)$ 中的一个子集,若算法 M 提供 ε-差分隐私保护,则对于所有的 $S\in R_{\text{Range}}(M)$,有:

$$\Pr[M(D_1)\in S]\leqslant\exp(\varepsilon)\Pr[M(D_2)\in S] \quad (1)$$

其中,$\Pr[\ ]$代表某个事件泄露隐私的概率,其由算法 M 本身的性质所决定,ε 称为隐私保护预算,是由数据提供者视隐私保护程度所给定的参数,隐私保护程度与 ε 的取值成反比,ε 通常取小于 1 的常数。式(1)表示对算法 M 的任何一个输出结果,不管函数的输入是 D_1 还是 D_2,最后得到这个输出结果的概率差别很小。

1.2　差分隐私保护实现机制

差分隐私保护的实现原理是通过添加噪声来提供保护,经常使用的 2 种噪声模型为拉普拉斯机制[21]和指数机制[22]。某个算法提供差分隐私保护所添加噪声的多少与全局敏感度有关。

定义 2(全局敏感度)　设函数 $f:D\rightarrow R^d$ 的输入为数据集 D,输出为任一实体对象,数据集 D_1 和 D_2 之间不同的记录个数不大于 1,则 f 的全局敏感度定义为:

$$\Delta f=\max_{D_1,D_2}\|f(D_1)-f(D_2)\|_1 \quad (2)$$

其中,$\|f(D_1)-f(D_2)\|_1$ 表示 $f(D_1)$ 和 $f(D_2)$ 之间的 1-阶范数距离。

定义 3(拉普拉斯机制)　对于数据集 D 上的任一函数 $f:D\rightarrow R^d$,算法 M 添加由拉普拉斯分布产生的相互独立的噪声变量,以提供 ε-差分隐私保护,具体如下:

$$M(D)=f(D)+L^1_{\text{Lap}}(\Delta f/\varepsilon)+L^2_{\text{Lap}}(\Delta f/\varepsilon)+\cdots+L^d_{\text{Lap}}(\Delta f/\varepsilon) \quad (3)$$

定义 4(指数机制)　设算法 M 的输入为数据集 D,输出为一个具体对象 $r\in R_{\text{Range}}(M)$,$q(D,r)$ 为可用性函数,Δq 为函数 $q(D,r)$ 的敏感度。如果输出结果满足式(4),则称算法 M 提供 ε-差分隐私保护。

$$M(D,u)=\left\{u:\Pr[r\in R_{\text{Range}}]\propto\exp\left(\frac{\varepsilon q(D,r)}{2\Delta q}\right)\right\} \quad (4)$$

1.3　差分隐私的性质

定义 5(并行组合性)[22]　设算法 M 提供 ε-差分隐私保护,给定数据集 D 被切割成互不相交的子集 $D_1,D_2,\cdots,D_n$,则 M 作用在 $\{D_1,D_2,\cdots,D_n\}$ 上的组合算法 $\{M(D_1),M(D_2),\cdots,M(D_n)\}$ 具有 ε-差分隐私。

定义 6(序列组合性)[23]　设有算法 $M_1,M_2,\cdots,M_n$,分别提供 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 的差分隐私保护,则对于给定数据集 D,由这些算法组成的组合算法 $\{M(D_1),M(D_2),\cdots,M(D_n)\}$ 具有 $\sum_{i=1}^{n}\varepsilon_i$-差分隐私。

1.4　ExtraTrees 算法

ExtraTrees 是一种集成学习算法,包含众多决策树,分类结果最终由其包含的众多决策树的输出结果共同投票决定,回归则由这些决策树输出值的均值来决定。ExtraTrees 算法最大的特点是在分裂特征的选择上采用随机的方式,随机选择某个特征的某个取值作为该特征的分裂点。由于这 2 个随机性特点,使得 ExtraTrees 算法泛化能力较强,具有抵抗噪声的能力。

1.5　已有相关研究

在差分隐私保护决策树分类算法研究中,文献[24]提出了基于线性查询函数的分类器算法 SuLQ-based ID3。文献[25]为解决 SuLQ-based ID3

算法添加噪声过多的问题,提出一种基于决策树建造的差分隐私保护算法 DiffP-C4.5。文献[26]沿用数据泛化的思想,在决策树的基础上,提出一种非交互式的差分隐私数据发布算法 DiffGen。实验结果表明,在隐私预算相同的情况下,DiffGen 的实验效果相较于其他算法更好。文献[27]在 DiffGen 算法的基础上,通过对细分方案进行改进,为连续型数值加上权重,从而更充分地利用隐私预算,相较于 DiffGen 具有更好的分类效果。文献[28]将差分隐私和随机决策树模型相结合,提出了 RDT 算法,但是该算法仅使用拉普拉斯机制,在叶子节点的向量添加噪声。文献[29]提出了差分隐私随机森林算法,该算法的缺点在于仅适合标称型属性的数据集,对于包含连续属性的数据集需要先将连续属性转化成离散属性,然后对数据集进行分类。

在差分隐私保护回归分析算法研究中,文献[30]提出了一个利用拉普拉斯机制和指数机制进行统计分析的总体框架,但是该框架要求统计分析的输出空间有界,导致不适用于线性回归,并且其直接在计算结果上添加拉普拉斯噪声,虽然在一定程度上保护了用户隐私,但是由于添加过多的噪声,造成计算结果的可用性非常低。文献[31-32]研究表明,当回归任务的成本函数是凸函数并且可微时,回归才能满足差分隐私保护。相较于文献[30]直接在输出结果上添加噪声,文献[31-32]在 n 个目标函数的输出平均值上添加噪声以实现隐私保护,其添加的噪声量有所减少,但该方法普适性较低。文献[33]为了解决回归结果误差大的问题,提出了一种函数机制(Functional Mechanism,FM)。FM 机制不是在输出结果上添加噪声,而是通过对回归的目标函数添加噪声,得到加噪的目标函数,然后求解出最优回归模型参数。因此,该机制普适性较高,在多种回归模型上都能应用。但是,由于其计算出来的全局敏感度与数据集的维度成正比,造成噪声添加过多,使得回归结果误差大。文献[34]提出了一种根据回归目标函数敏感度大小来分配隐私预算的算法 Diff-LR。该算法将线性回归分析的目标函数拆分成不同的模块,再分别求解出不同模块的敏感度。对模块分配隐私预算时,敏感度大的模块分派隐私预算多,敏感度小的模块分派隐私预算少。虽然相对于函数机制 FM,该算法有效降低了敏感度,但是其敏感度依然和数据集维度正相关,导致在数据集维度较大的情况下带来过多的噪声。

2 DiffPETs 算法的框架描述与分析

本文提出一种基于 ExtraTrees 的差分隐私保护算法 DiffPETs,并将其应用于决策树分类和回归分析。

2.1 算法描述

DiffPETs 差分隐私算法伪代码如下:

算法 1 DiffPETs(D,A,P_ε,t,h)

输入 数据集 D,特征集 A,总隐私预算 P_ε,产生决策树的个数 t,单棵树的高度 h

输出 满足 ε-差分隐私的学习器 $T=\{T_1,T_2,\cdots,T_t\}$

1. $\varepsilon_1=\frac{P_\varepsilon}{t}$

2. $\varepsilon_2=\frac{\varepsilon_1}{2(h+1)}$

3. for i = 1 to t:

4. 从 D 中有放回地随机抽样大小为 $|D|$ 的训练集 D(i)

5. 调用算法 2 生成树 T_i = Build_DiffPET(D(i),A,ε_2,1)

6. end for

7. 返回提供 ε-差分隐私的学习器 $T=\{T_1,T_2,\cdots,T_t\}$

算法 2 Build_DiffPET(D,A,ε,h)

输入 数据集 D,特征集 A,隐私预算 ε,单棵树的高度 h

输出 满足 ε-差分隐私的学习器 T_i

终止条件 节点全部记录的标签一致,误差达到阈值或树达到最大高度 h

1. if 节点达到了终止条件:

2. 返回叶子节点 N_D = Noise(|D|)

3. end if

4. 随机地从特征集 A 中挑选 K 个特征 $\bar{A}=\{A_1,A_2,\cdots,A_K\}$

5. 从 K 个特征中分别随机选择 K 个分裂点 $\{s_1,s_2,\cdots,s_K\}$

6. 从 K 个特征和分裂点中,使用指数机制用以下概率选择最优特征和分裂点:

$$P_i=\frac{\exp\left(\frac{\varepsilon q(D,A')}{2\Delta q}\right)}{\sum_{A'\in\bar{A}}\exp\left(\frac{\varepsilon q(D,A')}{2\Delta q}\right)}$$

其中,$q(D,A')$ 表示可用性函数,Δq 为可用性函数的敏感度.

7. 根据最佳分裂特征和分裂点将当前数据集分割成 2 个子集 D_l 和 D_r

8. 在最佳分裂特征和分裂点的基础上分别建立左右子树 t_l = Build_DiffPET(D_l,A,ε,h+1),t_r = Build_DiffPET(D_r,A,ε,h+1)

9. 返回提供 ε-差分隐私的学习器 T_i

算法 1 展示了 DiffPETs 的基本框架,其涉及 5 个主要的输入参数:数据集 D,特征集 A,隐私预算 P_ε,产生决策树的数目 t 和单棵树的高度 h。DiffPETs 首先根据生成决策树的数量 t 和树的高度 h 将隐私预算 P_ε 进行分配(第 1 ~ 第 2 行),然后根据生成决策树的数量 t 循环建立单棵决策树,每次建立单棵决策树时使用不同的抽样样本,以提高算法的泛化能力(第 3 ~ 第 6 行)。算法 2 展示了生成单棵决策树的详细步骤,首先根据如下规则判断当前节点是否达到终止条件:1)分类树上节点所属记录的分类特征是否相同;2)回归树上节点误差是否小

于阈值;3)树是否达到最大高度 h。当满足终止条件以后,使用拉普拉斯机制返回添加了噪声的叶子节点(第 1 ~ 第 3 行),然后从特征集 A 中随机选择 K 个特征和分裂点,使用指数机制选择最佳分裂特征和分裂点(第 4 ~ 第 6 行)。本文指数机制的选择方法为:将算法 2 第 6 行计算出的各概率值按顺序排列放在坐标轴上,分别对应 0 ~ 1 上的某个区间,之后随机生成一个 0 ~ 1 之间的数,随机数对应哪个区间,指数机制就挑选相应备选项作为输出。最后,根据最佳分裂特征和分裂点递归构造左右子树,返回单棵决策树 T_i(第 7 ~ 第 9 行)。

2.2　算法隐私性分析

DiffPETs 算法首先将总隐私预算 P_ε 平均分配给每一棵决策树 $\varepsilon_1 = P_\varepsilon / t$。根据差分隐私的性质[22],由于单棵树的样本有交集,因此算法耗费的隐私预算为单棵决策树耗费隐私预算的总和,然后按照同样规则递归产生每一棵决策树。具体步骤如下:首先从 D 中有放回地抽取大小为 $|D|$ 的训练集 $D(i)$,随后将单棵树的隐私预算平均分配给各层 $\varepsilon_2 = \varepsilon_1/(2(h+1))$,每层的隐私预算平均分成两部分,一部分应用拉普拉斯机制对划分到该节点的样本数添加噪声,然后判断是否满足终止条件,若满足则将此节点标记为叶子节点,若没有达到终止条件,随机地从 $|A|$ 个特征集中选择 K 个特征,之后再随机产生 K 个分裂点,利用指数机制和剩余的隐私预算选择最佳分裂特征和分裂点。因此,算法所耗费的全部隐私预算是 P_ε,提供 ε-差分隐私保护。在分类中,指数机制中的可用性函数为基尼指数,而在回归中其可用性函数为方差。

2.3　敏感度分析

2.3.1　基尼指数的敏感度

本文进行如下的符号定义:$\tau = |D|$ 代表数据集 D 的记录数,特征集为 $A = \{A_1, A_2, \cdots, A_n\}$,标签集为 C,用 $T_c = \{(t \in D) \cap (c \in C)\}$ 表示某条记录属于标签 c,$T_j^A = \{(t \in D) \cap (A = A_j)\}$ 表示某条记录在特征 A 上的取值为 A_j,$T_{j,c}^A = \{(t \in D) \cap (A = A_j) \cap (c \in C)\}$ 表示记录 t 在特征 A 上的取值为 A_j 并且其标签为 c。算法需要判断用不同特征进行分类和对特征上不同分裂点进行划分的优劣,挑选基尼指数作为可用性函数,基尼指数是 CART 算法中的划分标准,具体如下:

$$\Delta G_{\text{Gini}}(A) = G_{\text{Gini}}(D) - G_{\text{Gini}}^A(D)$$

$$q_{\text{Gini}}(T,A) = -\sum_{j \in A} T_j^A \left(1 - \sum_{c \in C}\left(\frac{T_{j,c}^A}{T_j^A}\right)^2\right) \tag{5}$$

最小化基尼指数等价于最大化 $G_{\text{Gini}}^A(D)$,其敏感度 $\Delta q_{\text{Gini}} = 2$。使用信息增益(ID3 算法中使用的划分标准)作为指数机制的打分函数,其敏感度为 $\ln(|D|+1) + 1/\ln 2$。在隐私预算相同的情况下,敏感度越高,指数机制的效率越低。因此,本文挑选基尼系数作为指数机制的打分函数。

将 DiffPETs 算法应用到回归中,在选择最佳分裂特征和分裂点时,指数机制的可用性函数为方差。而在标记叶子节点时,需要对其均值添加由拉普拉斯分布产生的噪声来提供差分隐私保护。设 D 为给定的数据集,包括 n 条记录 $t_1, t_2, \cdots, t_n$,每条记录 t_i 有 $d+1$ 个特征值,即 $t_i = (x_{i1}, x_{i2}, \cdots, x_{id}, y_i)$,其中,$\sqrt{\sum_{i=1}^{d} x_{id}^2} \leqslant 1$,$y \in [0,1]$。本文数据集的组成形式如下:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} \tag{6}$$

2.3.2　均值的敏感度

由于回归树的均值主要用于划分叶子节点,因此本文考虑数据集的最后一列。根据相邻数据集的定义,假设最后一条数据不一样,均值的计算公式为 $M_{\text{mean}} = S_{\text{sum}}/N_{\text{num}}$,$S_{\text{sum}}$ 与 N_{num} 的敏感度均为 1。因此,分别添加噪声得到 S'_{sum} 和 N'_{num},均值的敏感度为 2。

2.3.3　方差的敏感度

决策树方差的定义如下:

$$V_{\text{var}} = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{n} y_j/n\right)^2 / n \tag{7}$$

分别计算分母和分子的敏感度。根据相邻数据集的定义,分母 n 的敏感度为 1。由于数据集都被归一化到 $[0,1]$ 之间,分子 $\left(y_i - \sum_{j=1}^{n} y_j/n\right)^2$ 的取值范围也在 $[0,1]$ 之间,因此,分子的敏感度也为 1,回归树方差的总敏感度为 2。

3　实验结果与分析

本文分类器训练算法、测试算法和数据处理均通过 Python3.5 版本实现,无差分隐私的算法调用 scikit-learn 库实现。硬件环境为 Intel(R) Core(TM)i5 2.5 GHz,内存为 8 GB 1 600 MHz DDR3。为检验 DiffPETs 算法的性能,采取如下 5 个 UCI 公开数据集进行实验:1)Mushroom 数据集,根据蘑菇的各种特征预测蘑菇是否可以食用;2)Nursery 数据集,根据家庭背景等特征预测申请幼儿园的等级;3)Congressional Vote Records(CongVote)数据集,美国国会投票记录,预测共和党或民主党是否能够当选;4)WineQuality 数据集,根据物理化学特性预测红葡萄酒的质量;5)Daily_Demand 数据集,预测巴西某一家物流公司订单的总量。表 1 所示为 5 个数据集的基本信息。

表 1 实验数据集信息

Table 1 Experimental dataset information

数据集名称	记录数	维度	任务
Mushroom	8 124	22	分类
Nursery	12 960	8	分类
CongVote	435	16	分类
WineQuality	4 898	12	回归
Daily_Demand	60	13	回归

3.1 分类实验结果

为验证 DiffPETs 算法在分类预测中的性能表现,分别令 $\varepsilon=\{0.50,0.75,1.00\}$(其对应森林中树的个数分别为 $t=\{10,10,5\}$),设置每棵树的高度为 $h=k/2$,k 为对应数据集的维度,可用性函数为基尼指数。本文首先利用 DiffPETs 算法在训练集上生成算法模型,之后在不同的隐私预算、树的棵数和单棵树的高度下进行实验,最后在测试集上作预测,并计算相应的准确率。每组实验进行 10 次,以平均实验结果作为最后的准确率,将 DiffPETs 算法与无差分隐私、RDT[28] 和 DiffPRF[29] 算法进行比较,实验结果如图 1 ~ 图 4 所示。

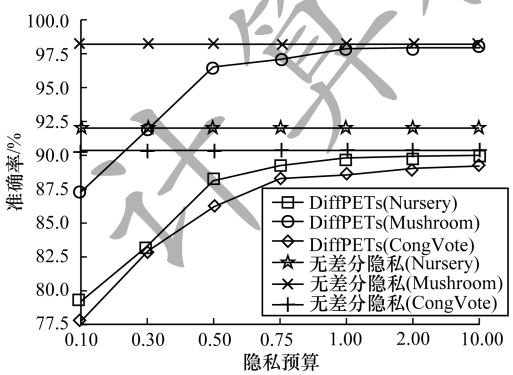

图 1 不同隐私预算下的分类结果

Fig. 1 Classification results under different privacy budgets

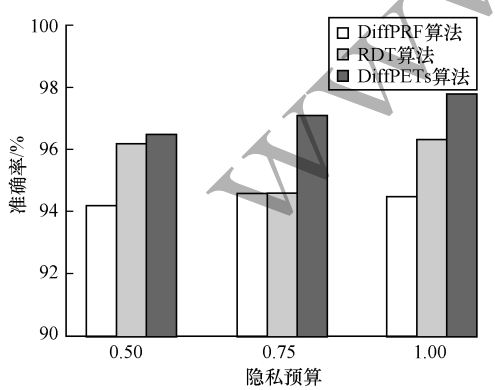

图 2 Mushroom 数据集在不同隐私预算下的实验结果

Fig. 2 Experimental results of Mushroom dataset under different privacy budgets

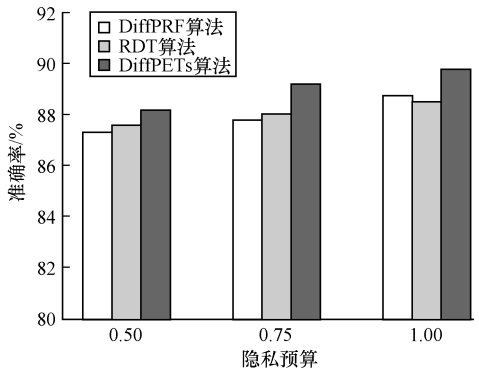

图 3 Nursery 数据集在不同隐私预算下的实验结果

Fig. 3 Experimental results of Nursery dataset under different privacy budgets

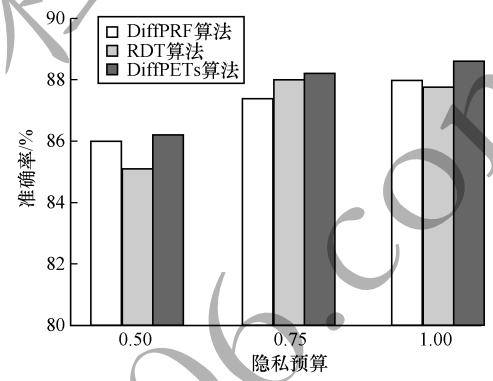

图 4 CongVote 数据集在不同隐私预算下的实验结果

Fig. 4 Experimental results of CongVote dataset under different privacy budgets

从图 1 可以看出,在隐私预算较小时,DiffPETs 算法的准确性不高,可以理解为为了实现高级别的隐私保护,算法不得不做出一些牺牲。随着隐私预算的增加,DiffPETs 算法的准确率逐渐提高,原因是添加到算法中的噪声量逐渐减少。通过与 DiffPRF 和 RDT 算法进行对比可以看出,本文算法在同等隐私预算下能够实现较低的分类误差。

3.2 回归实验结果

在回归实验中,本文采用均方误差(MSE)来判断算法的性能,其定义如下:

$$M_{\text{err}}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/n \tag{8}$$

其中,y_i 代表数据集中某条记录真实的值,$\hat{y}_i$ 代表根据某条记录的特征预测出的值,n 代表数据集中包含的记录数。

和分类实验一样,令 $\varepsilon=\{0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0\}$,森林中对应树的个数为 $t=10$。每组实验重复 10 次,采用最后的平均误差为实验结果,并将本文算法与无差分隐私、FM[33]、Diff_LR[34] 算法进行比较,实验结果如图 5、图 6 所示。

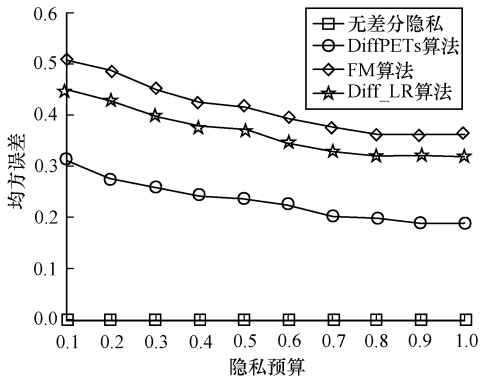

图 5　WineQuality 数据集的均方误差

Fig. 5　Mean square error of WineQuality dataset

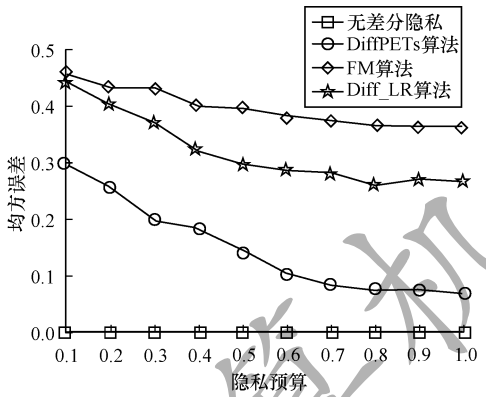

图 6　Daily_Demand 数据集的均方误差

Fig. 6　Mean square error of Daily_Demand dataset

从图 5、图 6 可以看出,在同等隐私预算 ε 的条件下,本文 DiffPETs 算法相比于 FM 和 Diff_LR 均方误差更小。随着 ε 的增加,DiffPETs 算法的均方误差与未加噪声的模型逐渐接近。此外,可以看出,没有添加噪声的模型产生的均方误差与隐私预算 ε 的大小无关。

4　结束语

本文提出一种基于 ExtraTrees 的差分隐私保护算法 DiffPETs,并将其应用于分类和回归分析。在决策树分类时,采取随机的方式选择特征使算法的泛化能力得到有效提升。在回归中,算法的全局敏感度与所添加的噪声均较低,回归预测的结果更准确。实验结果表明,DiffPETs 算法具有较高的可行性与准确性。下一步考虑将本文算法应用于交通、医疗等大数据问题。

参考文献

[1] SWEENEY L. k-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[2] MACHANAVAJJHALA A, KIFER D, GEHRKE J, et al. L-diversity: privacy beyond k-anonymity [J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 3.

[3] LI N H, LI T C, VENKATASUBRAMANIAN S. t-closeness: privacy beyond k-anonymity and l-diversity [C]// Proceedings of 2007 IEEE International Conference on Data Engineering. Washington D. C., USA: IEEE Press, 2007: 106-115.

[4] WONG R C W, LI J Y, FU A W C, et al. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing[C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2006: 754-759.

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

[6] XIONG Ping, ZHU Tianqing, WANG Xiaofeng. A survey on differential privacy and applications [J]. Chinese Journal of Computers, 2014, 37(1): 101-122. (in Chinese)
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.

[7] SU Sen, TANG Peng, CHENG Xiang, et al. Differentially private multi-party high-dimensional data publishing[C]// Proceedings of 2016 IEEE International Conference on Data Engineering. Washington D. C., USA: IEEE Computer Society, 2016: 205-216.

[8] ZHANG Xiaojian, MENG Xiaofeng. Streaming histogram publication method with differential privacy[J]. Journal of Software, 2016, 27(2): 381-393. (in Chinese)
张啸剑,孟小峰. 基于差分隐私的流式直方图发布方法[J]. 软件学报,2016,27(2):381-393.

[9] LI Wanjie, ZHANG Xing, CAO Guanghui, et al. Hierarchical data fusion publishing mechanism based on differential privacy protection [J]. Journal of Chinese Computer Systems, 2019, 40(10): 2252-2256. (in Chinese)
李万杰,张兴,曹光辉,等. 基于差分隐私保护的数据分级融合发布机制[J]. 小型微型计算机系统,2019,40(10):2252-2256.

[10] WANG Qian, ZHANG Yan, LU Xiao, et al. Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy[C]//Proceedings of the 35th Annual IEEE International Conference on Computer Communications. Washington D. C., USA: IEEE Press, 2016: 52-60.

[11] HUO Zheng, MENG Xiaofeng. A trajectory data publication method under differential privacy [J]. Chinese Journal of Computers, 2018, 41(2): 400-412. (in Chinese)
霍峥,孟小峰. 一种满足差分隐私的轨迹数据发布方法[J]. 计算机学报,2018,41(2):400-412.

[12] CUI Lei. Research on some key security issues for social network information dissemination [D]. Taiyuan: Taiyuan University of Technology, 2019. (in Chinese)

崔磊. 面向社交网络信息传播的若干关键安全问题研究[D]. 太原:太原理工大学,2019.

[13] SHI Xiujin, LI Hanyue. A social network publishing graph model based on differential privacy protection[J]. Intelligent Computer and Application, 2019, 9(1): 28-30,35. (in Chinese)
石秀金,李寒悦. 一种基于差分隐私保护的社交网络发布图模型[J]. 智能计算机与应用,2019,9(1):28-30,35.

[14] BUN M, STEINKE T, ULLMAN J. Make up your mind: the price of online queries in differential privacy[C]// Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms. New York, USA: ACM Press, 2017: 1306-1325.

[15] YUAN Ganzhao, YANG Yin, ZHANG Zhenjie, et al. Convex optimization for linear query processing under approximate differential privacy[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2016: 2005-2014.

[16] NI Lina, LI Chao, LIU Haoren, et al. Differential private preservation multi-core DBScan clustering for network user data[J]. Procedia Computer Science, 2018, 129: 257-262.

[17] LI Tong, LI Jin, LIU Zheli, et al. Differentially private naive Bayes learning over multiple data sources[J]. Information Sciences, 2018, 444: 89-104.

[18] QU Jingjing, CAI Ying, XIA Hongke. Summary of research on differential privacy protection for dynamic data publishing[J]. Journal of Beijing Information Science and Technology University, 2019, 34(6): 30-36. (in Chinese)
屈晶晶,蔡英,夏红科. 面向动态数据发布的差分隐私保护研究综述[J]. 北京信息科技大学学报(自然科学版),2019,34(6):30-36.

[19] LI Hongcheng, WU Xiaoping. Network intrusion correlation method with differential privacy protection of alerts sequence[J]. Computer Engineering, 2018, 44(5): 128-132. (in Chinese)
李洪成,吴晓平. 支持告警序列差分隐私保护的网络入侵关联方法[J]. 计算机工程,2018,44(5):128-132.

[20] MU Hairong, DING Liping, SONG Yuning, et al. DiffPRFs: random forest under differential privacy[J]. Journal on Communications, 2016, 37(9): 175-182. (in Chinese)
穆海蓉,丁丽萍,宋宇宁,等. DiffPRFs:一种面向随机森林的差分隐私保护算法[J]. 通信学报,2016,37(9):172-182.

[21] DWORK C. Calibrating noise to sensitivity in private data analysis[J]. Lecture Notes in Computer Science, 2012, 3876(8): 265-284.

[22] MCSHERRY F, TALWAR K. Mechanism design via differential privacy[C]//Proceedings of FOCS'07. Washington D. C., USA: IEEE Press, 2007: 94-103.

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

[24] BLUM A, DWORK C, MCSHERRY F, et al. Practical privacy: the SuLQ framework[C]//Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. New York, USA: ACM Press, 2005: 128-138.

[25] FRIEDMAN A, SCHUSTER A. Data mining with differential privacy[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2010: 493-502.

[26] MOHAMMED N, CHEN R, FUNG B, et al. Differentially private data release for data mining[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2011: 493-501.

[27] ZHU Tainqing, XIONG Ping, XIANG Yang, et al. An effective deferentially private data releasing algorithm for decision tree[C]//Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications. Washington D. C., USA: IEEE Press, 2013: 362-374.

[28] JAGANNATHAN G, PILLAIPAKKAMNATT K, WRIGHT R N. A practical differentially private random decision tree classifier[C]// Proceedings of ICDMW'09. Washington D. C., USA: IEEE Press, 2009: 114-121.

[29] PATIL A, SINGH S. Differential private random forest[C]// Proceedings of ICACCI'14. Washington D. C., USA: IEEE Press, 2014: 2623-2630.

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

[31] CHAUDHURI K, MONTELEONI C. Privacy-preserving logistic regression[C]//Proceedings of the 20th Annual Conference on Neural Information Processing Systems. Washington D. C., USA: IEEE Press, 2008: 289-296.

[32] CHAUDHURI K, MONTELEONI C, SARWATE A D. Differential private empirical risk minimization[J]. The Journal of Machine Learning Research, 2011, 12: 1069-1109.

[33] ZHANG Jun, ZHANG Zhenjie, XIAO Xiaokui, et al. Functional mechanism: regression analysis under differential privacy[J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1364-1375.

[34] ZHENG Jian, ZOU Hongzhen. Linear regression analysis algorithm of differential privacy budget allocation[J]. Computer Applications and Software, 2016, 33(3): 275-278. (in Chinese)
郑剑,邹鸿珍. 差异化隐私预算分配的线性回归分析算法[J]. 计算机应用与软件,2016,33(3):275-278.

编辑 吴云芳