差分隐私保护约束下集成分类算法的研究

贾俊杰[1], 邱万勇[1*], 马慧芳[1,2]

[1]西北师范大学计算机科学与工程学院 兰州 中国 730070
[2]桂林电子科技大学 广西可信软件重点实验室 桂林 中国 541004

摘要 机器学习中的隐私保护问题是目前信息安全领域的研究热点之一。针对隐私保护下的分类问题, 该文提出一种基于差分隐私保护的 AdaBoost 集成分类算法: CART-DPsAdaBoost(CART-Differential Privacy structure of AdaBoost)。算法在 Boosting 过程中结合 Bagging 的基本思想以增加采样本的多样性, 在基于随机子空间算法的特征扰动中利用指数机制选择连续特征分裂点, 利用 Gini 指数选择最佳离散特征, 构造 CART 提升树作为集成学习的基分类器, 并根据 Laplace 机制添加噪声。在整个算法过程中合理分配隐私预算以满足差分隐私保护需求。在实验中分析不同树深度下隐私水平对集成分类模型的影响并得出最优树深值和隐私预算域。相比同类算法, 该方法无需对数据进行离散化预处理, 用 Adult、Census Income 两个数据集实验结果表明, 模型在兼顾隐私性和可用性的同时具有较好的分类准确率。此外, 样本扰动和特征扰动两类随机性方案的引入能有效处理大规模、高维度数据分类问题。

关键词 隐私保护; 差分隐私; 机器学习; AdaBoost; CART 分类树
中图法分类号 TP309.2 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2021.07.07

Research on an Ensemble Classification Algorithm under Differential Privacy

JIA Junjie[1], QIU Wanyong[1*], MA Huifang[1,2]

[1]College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China
[2]Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China

Abstract In the field of information security, privacy protection based on machine learning is currently a hot topic. For classification issues under privacy protection, this paper proposes an AdaBoost ensemble classification algorithm based on differential privacy protection: CART-DPsAdaBoost (CART-Differential Privacy structure of AdaBoost). The algorithm combines the idea of Bagging in the Boosting process to increase the diversity of sampling. In the feature perturbation based on the random subspace algorithm, an exponential mechanism is used to select continuous attribute split points to construct a CART boosting tree as a base classifier for ensemble learning. And add noise according to the Laplace mechanism. In the whole algorithm process, the privacy budget is allocated reasonably to meet the differential privacy protection needs. In the experiment, the impact of the privacy level on the ensemble classification model under different tree depths is analyzed, and the optimal tree depth value and privacy budget domain are obtained. Compared with similar algorithms, this method does not need discretization preprocessing of data. The experimental results of Adult and Census Income show that the model has good classification accuracy while taking into account privacy and usability. Moreover, the introduction of two types of random schemes, sample perturbation and feature perturbation can effectively deal with large-scale and high-dimensional data classification problems.

Key words privacy protection; differential privacy; machine learning; AdaBoost; CART classification tree

1 引言

互联网“颠覆性”的发展, 使各种信息系统采集且积累了丰富的数据。机器学习即对海量的数据进行分析以获得各类可用的数据模型[1]。而数据集中通常都含有私有或敏感信息, 如医疗诊断信息、电子商务购物信息等, 由此引发人们对自身隐私泄露的担忧。

隐私保护技术[2-4]能有效处理数据模型发布当中个人隐私信息泄露问题。匿名技术依赖于攻击者的背景知识, 所提供的隐私是“无法保证的”。Dwork 等人[5]提出差分隐私定义, 解决了传统隐私

通讯作者: 邱万勇, 硕士生, Email: Qiuwy8023@163.com。
本课题得到国家自然科学基金项目(No.61967013), 甘肃省高等学校创新能力提升项目(No.2019A-006), 的资助。
收稿日期: 2020-10-18; 修改日期: 2020-12-18; 定稿日期: 2021-06-24

保护模型的两个缺陷: 1)差分隐私保护与背景知识无关; 2)差分隐私建立在严格的数学基础上, 对隐私保护提供了量化的评估方法。差分隐私的核心概念涵盖了从隐私保护领域到数据科学(如机器学习、数据挖掘、统计和学习理论)等领域的一系列研究[6-10]。

针对机器学习中数据模型发布、分析时的个人隐私泄露问题, 在特定算法中引入差分隐私, 保护数据私有或敏感信息的同时, 使得发布的模型发挥其最大的可用性[7-8]。差分隐私数据分析的基本任务是将现有的非隐私算法扩展到差分隐私算法。在集中学习模式下, 基于数据分析的差分隐私模型研究中, 由于模型本身会泄露训练数据中个体属性或敏感信息。对此, 需设计相应扰动机制使模型训练过程实现差分隐私保护。一般为输入扰动、目标扰动、梯度扰动和输出扰动等[8,11]。

差分隐私已被广泛应用于传统机器学习中, 如逻辑回归、支持向量机以及决策树[12-14]等分类模型, 并能较好的平衡可用性和隐私性。本文以决策树作为基分类器进行差分隐私约束下的集成分类模型研究, 表 1 汇总了差分隐私约束下基于决策树分类方法的相关研究[12-18]。在集成学习中, Patil 和 Singh[15]将差分隐私引入随机森林并提出 DiffPRF 算法, 其使用 ID3 作为基分类器且只能处理离散属性。穆海蓉等人[16]对其改进并提出 DiffPRFs 算法, 引入指数机制处理连续属性。差分隐私保护下的集成分类模型 DP-AdaBoost[18]只能处理离散属性的 ID3 决策树, 算法以决策树桩作为基分类器并在叶结点添加相应噪声, 没有考虑树深度与隐私保护下分类模型的关系。

表 1 差分隐私约束下基于决策树分类方法对比

Table 1 Comparison of classification methods based on decision tree under differential privacy constraints

方法	典型机制	具体方法	噪声	框架	数据类型	特点
SuLQ-based ID3	Laplace 机制	使用含 Laplace 噪声的计数值计算属性信息增益	高	接口模式	离散	噪声大, 数据可用性下降
PinQ-based ID3	Laplace 机制	利用 Partition 算子将数据集分割成不相交子集, 再实现 ID3 算法	高	接口模式	离散	提高了隐私预算利用率, 但未能降低噪声
DiffP-ID3	Laplace 机制 指数机制	利用指数机制划分属性	低	接口模式	离散	一次分裂只需消耗一次预算, 降低了噪声
DiffP-C4.5	Laplace 机制 指数机制	将指数机制扩展到连续属性	低	接口模式	离散 连续	使用两次指数机制, 隐私预算消耗过多
DiffGen	Laplace 机制 指数机制	先使用泛化技术之后结合指数机制和信息增益对属性进行分割	低	完全访问模式	离散 连续	在属性类型较少时, 隐私保护效果较好
DT-Diff	Laplace 机制 指数机制	在 DiffGen 基础上提出特征模型选择策略, 通过建立特征模型对样本分组添加噪声	低	完全访问模式	离散 连续	充分利用隐私预算, 提高了分类准确率
DiffPRF	Laplace 机制 指数机制	利用 ID3 决策树构建集成学习随机森林算法	低	接口模式	离散	利用随机森林解决维度问题, 需要预先将连续属性离散化
DiffPRFs	Laplace 机制 指数机制	改进 DiffPRF 算法, 将指数机制扩展到连续属性	低	接口模式	离散 连续	消除多维度大数据离散化预处理问题
DP-AdaBoost	Laplace 机制	使用单层 ID3 决策树构建自适应提升算法 AdaBoost	低	完全访问模式	离散	降低模型复杂度, 无需在属性划分时引入噪声, 提高了分类准确率

接口模式下, 数据挖掘工作者被视为是不可信的, 数据管理者不会发布原始数据集, 而只是提供访问接口。数据挖掘者只能获取差分隐私保护后的查询函数的结果。在这种模式下, 隐私保护的功能完全由接口来提供。

完全访问模式下, 数据挖掘工作者被认为是可信的, 能够直接访问数据集并对执行算法进行修改使其满足差分隐私保护的需求, 从而保证最终发布的模型不会泄露数据集中的隐私信息[7-9]。

本文在分析已有差分隐私保护决策树相关研究的基础上构建CART提升树, 提出一种基于差分隐私保护的 AdaBoost 集成分类模型 CART-DPsAdaBoost(CART-Differential Privacy structure of AdaBoost), 在树深度增加而模型复杂度增加的情况下, 分析不同隐私保护水平对集成模型分类

性能的影响, 通过合理分配利用隐私预算, 设计分类性能良好且安全有效的隐私保护集成模型。

本文的主要贡献包括 3 个方面:

1) 算法在 Boosting 过程中结合 Bagging 的思想和属性扰动中的随机子空间算法, 增加基分类器多样性的同时使迭代过程中每个元组被选中的概率由它的权重决定。若添加噪声大小小于样本和特征采样的随机性, 噪声对隐私的影响微乎其微。该方法对大数据集、高维度属性数据分类效率较好。

2) 算法构建 CART 提升树作为集成学习的基分类器, 在建树过程中利用指数机制选择连续特征分裂点, 利用 Gini 指数选择最佳分裂特征, 并保证隐私预算的合理分配与利用。

3) 算法无需对数据进行离散化预处理, 降低了分类系统性能的消耗。在满足差分隐私保护的需求下, 分析最佳参数域, 保持较高分类准确率, 保证分类模型的隐私性。

本文第 2 节描述差分隐私理论基础和改进后的 CART 提升树算法以及集成学习 AdaBoost 理论背景; 接下来, 第 3 节给出本文算法框架, 并详细讨论提出的 CART-DPsAdaBoost 算法, 对其隐私性进行合理证明与分析; 第 4 节提供完整的检验和讨论结果, 这些结果来自两个标准数据集上的一系列实验分析; 文章结论部分总结了该方法的优点和局限性, 并期待进一步的研究。

2 定义与理论基础

2.1 差分隐私理论基础

2.1.1 差分隐私定义

定义 1. ε-差分隐私[5-6]。给定相邻数据集 D 和 D' 至多相差一条记录, 即($|D\Delta D'|\leqslant 1$)。设有随机机制 M, $Range(M)$ 为 M 的值域, M 的任意输出结果 $O(O\in Range(M))$ 若满足下列不等式, 则 M 满足 ε-差分隐私:

$$Pr[M(D_1)\in O]\leqslant e^{\varepsilon}Pr[M(D_2)\in O], \quad (1)$$

其中 $Pr[\cdot]$ 表示隐私被披露的风险, 隐私预算 ε 也称隐私保护水平指数。ε 越小数据安全需求越高。

定义 2. 全局敏感度[5-6]。对于任意满足 $|D\Delta D'=1|$ 的相邻数据集 D 和 D', 给定查询函数 $f:D\to R^d$, 函数 f 的敏感度为:

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

其中 R 表示所映射的实数空间, d 表示 f 的查询维度, $\|f(D_1)-f(D_2)\|$ 为 $f(D_1)$ 和 $f(D_2)$ 之间 1-阶范数距离。

2.1.2 噪声机制

定义 3. Laplace 机制[5-6]. 给定任意函数 $f:D\to R^d$, 表达式 $K(D)$ 的输出满足下列等式, 则 $K(D)$ 满足 ε-差分隐私。

$$K(D)=f(D)+\left(\text{Laplace}\left(\frac{\Delta f}{\varepsilon}\right)\right)^d, \quad (3)$$

其中 $\text{Laplace}\left(\frac{\Delta f}{\varepsilon}\right)$ 是服从尺度参数为 $\frac{\Delta f}{\varepsilon}$ 的 Laplace 分布, 噪声量与 Δf 和 ε 取值相关。

Laplace 机制通过 Laplace 分布产生的噪声, 扰动真实输出值来实现差分隐私保护, Laplace 机制仅适用于对数值结果的保护。对于非数值型数据, 例如实体对象, McSherry 等人[6,11]提出了指数机制。

定义 4. 指数机制[5-6]. 给定一个效用函数 $q:(D\times O)\to r(r\in Range)$, 函数 $F(D,q)$ 满足下列等式, 则 $F(D,q)$ 满足 ε-差分隐私。

$$F(D,q)=\{r:Pr[r\in O]\}\propto\exp\left(\frac{\varepsilon q(D,r)}{2\Delta q}\right), \quad (4)$$

其中输入为数据集 D, 输出为实体对象 r, Δq 是效用函数 $q(D,r)$ 的全局敏感度。函数 F 以正比于 $\exp\left(\frac{\varepsilon q(D,r)}{2\Delta q}\right)$ 的概率从 *Range* 中选择并输出 r。

2.1.3 相关性质

性质 1. 序列组合性[5-6]。设有算法 $M_1, M_2, \cdots, M_n$ 其隐私预算分别为 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$, 对于同一数据集 D, 由这些算法构成的组合算法 $M(M_1(D), M_2(D), \cdots, M_n(D))$ 提供 $\left(\sum_{i=1}^{n}\varepsilon_i\right)$-差分隐私保护。

性质 2. 并行组合性[5-6]。设有算法 $M_1, M_2, \cdots, M_n$ 其隐私预算分别为 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$, 对于不相交的数据集 $D_1, D_2, \cdots, D_n$, 由这些算法构成的组合算法 $M(M_1(D_1), M_2(D_2), \cdots, M_n(D_n))$ 提供 $(\max\varepsilon_i)$-差分隐私保护。

2.2 分类与回归树

构建 CART 提升树

CART(Classification and Regression Tree)分类

与回归树[1,14]。本文以增加模型的多样性来提高基分类器一定程度的准确率, 使基分类器“好而不同”, 对此引入样本扰动和特征扰动两类随机性方案, 即在集成算法迭代的过程中引入自适应采样方案, 以及在特征选择中使用随机子空间算法来构建 CART 提升树。穆海蓉等人[16]提出一种基于差分隐私的随机森林算法 DiffPRFs, 在树的构建过程中采用指数机制选择分裂点和分裂特征, 并根据 Laplace 机制添加噪声。该算法虽无需对数据进行离散化预处理, 但每次迭代需调用两次指数机制, 导致大量隐私预算消耗。

本文以改进后的 CART 提升树作为集成学习的基分类器。对于连续特征使用指数机制选择分裂点, 使用 Gini 指数选择分裂特征, 迭代中只需调用一次指数机制, 整个算法保证了隐私预算的充分利用。

假设有样本集 $D(i)$, 样本数 N, 属性集合 $\Lambda(A\in\Lambda)$, 类标签集 $Y(\forall y\in Y)$, 分类类别数 k, 实体对象 r, 记录属于类别 $k(k\in K)$ 的概率 Pr_k, 概率分布的 Gini 值为 $Gini(Pr)=\sum_{k=1}^{K}Pr_k\left(1-Pr_k\right)=1-\sum_{k=1}^{K}Pr_k^2$, 则本文 CART 提升树的构建如算法 1 所示。

算法 1. $Build\_DPTree\left(D(i),\Lambda,\varepsilon_c,d\right)$

输入: $D(i)=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$; $x_i\in\Lambda$; $y_i\in\{-1,1\}$;树深度 d;隐私预算 ε_c

输出: 添加噪声的基分类器 $g_t(x)$

1. REPEAT

2. $\varepsilon'=\dfrac{\varepsilon_c}{d+1}$

3. 调用随机子空间算法并依据 Gini 指数最小化原则,从属性集 $\overline{\Lambda}=\{\Lambda_1,\cdots,\Lambda_k\}$ 中选取属性 A 划分;

4. IF 节点达到终止条件:THEN

令其为叶结点,对该结点实例数添加 Laplace 噪声 $N_{\tilde{D}_i}=NoisyCount\left(\left|\tilde{D}_i\right|\right)$,噪声量为 $\text{Laplace}\left(\dfrac{\max(\omega_t,i)}{(1/2)\varepsilon'}\right)$;

5. 对当前叶结点分类:

$\forall y\in Y$:

$D_y=Partition\left(\tilde{D}_i,\forall y\in[Y]:\ y=y_i\right)$,类计数加噪 $N_y=NoisyCount\left(\left|D_y\right|\right)$,噪声量为 $\text{Laplace}\left(\dfrac{\max(\omega_t,i)}{(1/2)\varepsilon'}\right)$;

6. ELSE 节点未达到中止条件:THEN

划分当前分裂节点,添加噪声量为 $\text{Laplace}\left(\dfrac{\max(\omega_t,i)}{\varepsilon'}\right)$;

7. IF 子空间 $\overline{\Lambda}$ 包含 n 个连续属性,执行步骤 8;

8. $\varepsilon''=\dfrac{\varepsilon'}{2(n+1)}$

9. 用以下概率选择每个连续属性分裂点:

$$\frac{\exp\left(\dfrac{\varepsilon''}{2\Delta q}q(D_i,A)\right)|R_i|}{\sum_i\exp\left(\dfrac{\varepsilon''}{2\Delta q}q(D_i,A)\right)|R_i|}$$

/*其中 $q(D_i,A)$ 为效用性函数, Δq 为效用性函数的全局敏感度, $|R_i|$ 为区间集合的大小*/

10. END IF

11. 将数据集 $D(i)$ 划分为两个子集 $D_l(i)$ 和 $D_r(i)$;

12. 建立左右子树:

$t_l=Build\_DPTree\left(D_l(i),\Lambda,\varepsilon_{D_l},d+1\right)$

$t_r=Build\_DPTree\left(D_r(i),\Lambda,\varepsilon_{D_r},d+1\right)$

13. END IF

14. UNTIL 节点达到中止条件即全部记录标签一致;达到最大深度 d;隐私预算耗尽;

15. RETURN $g_t(x)$

2.3 集成学习之 AdaBoost

AdaBoost[1,19]核心思想是针对同一数据分布将逐步强化生成的多个弱分类器进行集成, 最终构成一个分类效果较好的强分类器。而强化的过程即算法核心步骤如下所述: 给定训练集, 初始化样本分布, 调用弱学习算法获得基分类器, 根据其误差率调整训练样本权重。基于改变的样本分布, 经过多次迭代, 得到一组具有互补性的基分类器并将其线性组合成强分类器, 以提高集成模型的准确性和稳定性。具体如算法 2 所示。

算法 2．AdaBoost

输入： $\tilde{D}=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$；基学习算法 $\mathfrak{L}$；训练轮数 $\tilde{T}$；

1. 初始化样本权重分布: $\tilde{D}_1(x)=1/n$；

2. FOR $t=1$ TO $\tilde{T}$ DO

3. 调用算法 $\mathfrak{L}$ 生成基分类器: $\tilde{g}_t=\mathfrak{L}(\tilde{D},\tilde{D}_t)$；

4. 估计 $\tilde{g}_t$ 的误差: $\tau_t=P_{x\sim\tilde{D}_t}(\tilde{g}_t(x)\neq y_i)$；

5. IF $\tau_t>0.5$ THEN

6. break;

7. END IF

8. 确定 $\tilde{g}_t$ 的权重: $\tilde{\alpha}_t=\frac{1}{2}\ln\left(\frac{1-\tau_t}{\tau_t}\right)$；

9. 更新样本分布:

$$\tilde{D}_{t+1}(x)=\frac{\tilde{D}_t(x)}{Z_t}\times\begin{cases}\exp(-\tilde{\alpha}_t), & \text{if } \tilde{g}_t(x)=y_i\\ \exp(\tilde{\alpha}_t), & \text{if } \tilde{g}_t(x)\neq y_i\end{cases}=\frac{\tilde{D}_t(x)\exp(-\tilde{\alpha}_t y_i\tilde{g}_t(x))}{Z_t}$$

其中 Z_t 是规范化因子,以确保 $\tilde{D}_{t+1}$ 是一个分布;

10. END FOR

输出: $\tilde{G}(x)=\mathrm{sign}\left(\sum_{t=1}^{\tilde{T}}\tilde{\alpha}_t\tilde{g}_t(x)\right)$

3 CART-DPsAdaBoost 算法

发布最终的集成分类模型，要确保基分类器即决策树的结构不被泄露，因为树结构含有核心的分类规则和结果。假设数据分析者对数据集有完全访问的权限，则可以在基分类器的生成过程加入噪声。即在构建 CART 提升树过程中向分裂节点及叶结点添加相应的 Laplace 机制噪声，并合理分配隐私预算来控制噪声量大小，最后进行隐私性的合理证明和适用性分析，以完善算法性能评估，分析隐私保护与模型效用之间的平衡。

3.1 算法与模型

给定算法的迭代次数为 T，总的隐私预算为 ε_p，每迭代一次生成新的基分类器。

步骤1，初始化数据样本权重分布。

步骤2，递归的使用采样策略，并根据权值分布找出最大的权重值作为敏感度参数。调用算法1在生成基分类器的过程中，根据噪声公式计算每个分裂节点和叶结点所需的噪声值并加入到基分类器中，得到满足差分隐私的基分类器。

步骤3，对生成的基分类器即决策树进行剪枝操作，计算被剪枝子树的隐私预算总量，并将其添加到被剪枝子树的父节点中。

步骤4，计算错误分类的数据比例，并根据误差率计算该基分类器在总体分类器中对应的权重系数。

步骤5，依据误差率更新数据样本的权值分布。

步骤6，判断当前迭代次数是否已满足设定值，若满足则终止，否则执行下一步。

步骤7，返回步骤2，继续构建新的基分类器。

步骤8，将所得到的基分类器进行线性组合，得到最终满足差分隐私保护的集成分类器。

通过上述步骤生成的强分类器即为带差分隐私保护的集成分类算法的分类模型，可以直接发布给数据挖掘工作者而不用担心隐私的泄露。

假定初始训练集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，属性集合 Λ，类标签集 Y，数据集大小 n，初始化样本权重分布 $D_1=(\omega_{1,1},\cdots,\omega_{1,i},\cdots,\omega_{1,n}),\omega_{1,i}=\frac{1}{n}$，$i=1,\cdots,n$。

算法 3．CART-DPsAdaBoost

输入： $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$；$x_i\in\Lambda$；$y_i\in\{-1,1\}$；迭代次数 T；隐私预算 ε_p；

输出: 满足 ε_p -差分隐私保护的集成分类器 $G(x)$;

1. 初始化权重分布: $D_1(i)=\frac{1}{n},i\in[1,n]$；

2. 损失函数: $E(g(x),y,i)=\exp(-y_i g(x_i))$；

3. $\varepsilon_c=\frac{\varepsilon_p}{T}-\log(\alpha_t)$

4. FOR $t=1$ TO T DO

5. 从 D 中采样选取大小为 $|D|$ 的训练集 $D(i)$;

6. 调用算法 1 生成基分类器 $g_t(x)$

$g_t(x)=Build\_DPTree(D(i),\Lambda,\varepsilon_c,d)$;

7. 对生成的决策树进行剪枝:

$g_t'(x)=CART\_Prun(g_t(x))$;

8. 计算被剪枝子树的隐私预算总量,并将其添加到被剪枝的子树的父节点中;

9. 计算 $g_t'(x)$ 的误差率:

$\tau_t = \sum \omega_t(i) I\left[g_t'(x_i) \neq y_i \right], i \in [1,n]$;

10. IF $\tau_t > 0.5$ THEN

11. break;

12. END IF

13. 计算 $g_t'(x)$ 的权重系数: $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\tau_t}{\tau_t}\right)$;

14. 更新样本权重分布:

$$D_{t+1}(i) = \frac{D_t(i)\exp\left(-\alpha_t y_i g_t'(x_i)\right)}{Z_t}$$

其中 Z_t 是规范化因子 $Z_t = \sum_{i=1}^{n} D_t(i)\exp\cdot\left(-\alpha_t y_i g_t'(x_i)\right)$;

15. END FOR

16. RETURN $G(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t g_t'(x)\right)$

通过 Laplace 噪声机制对隐私信息进行保护, 噪声量大小由隐私预算 ε 来控制.计算噪声时需考虑每一条记录权重值 $\omega_t(i)$, 使得差分隐私噪声计算公式中函数敏感度动态变化[18], 根据敏感度公式得出 $\Delta f = \max(\omega_t, i)$, 同时 Laplace 机制噪声公式转化为 $K(D) = f(x) + \left(\text{Laplace}(\Delta f/\varepsilon)\right)^d$。算法中指数机制的效用函数使用 Gini 指数.最小化 Gini 指数等价于最大化 $Gini_A(D)$, 其敏感度 $\Delta q_{Gini} = 2$, 选择 Gini 指数作为效用函数敏感度较低使得指数机制的效率有所增高[12]。本文预先指定迭代次数, 在隐私预算分配中同时考虑每个弱分类器的权重值 α_t, 即每个弱分类器分配到的预算为 $\frac{\varepsilon_p}{T} - \log(\alpha_t)$。

3.2 分类

对新样本集中的每一条记录, 应用最终的集成模型进行分类。最终分类结果是加入差分隐私噪声的每个基分类器结果与权重的乘积线性组合。基分类器的构建包含建树和剪枝两步骤。算法的集成分类器简化模型如图 1 所示。对新样本分类过程见算法 4。

算法 4. 使用 $G(x)$ 对新样本分类

输入: 预测集 $\hat{M}$;满足 ε_p -差分隐私的集成分类器 $G(x)$

输出: 预测集中样本的分类结果

1. REPEAT

2. 对每条待分类样本;

3. FOR $i = 1$ TO m DO

4. 从当前树的根节点开始, 根据当前节点类型, 判断进入哪个子节点, 直至到达叶子结点;

5. 得到当前树的预测结果与其权重值相乘;

6. 线性组合每棵树的结果: $\text{sign}\left(\sum_{t=1}^{T} \alpha_t g_t'(x)\right)$;

7. END FOR

8. UNTIL 到达叶子结点

9. RETURN 输出样本集 $\hat{M}$ 的分类结果 $Y(\hat{M})$

通过本文算法构建满足隐私保护的集成分类器模型 $G(x)$, 可将该分类模型直接发布给数据挖掘工作者而不必担心数据集中的个体信息会被泄露。

3.3 算法的性能

隐私性

集成分类器生成过程中, 预先定义了算法的迭代次数, 故采用层次均分的方法[16-20]。首先将总的隐私预算 ε_p 平均分配给每一棵树 $\varepsilon_c = \frac{\varepsilon_p}{T} - \log(\alpha_t)$。基于 Bagging 思想采样的每棵树样本集有交集, 根据差分隐私性质 1, 总的隐私预算为每棵树消耗预算的叠加。对于每一棵树将隐私预算平均分配到每一层 $\varepsilon' = \frac{\varepsilon_c}{d+1}$, 由于每层节点样本之间不相交, 根据差分隐私性质 2, 每个节点预算为 $\varepsilon' = \frac{\varepsilon_c}{d+1}$。用每个节点预算的一半 $\frac{\varepsilon'}{2}$ 来估计该节点实例数, 之后判断是否到达终止条件, 若满足则标记为叶结点并用另一半预算确定叶结点类计数。若当前节点有 n 个连续属性, 则将另一半的预算均分为 $n+1$ 份, 用于选择每个连续属性的分裂点, 每次指数机制消耗预算为 $\frac{\varepsilon'}{2(n+1)}$, 由序列组合性可知, 多次指数机制消耗的

错误!

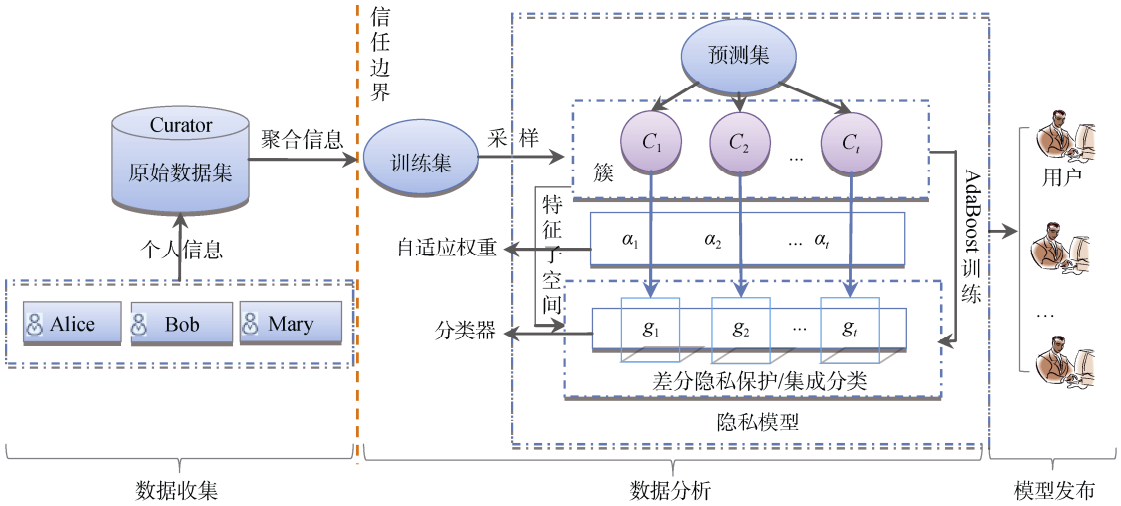


图 1　差分隐私保护下的集成分类模型结构图

Figure 1　Structure diagram of ensemble classification model under differential privacy protection

预算为各次的叠加。

性质 3．$G(x)$ 提供 ε_p -差分隐私保护。

证明:

假设相邻数据集 D 和 D' $(|D\Delta D'=1|)$，$M(D)$ 和 $M(D')$ 分别表示随机算法 M 的输出结果，总的隐私预算为 ε_p，基分类器 $g_t(x)$ 的权重为 α_t，基分类器数量为 T。

1) 对于每个连续属性有 r_i 种划分方法，r_i 被以指数机制选中的概率为 $p(r_i)=\dfrac{E(D,r_i)}{\sum\limits_{r_i\in Range}E(D,r_i)}$。

其中 $E(D,r_i)$ 表示 $\exp\left(\dfrac{\hat{\varepsilon}q(D,r_i)}{2\Delta q}\right)$，以 $p(r_i)$ 为权重，则连续属性划分方案 r_i 以正比于 $p(r_i)E(D,r_i)$ 的概率参与全局选择。

2) 对于离散属性使用 Gini 指数划分，若属性 A 有 v 个值，则有 $\dfrac{2^v-2}{2}$ 个划分子集，以属性 A 划分所计算 Gini 指标值为:

$$Gini(D,A)=\frac{N_1}{N}Gini(D_1)+\frac{N_2}{N}Gini(D_2), \tag{5}$$

其中 N_1、N_2 为不相交的子集，根据差分隐私性质 1 得 $\hat{\varepsilon}=\varepsilon_c/(d+1)\left|\dfrac{2^v-2}{2}\right|$。则 Gini 指数的差分隐私预算可根据性质 2 转化为:

$$\begin{aligned}\frac{prob(M(D)=r_i)}{prob(M(D')=r_i)}&=\frac{\prod_{i=1}^{\left|\frac{2^v-2}{2}\right|}p(r_i)E(D,r_i)}{\prod_{i=1}^{\left|\frac{2^v-2}{2}\right|}p(r_i)E(D',r_i)}\\&\leqslant\prod_{i=1}^{\left|\frac{2^v-2}{2}\right|}\left(e^{\frac{\hat{\varepsilon}}{2}}\right)^2\frac{e^{\frac{-\hat{\varepsilon}}{2}}\sum\limits_{r_i\in Range}\exp\left(\frac{\hat{\varepsilon}q(D,r_i)}{2\Delta q}\right)}{\sum\limits_{r_i\in Range}\exp\left(\frac{\hat{\varepsilon}q(D,r_i)}{2\Delta q}\right)}\\&=\prod_{i=1}^{\left|\frac{2^v-2}{2}\right|}e^{\frac{\varepsilon_c}{2(d+1)\left|\frac{2^v-2}{2}\right|}}\\&=e^{\frac{\varepsilon_c}{d+1}}\end{aligned} \tag{6}$$

根据差分隐私性质 1，每棵树的隐私保护程度为:

$$\prod_{i=1}^{|d+1|}e^{\frac{\varepsilon_c}{d+1}}=e^{\varepsilon_c} \tag{7}$$

因此，提供 ε_c -差分隐私保护。

3) 对于集成分类模型 $G(x)$，总的差分隐私预算满足:

$$\begin{aligned}\frac{Prob(M(D)=R)}{Prob(M(D')=R)}&\leqslant\frac{\prod_{i=1}^{T}\alpha_t\exp\left(\frac{q(D,r_i)}{2\Delta q}\right)\cdot\left(\frac{\varepsilon_p}{T}-\log\alpha_t\right)}{\prod_{i=1}^{T}\alpha_t\exp\left(\frac{q(D',r_i)}{2\Delta q}\right)\cdot\left(\frac{\varepsilon_p}{T}-\log\alpha_t\right)}\\&=\prod_{i=1}^{T}\left(\alpha_t\exp\left(\frac{\varepsilon_p}{T}-\log\alpha_t\right)\cdot\left(\left|q(D,r_i)-q(D',r_i)\right|\right)\right)\\&=e^{\sum_{i=1}^{T}\left(\log\alpha_t+\frac{\varepsilon_p}{T}-\log\alpha_t\right)}\\&=e^{\varepsilon_p}\end{aligned} \tag{8}$$

因此，$G(x)$ 提供 ε_p -差分隐私保护。得证。

4 实验结果与分析

4.1 实验数据

分类器训练、预测及数据预处理均使用python3.7实现，编辑器为PyCharm。采用UCI Machine Learning Repository中的Adult数据集设计对照实验，并在Census Income大规模、高维度数据集上验证算法的有效性。表2展示了相关数据集的基本信息。

Adult数据集包含6个连续属性、8个离散属性，类别属性为income lever，类别值为“<=50K”“>50K”；Adult_Train共32561个元组，其中70%作为训练集，30%作为测试集对训练模型进行剪枝操作；Adult_Predict为预测数据集其包含16281个元组。Census Income共41个属性，其中8个数值属性，33个离散属性，Census Income_Train共199523个元组，Census Income_Predict共99763个元组。为检验算法的有效性，设置$\varepsilon=(0.05,0.1,0.25,0.5,0.75,1)$，决策树数量$T=10$，每棵树的深度$d=(2,3,4,5,6)$，效用函数使用Gini指数，随机子空间算法中节点分裂时随机选取的特征数为$k=5$进行多组实验。

表2 实验数据集

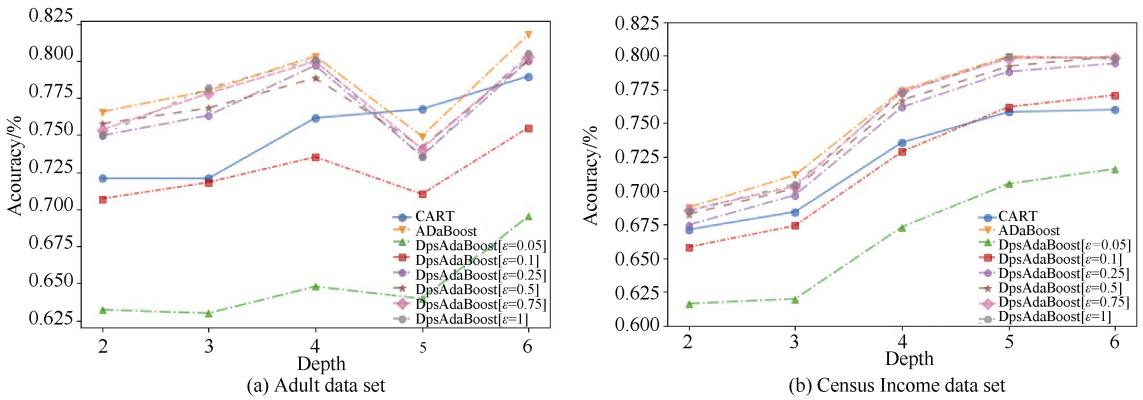
Table 2 Experimental data set

Data set	Number of samples	Number of features	Task
Adult_Train	32561	14	classification
Adult_Predict	16281	14	classification
Census Income_Train	199523	41	classification
Census Income_Predict	99763	41	classification

4.2 实验结果

1) 噪声对模型分类准确率的影响

由定义3可知，Laplace分布的尺度参数为$\Delta f/\varepsilon$时，噪声大小由函数敏感度Δf和隐私预算ε共同决定。Δf根据样本权重更新而动态变化，ε与噪声量成反比。要降低噪声量，则Δf尽量小以及ε分配较多。图2(a-b)所示树深度d不变时，ε取值增大分类准确率总体趋势逐渐提高。$\varepsilon\in[0.25,1]$时两个数据集上噪声对模型准确率的影响波动在某一较小且可接受的范围内。$\varepsilon\in[0.75,1]$当噪声大小小于样本采样的随机性时，噪声对分类性能的影响微乎其微。以上结果表明在添加一定量的噪声时该集成模型仍能保持较好的分类性能。



(a) Adult data set (b) Census Income data set

图2 不同隐私水平下随树深度的增加分类准确率变化情况

Figure 2 Changes in classification accuracy with increasing tree depth under different privacy levels

2) 树深度对模型分类准确率的影响

在隐私预算ε不变的情况下，随着树深度的增加，集成模型的准确率呈现上升趋势，这是因为树越深建立的规则越精确分类误差便降低。当然，随着树深度的增加，集成模型复杂度也随之增大，图2(a)在Adult数据集上当$d>4$时，模型已发生过拟合。实验以样本量与模型分类准确率变化情况做进一步解析。

3) 两个数据集上模型最优树深值的分析

图3(a-f)实验中取值为$\varepsilon=1, T=10$。一个共同趋势是样本量较少时训练集模型准确率较高，测试集模型准确率低，随着样本量增加初始的过拟合现象逐渐

好转, 模型准确率逐渐趋于平稳。此外, 两个数据集上随树深度的增加分类模型波动随之增大, 这是由于模型越复杂所添加的噪声越多干扰越明显。Census Income 数据集上的噪声干扰总体上较大, 这是由于该数据集属性维度较高, 建树相比较复杂从而使得噪声的影响更加显著。

如图 3(a~c)在 Adult 数据集上, 模型准确率从 35%上升到了 81%左右, 当隐私预算越大即 $\varepsilon \in [0.75,1]$时越趋近于未加噪的模型准确率。对于 Census Income 高维度数据集从 60%上升到了 80%左右, 由于该数据集特征数量较多模型更易于学习, 400 左右样本量已达到模型的最优值, 准确率为 80%左右。该算法在大规模、高维度数据集上表现更佳, 时间效率更好, 具有较好的数据扩展性。

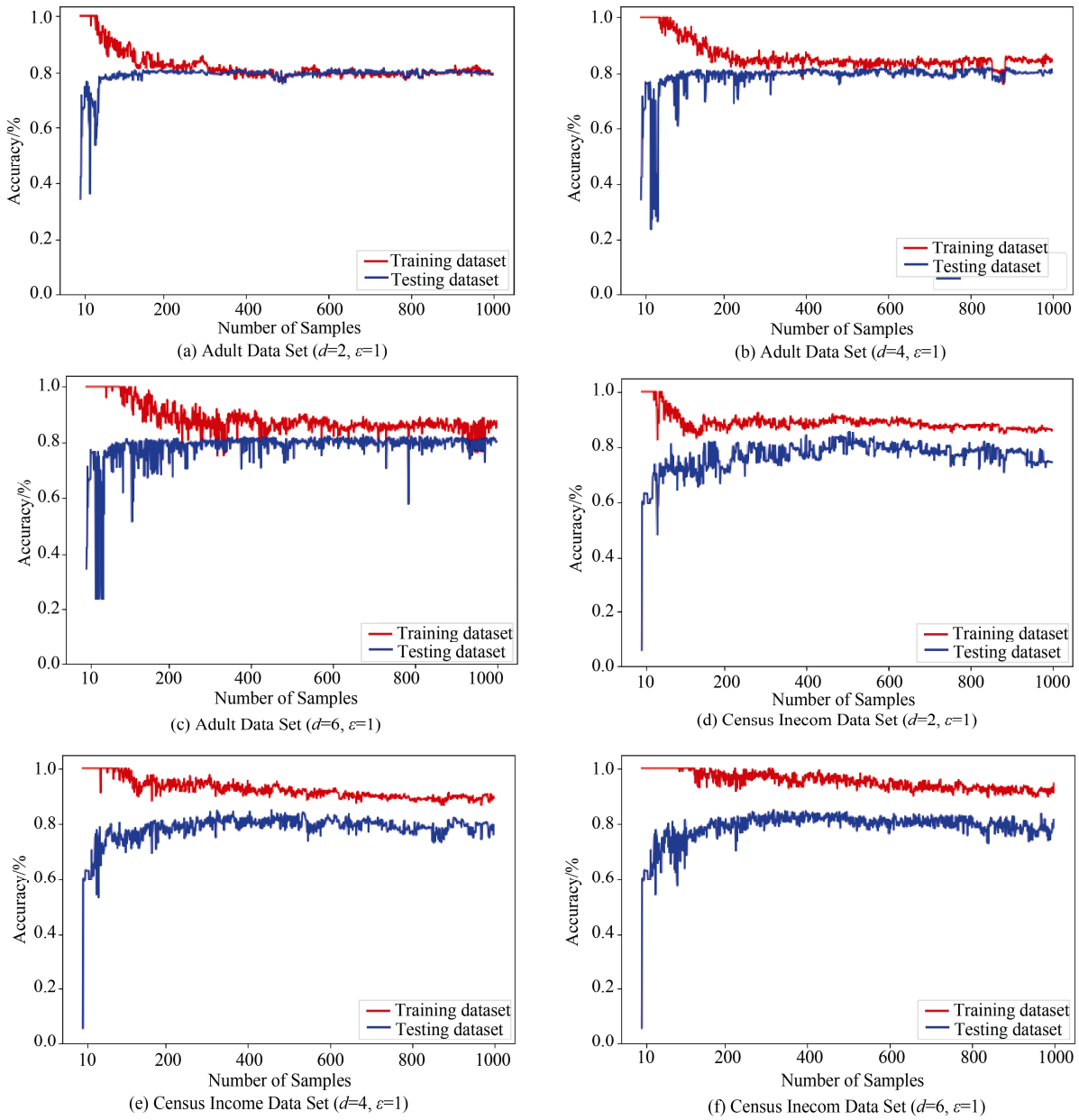


图 3　两个数据集上不同树深度的模型在训练集和测试集上样本量与模型准确率的关系

Figure 3　The relationship between the sample size of the models with different tree depths on the training set and the test set and the accuracy of the model

综上实验进一步分析, 在 Adult 数据集上, 模型分类准确率稳定在 81%左右。而由图 2(a)实验数据分析可知, 当隐私预算不变时 $d=4$ 模型准确率达到最佳, 为 0.8081813156441。在 Census Income 数据集上, 准确率稳定在 80%左右, 由图 2(b)实验数据可知, $d=5$ 模型准确率达到最佳,为 0.7993123634249。

4.2.1　评估

F-Measure[1,19]是用来评价分类器性能的常用指

标, 综合衡量模型的精准度和召回率, 定义如下:

$$Precision = \frac{TP}{TP + FP}, \tag{9}$$

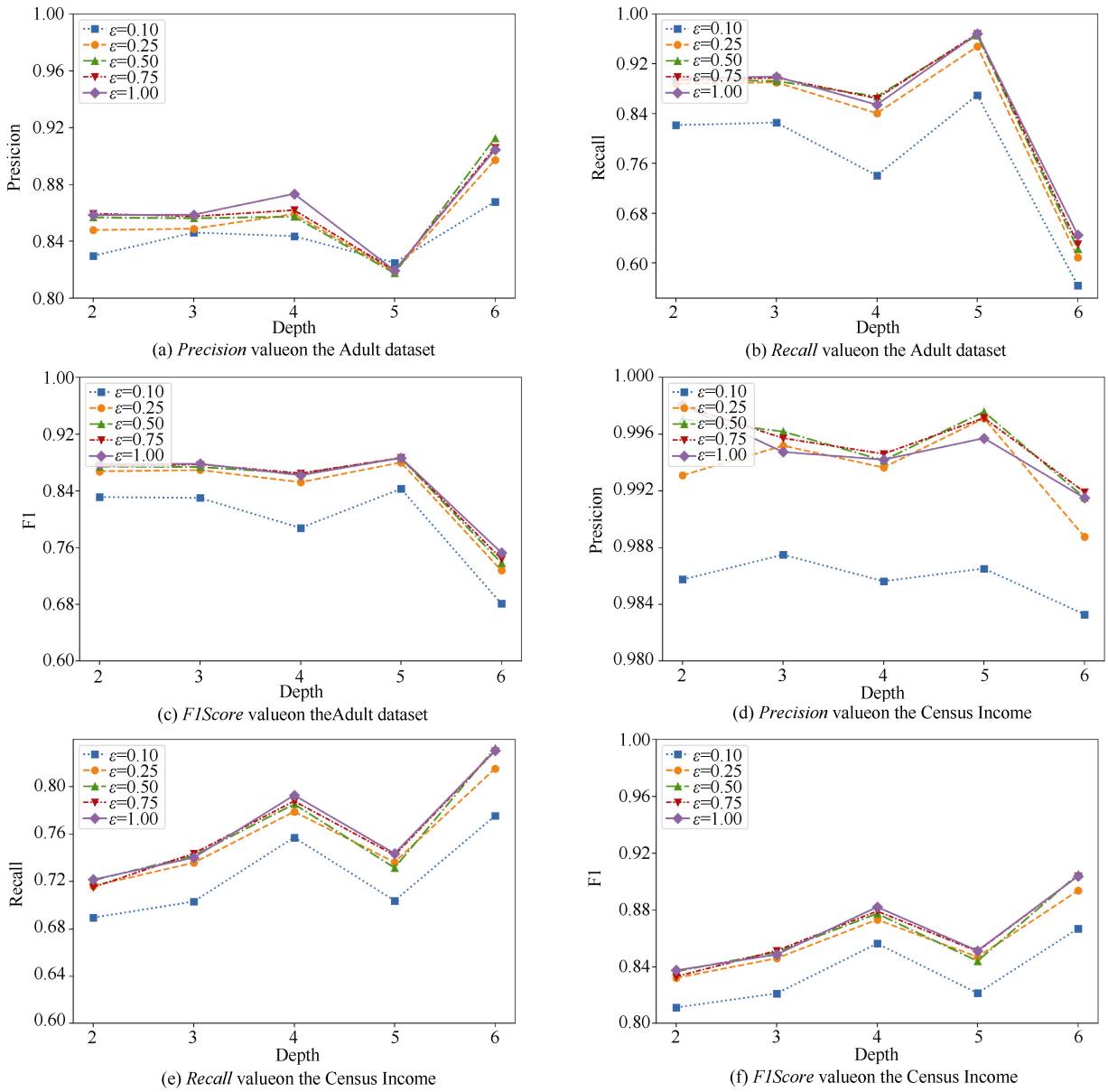
$$Recall = \frac{TP}{TP + FN}, \tag{10}$$

$$F\text{-}Measure = \frac{\left(1+\beta^2\right)\cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \tag{11}$$

其中 β 为调节 *Precision* 和 *Recall* 相对重要性的系数(取 1 即为 *F1Score*)。*F1Score* 的取值范围为[0,1], 该值越大算法可用性越好。

同时使用 *ROC*[13-14,19](Receiver Operating Characteristic)曲线下的 *AUC*(Area Under Curve)值来评估分类器泛化性能。该值越大分类器泛化性能越高, *AUC* 取值范围为[0,1]。两个数据集上实验结果如图 4-5 以及表 3-4 所示。

树深度设定为 $d=(2,3,4,5,6)$, 分配的隐私预算为 ε=(0.1,0.25,0.5,0.75,1)。图 4(a~c)为 Adult 数据集上的实验结果, 图 4(a)当 d 固定 ε 越大 *Precision* 值也随着增大。此外, 如图 4(a)(d)所示, Adult 数据集上 $d=4$ 时 *Precision* 最佳, Census Income 数据集上 $d=5$ 时 *Precision* 最佳。如图 4(b)(e)所示为模型在数据集上的召回率变化情况, 随着 d 的变化模型



(a) *Precision* valueon the Adult dataset

(b) *Recall* valueon the Adult dataset

(c) *F1Score* valueon theAdult dataset

(d) *Precision* valueon the Census Income

(e) *Recall* valueon the Census Income

(f) *F1Score* valueon the Census Income

图 4　不同隐私水平下随树深度的增加 *Precision*/*Recall*/*F1Score* 值的变化情况

Figure 4　The change of *Precision/Recall/F1Score* with the increase of tree depth at different privacy levels

精准度和召回率互相影响, *Precision* 值大时 *Recall* 较小, 反之亦然。图 4(c)(f)为 *F1Score* 值的变化情况, *F1Score* 综合衡量了 *Precision* 和 *Recall* 的结果, $\varepsilon \geqslant 0.25$ 时, *F1Score* 值在 86%上下波动, 模型的可用性较好。如图 5(a) Adult 数据集实验结果显示 $\varepsilon \in [0.1, 0.5]$ *AUC* 变化波动较大, 这是由于 Adult 数据集特征数量较少, 调节隐私水平对 *AUC* 产生的扰动较大。而图 5(b)所示在 Census Income 特征维度高的数据集上 *AUC* 表现较好。

4.2.2 对比

由于同类算法 DP-AdaBoost 是树深度为 1 的集成分类模型, 相关对比实验设定 $\varepsilon = 0.05, 0.1, 0.25, 0.5, 0.75, 1$, $T = 10$, $d = 1$, 在不同隐私水平下, 模型分类准确率比较结果如图 5(c)所示。

表 3 随隐私水平 ε 的变化 *Precision*, *Recall*, *F1Scores* 值的变化情况

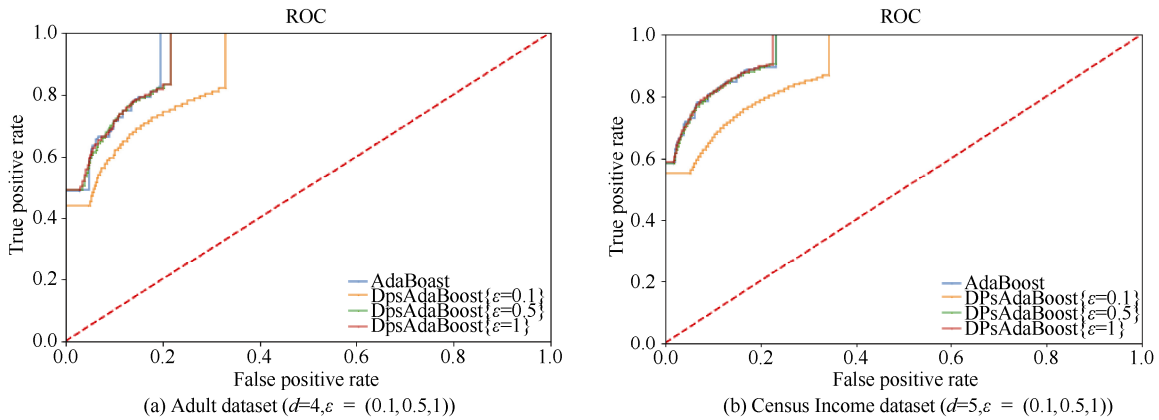
Table 3 How *Precision*, *Recall*, and *F1Scores* vary with ε

Value of ε	Adult (d=4)			Census Income (d=5)		
	Precision	*Recall*	*F1Score*	*Precision*	*Recall*	*F1Score*
0.1	0.84468	0.72529	0.78045	0.99106	0.71032	0.82752
0.25	0.85507	0.83650	0.84568	0.99468	0.73238	0.84361
0.5	0.86235	0.86885	0.86559	0.99711	0.73736	0.84779
0.75	0.86325	0.85866	0.86095	0.99618	0.74270	0.85096
1	0.86813	0.85467	0.86135	0.99714	0.74590	0.85342

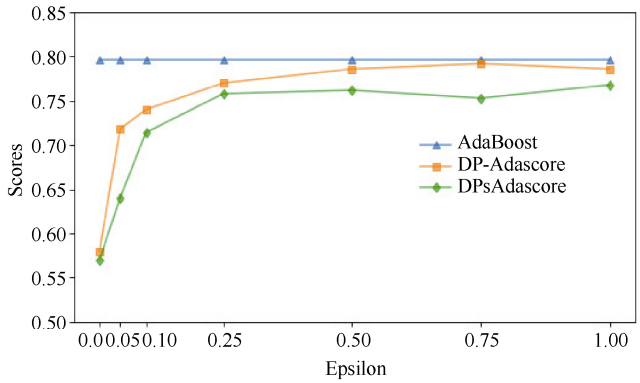
表 4 *AUC* 性能对比

Table 4 *AUC* performance comparison

Data Set	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.25$	$\varepsilon = 0.5$	$\varepsilon = 0.75$	$\varepsilon = 1$
Adult (d=4)	0.64267	0.73323	0.74343	0.82035	0.87207	0.87240
Census Income (d=5)	0.68524	0.89724	0.90804	0.91011	0.91414	0.91441



(a) Adult dataset (d=4,ε = (0.1, 0.5, 1)) (b) Census Income dataset (d=5,ε = (0.1, 0.5, 1))



(c) Adult数据集上d=1时同类算法对比

图 5 模型的 *ROC* 曲线部分图示以及同类算法对比

Figure 5 Part of the model's *ROC* curve diagram and comparison of similar algorithms

当$d=1$时, 相比 DP-AdaBoost 同类算法旨在叶结点添加相应的噪声, 本文算法在树的构建过程中对分裂节点和叶结点同时添加相应的 Laplace 噪声, 树深度较低时对模型分类性能影响较大, 即以牺牲一定的模型准确率来换取更佳的数据隐私性。同时本文在研究不同树深度下隐私水平对分类模型的影响时, 通过相关实验结果表明, 算法在隐私保护策略下选择最优树深及隐私域时模型可以达到较好的分类效果。在分类树构建过程中消耗了一部分隐私预算用于处理连续特征, 但无需对连续特征进行离散化预处理, 一定程度上提高了分类效率。

5 结束语

在已有差分隐私决策树各类算法基础上, 研究了差分隐私保护下 AdaBoost 集成分类算法, 保护数据隐私信息的同时保持集成模型较高的分类性能。实验表明, 在集成当中提高了隐私预算的利用和算法的执行效率, 能有效处理大规模、高维度数据分类问题。此外由于其可扩展性, 该方法将可以适用于大数据环境.存在的不足之处: (1)隐私保护水平采用隐私预算ε定量分析, 但实际算法和应用中的取值缺乏一个标准; (2)通常差分隐私假设数据记录相互独立, 但在实际中数据库的记录可能存在关联性, 会增加隐私泄露的风险。在接下来的工作中, 会继续对算法做进一步优化, 提升泛化性能。该研究的实验数据集和程序代码将在GitHub平台更新。

致　谢　本课题得到国家自然科学基金项目(No.61967013), 甘肃省高等学校创新能力提升项目(No.2019A-006), 的资助, 在此表示感谢。

参考文献

[1] Zhou Z H. Machine learning[M]. Beijing, Tsinghua University Press, 2016.
(周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.)

[2] Wu Y J. Privacy Preserving Data Publishing: Models and Algorithms[M]. Beijing, Tsinghua University Press, 2016.
(吴英杰. 隐私保护数据发布: 模型与算法[M]. 北京: 清华大学出版社, 2016.)

[3] Ding W X, Yan Z, Deng R H. Privacy-Preserving Data Processing with Flexible Access Control[J]. *IEEE Transactions on Dependable and Secure Computing*, 2020, 17(2): 363-376.

[4] Ding W X, Hu R, Yan Z, et al. An Extended Framework of Privacy-Preserving Computation with Flexible Access Control[J]. *IEEE Transactions on Network and Service Management*, 2020, 17(2): 918-930.

[5] Dwork C .Differential Privacy[J]. *Lecture Notes in Computer Science*, 2006, 26(2):1-12.

[6] Dwork C, Feldman V, Hardt M, et al. The Reusable Holdout: Preserving Validity in Adaptive Data Analysis[J]. *Science*, 2015, 349(6248): 636-638.

[7] Zhu T Q, Li G, Zhou W L, et al. Differentially Private Data Publishing and Analysis: A Survey[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 29(8): 1619-1638.

[8] Wang A, Wang C, Bi M, et al. A Review of Privacy-Preserving Machine Learning Classification[C]. *International Conference on Cloud Computing and Security*. Springer, Cham, 2018: 671-682.

[9] Zhao L C, Ni L H, Hu S S, et al. InPrivate Digging: Enabling Tree-Based Distributed Data Mining with Differential Privacy[C]. *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018: 2087-2095.

[10] Liu D, Yan Z, Ding W X, et al. A Survey on Secure Data Analytics in Edge Computing[J]. *IEEE Internet of Things Journal*, 2019, 6(3): 4946-4967.

[11] Kairouz P, Oh S, Viswanath P. The Composition Theorem for Differential Privacy[J]. *IEEE Transactions on Information Theory*, 2017, 63(6): 4037-4049.

[12] Liu X Q, Li Q M, Li T, et al. Differentially Private Classification with Decision Tree Ensemble[J]. *Applied Soft Computing*, 2018, 62: 807-816.

[13] Fletcher S, Islam M Z. Decision Tree Classification with Differential Privacy[J]. *ACM Computing Surveys*, 2019, 52(4): 1-33.

[14] Li Q B, Wu Z M, Wen Z Y, et al. Privacy-Preserving Gradient Boosting Decision Trees[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(1): 784-791.

[15] Patil A, Singh S. Differential Private Random Forest[C]. *2014 International Conference on Advances in Computing, Communications and Informatics*, 2014: 2623-2630.

[16] Mu H R, Ding L P, Song Y N, et al. DiffPRFs: Random Forest under Differential Privacy[J]. *Journal on Communications*, 2016, 37(9): 175-182.
(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): 175-182.)

[17] Lv C, Li Q M, Long H Q, et al. A Differential Privacy Random Forest Method of Privacy Protection in Cloud[C]. *2019 IEEE International Conference on Computational Science and Engineering and IEEE International Conference on Embedded and Ubiquitous Computing*, 2019: 470-475.

[18] Shen S Q. *Research on the Differential Privacy Classification Al-*

gorithms[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2017.
(沈思倩. 关于差分隐私保护分类算法的研究[D]. 南京: 南京航空航天大学, 2017.)

[19] Xiang T, Li Y, Li X G, et al. Collaborative Ensemble Learning under Differential Privacy[J]. *Web Intelligence*, 2018, 16(1): 73-87.

[20] Yan Z, Ding W X, Niemi V, et al. Two Schemes of Privacy-Preserving Trust Evaluation[J]. *Future Generation Computer Systems*, 2016, 62: 175-189.

贾俊杰 于 2009 年在长安大学获得博士学位, 硕士生导师, 现任西北师范大学计算机科学与工程学院副教授。主要研究方向为数据安全与隐私保护。E-mail: 654205992 @qq.com

马慧芳 于 2010 年在中国科学院计算技术研究所获得工学博士学位, 硕士生导师, 现任西北师范大学计算机科学与工程学院教授。主要研究方向为数据挖掘与机器学习。E-mail: mahuifang@yeah.net

邱万勇 曾就职于蓝山软件有限责任公司担任 JAVA 工程师, 参与多个项目研发, 中国计算机协会(CCF)会员, 现于西北师范大学计算机科学与工程学院攻读硕士学位, 主要研究方向为机器学习与隐私保护。E-mail: qiuwy8023@163. com