•信息工程•

DOI:10.15961/j.jsuese.201801059

基于等差隐私预算分配的大数据决策树算法

尚　涛[1],赵　铮[2],舒王伟[2],刘建伟[1]

(1.北京航空航天大学 网络空间安全学院,北京 100083;2.北京航空航天大学 电子信息工程学院,北京 100083)

摘　要:针对传统差分隐私保护方案以剩余隐私预算的一半逐层分配,即等比分配隐私预算,被应用于决策树时,随着决策树高度的增加,分配至顶层的隐私预算过小,随机噪声过大,分类准确率受到影响的问题,作者提出以差分隐私保护结合主流决策树C4.5分类方法为基本思路,依据决策树高度等差分配隐私预算的方案。差分隐私中的Laplace机制和指数机制确保决策树分类的安全性。作者利用大数据Hadoop平台的MapReduce框架,主程序进行MapReduce参数配置以及外层循环。在执行到每一个节点时,主程序将数据集属性的统计任务交给Mapper类,Reducer类接收Mapper类的统计结果并利用Laplace机制添加随机噪声,加噪结果返回主程序中作为计算信息增益率的参数。主程序利用指数机制选择最佳细分方案,递归过程直至样本数为0时停止。实验采用UCI数据库的car数据集进行测试,在不同隐私预算下将等比分配与等差分配两种方案得到的分类结果准确率进行对比。实验结果表明:本文算法在可接受的分类准确率降低的情况下满足差分隐私保护;与传统隐私预算分配相比,本文算法在相同隐私预算下提高了分类准确率;对于car数据集,本文算法在隐私预算为0.7或0.8时可较好兼顾数据集的安全性和有效性。因此,在一定程度上依据决策树高度等差分配隐私预算的方案可改善分类准确率,可实际应用于决策树分类算法。

关键词:分类;决策树;差分隐私;大数据

中图分类号:TP309.2　　**文献标志码**:A　　**文章编号**:2096-3246(2019)02-0130-07

Big Data Decision Tree Algorithm Based on Equal-arrival Privacy Budget Allocation

SHANG Tao[1], *ZHAO Zheng*[2], *SHU Wangwei*[2], *LIU Jianwei*[1]

(1.School of Cyber Sci. and Technol., Beihang Univ., Beijng 100083, China; 2.School of Electronic and Info. Eng., Beihang Univ., Beijing 100083, China)

Abstract: In order to address the problem that the traditional differential privacy preservation scheme is distributed layer by layer by half of the remaining privacy budget, i.e., equal ratio allocation of privacy budget, and when it is applied to the decision tree, the privacy budget allocated to the top layer is too small, the random noise is too large, and the classification accuracy is affected with the increase of the height of decision tree, a scheme of equal-arrival privacy budget allocation based on decision tree height difference was proposed, which combined differential privacy protection preservation with mainstream decision tree C4.5 classification algorithm. The Laplace mechanism and the exponential mechanism can ensure the security of the decision tree. This scheme utilized the MapReduce framework of the big data Hadoop platform, and the main program performed parameter configuration of MapReduce and outer loop. When executed to each node, the main program passed the statistical task of the dataset attribute to the Mapper class. The Reducer class received the statistical result of the Mapper class. The Laplace mechanism was used to add random noise. The noise-added result was returned to the main program for calculation the information gain rate. The main program used the exponential mechanism to select the best subdivision scheme. The recursion process stopped until the number of samples was 0. The experiment used the car data set of UCI database to test, and compared classification results of two schemes under different privacy budgets. Experiment showed that this scheme can satisfy differential privacy with acceptable classification accuracy reduction, and improve the classification accuracy under the same privacy budget compared to the traditional privacy budget allocation. For the car data set, the algorithm can balance the security and effectiveness of the data set when the privacy budget was 0.7 or 0.8. Therefore, the scheme of equal-arrival privacy budget allocation based on

收稿日期:2018 – 09 – 25

基金项目:国家重点研发计划资助项目(2016YFC1000307)

作者简介:尚　涛(1976—),男,副教授,博士.研究方向:网络安全. E-mail: shangtao@buaa.edu.cn

网络出版时间:2019 – 03 – 13 10 : 54 : 23　　**网络出版地址**: http://kns.cnki.net/kcms/detail/51.1773.TB.20190312.2209.005.html

decision tree height difference can improve classification accuracy to a certain extent. It can be practically applied to decision tree classification.

Key words: classification; decision tree; differential privacy; big data

数据挖掘[1]为政府、企业等从大数据中获取有价值信息提供了有效的技术手段。分类技术是数据挖掘中的一个重要分支。决策树[2]作为分类技术的重要方法之一, 从根节点以递归原理逐层构建分类规则, 依据分类规则将未知类别的数据划分到已知的类标签集合中。目前数据挖掘中很多用户的医疗、个人信息等隐私数据没有得到正当使用, 安全问题越来越突出。用户对个人隐私泄露的担忧在一定程度上制约数据挖掘技术的发展。因此, 发展数据挖掘技术的同时应减少隐私泄露风险。

近年, 学者们提出许多应用于分类技术的隐私保护方法, 其中以决策树为基础的方法包括: Du等[3]提出的随机响应技术, 其局限性在于仅仅适用于布尔属性的数据集, 将属性采用二进制标记后再进行随机化处理。Szűcs[4]提出的随机响应森林, 结合决策树与随机化技术, 对经典随机森林进行扩展, 用于处理隐私保护数据挖掘; 随机相应森林使用二进制匿名度量将属性和名称转化为二进制。但其局限性在于二进制匿名度量与分类准确性有较大的依赖程度, 需要权衡取值。Tai等[5]提出隐私保护决策树评估协议, 利用决策树的结构避免在决策树深度的指数级加密, 显著提高效率。但其局限在于该协议对于深度大但稀疏的决策树有效。Dwork等[6]提出的差分隐私保护技术, 其与攻击者的背景知识无关, 并且具有严格的数学基础, 可为不同的参数提供量化评估方法。因此, 差分隐私受到数据挖掘领域研究者的广泛认可, 常被用于决策树研究, 比如, 陈煜等[7]提出基于决策树的隐私保护数据流分类算法(PPFDT), 通过阈值算法找到扰动数据流的最佳分裂属性和最佳分裂点, 直接在扰动数据集上建立决策树。但其不足在于随机噪声添加使用传统差分隐私的隐私预算按照等比数列的方式进行均匀分配[8], 当决策树高度较高时, 顶层节点分配的隐私预算就会极少, 导致分类准确率的下降。

作者在研究当前隐私保护构建决策树算法的基础上, 分析传统隐私预算分配方案, 提出依据决策树高度分配隐私预算的方案, 利用差分隐私建立有效的决策树隐私保护模型, 提高分类准确率。利用大数据平台Hadoop的MapReduce并行框架实现方案, 分析其性能和安全程度并与传统隐私分配方案进行对比。

1　相关知识

1.1　决策树C4.5算法

目前, 较为成熟的决策树构建算法主要有3种, 分别是ID3(iterative dichotomiser 3)算法、C4.5算法、CART算法[9]。3种算法主要区别是分裂节点的标准不同。

ID3算法递归时在高层节点处较易选择属性取值较多的属性。在极端情况下会产生一个测试属性极多但是深度较浅的树。CART算法采用Gini系数最小化准则作为特征选取的依据。CART最终生成二叉树, 分类结果较好, 但稳定性较差, 与类似方法构建的决策树差异较大。本文方案的设计主要基于分类效率和准确率均较高的C4.5算法。

C4.5算法在ID3算法基础上增加了对连续属性、属性值空缺情况的处理。其基本思想是: 假设数据集D是训练样本集, 构建决策树时选取信息增益率最大的属性作为分裂节点, 依据此标准可以将D分为N个子集。若第i个子集D_i内包含的元组类别一致时, 该节点作为决策树的叶子节点停止分裂。其余情况均按照上述分类标准依次递归, 直到所有子集内的元组均属于一个类别为止。具体原理如下:

定义1(训练集D类别信息熵[10])　设D有d个样本, 将训练集分为m个类, 第i类的样本数为d_i, 概率p_i为d_i/d, 则类别信息熵为:

$$Info(D) = -\sum_{i=1}^{m} p_i \mathrm{lb}(p_i) \tag{1}$$

定义2(某属性划分子集信息熵[10])　假设选择属性A划分训练集D。$Value(A)$为属性A的取值集合, V是A的其中一个属性值, 由A划分子集的信息熵可由式(2)、(3)得出:

$$Info_A(D) = -\sum_{Value(A)} \frac{|D_V|}{d} Info(D_V) \tag{2}$$

$$Info(D_V) = -\sum_{i=1}^{m} p_{iV} \mathrm{lb}(p_{iV}) \tag{3}$$

式中, D_V为D中属性A的值为V的样本集合, $|D_V|$为D_V中所含样本数, p_{iV}为D_V中样本为第i类的概率。

定义3(划分属性A的信息增益[10])

$$Gain(A, D) = Info(D) - Info_A(D) \tag{4}$$

定义4(属性A分裂信息熵[10])　假设以属性A的取值集合对样本进行分割, 则分类信息熵为:

$$Info(A) = -\sum_{Value(A)} \frac{|D_V|}{d} \mathrm{lb}\left(\frac{|D_V|}{d}\right) \tag{5}$$

定义5(划分属性A的信息增益率[10])

$$Gain\text{-}Ratio(A)=\frac{Gain(A,D)}{Info(A)} \tag{6}$$

C4.5算法每一个节点下的分支都是由该属性的离散值数目决定，生成的决策树为规则较乱的多叉树。

1.2 差分隐私模型

差分隐私假设了攻击者可能具有的最大背景知识。其定义如下：

定义6（差分隐私[6]） 假设D和D'为邻近数据集，即数据记录相差至多为1。设$Range(M)$为一个随机函数M的取值范围，$\Pr(Es)$为随机事件Es的披露风险。对于任意的$S\in Range(M)$，满足不等式：

$$\Pr[M(D)\in S]\le \exp(\varepsilon)\times\Pr[M(D')\in S] \tag{7}$$

则M提供了ε–差分隐私保护。ε为隐私保护预算，用于评价隐私保护水平，ε值越大，隐私保护水平越高。

定义7（全局敏感度[11]） 设查询函数$f:D\to R^d$，输入数据集D，输出d维实数向量，对于任意两个邻近数据集D和D'：

$$\Delta f=\max_{D_1,D_2}\|f(D)-f(D')\|_1 \tag{8}$$

式中，Δf即为查询函数的全局敏感度。

差分隐私保护通过加入随机噪声使数据失真。噪声机制是差分隐私实现的关键，Laplace噪声机制和指数机制是常用的噪声机制。

Laplace噪声机制通过向算法响应值中加入服从Laplace分布的随机噪声，使干扰后的结果满足ε–差分隐私保护。若添加的噪声服从Laplace分布，其概率密度函数为：

$$p(x)=\frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) \tag{9}$$

式中：μ为位置参数，一般设置$\mu=0$；$b(b>0)$为尺度参数，可记为$x\sim Lap(\mu,b)$

定义8（Laplace噪声机制[12]） 假设查询函数为f，$f(D)$是函数f在数据集D上的查询结果，向$f(D)$中加入随机噪声$Y\sim Lap(\Delta f/\varepsilon)$，此处一般设位置参数$\mu=0$，尺度参数$b=\Delta f/\varepsilon$。查询函数最终得到的响应结果为$f(D)+Y$，满足$\varepsilon$–差分隐私保护。

定义9（指数机制） 假设输入原始数据集D，在随机算法M的作用下输出实体对象$r\in Range$，$f(D,r)$为可用性函数。若算法M以正比于$\exp(\varepsilon f(D,r)/2\Delta f)$的概率从$Range$中选择并输出结果$r$，则称算法$M$提供$\varepsilon$–差分隐私保护。上述结果可表示为：

$$\{M(D,f)=r:|\Pr[r\in Range]\}\propto \exp(\varepsilon f(D,r)/2\Delta f) \tag{10}$$

McSherry等[13]研究得出，若隐私预算ε增长，查询函数返回正确值的概率随之增加，表明隐私保护水平降低。

1.3 MapReduce机制与HDFS数据管理

MapReduce和HDFS是Hadoop的两个重要组成部分[14]。MapReduce是一种并行数据处理模型，适应于大数据处理。在MapReduce模型中，数据处理过程分为Mapper和Reducer两个过程。Mapper任务读取HDFS中对应文件，每行以(k,val)的形式作为中间输出，并传递给Reducer。Reducer任务接收来自每个映射器的输出，对输入数据进行按照键值排序，合并相同键的值，将结果汇总到Namenode输出[15]。MapReduce数据处理过程如图1所示。

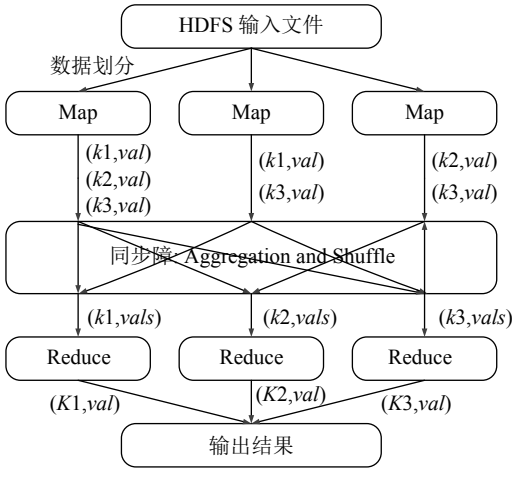

图 1 MapReduce并行数据处理

Fig. 1 Parallel data processing for MapReduce

HDFS是Hadoop的分布式文件系统。在HDFS中文件被分割为多个文件块（默认以64 MB为单位），每个文件块被复制分配到不同的存储节点DataNode上存放，即使某个节点出现故障或者中断连接，HDFS也可从其他节点中恢复出完整文件。

2 等差隐私预算大数据决策树算法

本文方案以差分隐私结合决策树C4.5算法为基础，作者提出等差隐私预算分配改善分类准确率，利用Laplace噪声机制、指数机制添加随机噪声及选择最佳细分方案以实现ε–差分隐私保护，并借助MapRedcue框架实现算法。

2.1 等差隐私预算分配

传统隐私预算分配方案以剩余隐私预算的一半逐层分配[16]，随决策树高度的增加，分配到高层节点的隐私预算以指数速度减少，导致高层节点的噪声量大，分类准确率下降。传统隐私预算分配方式如图2所示。

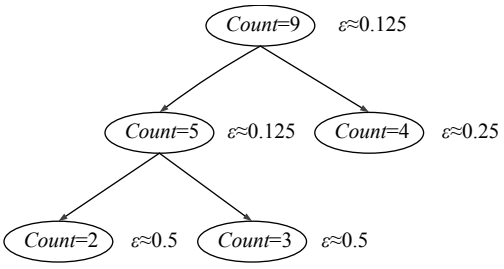

图 2　传统隐私预算分配示意图

Fig. 2　Schematic diagram of traditional privacy budget allocation

在决策树的递归过程中，等差分配按照等差数列的方式添加噪声。决策树的每次递归代表决策树某一层的全部节点，依据递归次数和分支数，隐私预算分配方式为相同层节点隐私预算相同，层与层节点之间隐私预算为等差数列。

假设算法整体隐私预算为ε，数据发布系统需有$\varepsilon/2$的隐私预算用于数据发布，因此用于添加Laplace噪声的隐私预算$\varepsilon_1=\varepsilon/2$。每一层分配的隐私预算为$a_0=d=2\varepsilon_1/[h(h+1)]$的等差数列。假设当前决策树的高度为$i$，分配给当前高度的隐私预算为$\varepsilon'=2i\varepsilon_1/[h(h+1)]$。该隐私预算分配方式可满足$\varepsilon$–差分隐私保护以确保安全性，并减少添加噪声量，提高分类准确率。等差分配隐私预算示意图如图3所示。

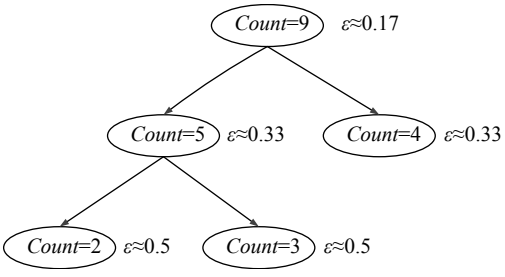

图 3　等差隐私预算分配示意图

Fig. 3　Schematic diagram of equal difference privacy budget allocation

2.2　Laplace噪声生成算法

Laplace噪声本质是生成一个概率密度函数服从Laplace分布的随机数，可利用Laplace概率累计函数的反函数式(11)得到随机数。

$$x=\begin{cases}\mu+b\ln(1+2\xi),\xi<0\\ \mu-b\ln(1-2\xi),\xi>0\end{cases}=\mu-b\times\mathrm{sgn}(\xi)\times\ln(1-2|\xi|)\qquad(11)$$

生成随机数的算法GenLaplaceNoise流程如下：

输入：本层隐私预算ε'，全局敏感度Δf，待加噪数据X；

输出：加噪数据Y。

Step 1：if($\varepsilon'>\varepsilon$)表示隐私预算不足

$noise=0$，return $Y=X$；

Step 2：else $\xi=random[-0.5,0.5]$，

$b=\Delta f/\varepsilon'=1/\varepsilon'$，

$noise=-b\times\mathrm{sgn}(\xi)\times\ln(1-2|\xi|)$；

return $Y=X+noise$。

每次节点进行分裂时，均调用GenLaplaceNoise算法。

2.3　基于差分隐私的决策树C4.5算法

在决策树构建算法中，每个属性都可作为决策数分裂的节点，因此有多种细分方案的选择。指数机制可评估查询函数f作用于每种细分方案返回值正确的概率是否与$\exp(\varepsilon f(D,r)/2\Delta f)$成正比，验证方案是否满足$\varepsilon$–差分隐私保护。指数机制的查询函数设为式(6)，即构建决策树时以信息增益率最大的方案作为最佳细分方案。本文算法基本流程如下：

输入：训练集D；

输出：决策树。

Step 1：训练集D单个节点作为根节点建树。

Step 2：若训练集D中的样本为同一类别，则记为叶子节点并标记类标签；否则，执行Step 3。

Step 3：若剩余属性为空或属性中剩余样本数小于给定的阈值，标记节点为D中样本数最多的类；否则，执行Step 4。

Step 4：根据指数机制，计算待选取属性的信息增益率，选择最大信息增益率作为当前节点的分裂属性，其中根节点的测试属性信息增益率最高。

Step 5：每次递归噪声为等差分配，调用GenLaplaceNoise算法对Step 4或Step 5中选择的属性值添加噪声。

Step 6：对于每一个新的节点，若节点所对应的样本为空，则标记为数据集中样本数最多的类，并停止迭代；否则，在该节点上递归执行该算法，继续进行分裂。

2.4　基于MapReduce的差分隐私决策树C4.5算法

主程序进行MapReduce参数配置以及外层循环。在执行到每一个节点时，主程序将数据集属性的统计任务交给MapReduce，并将执行结果作为参数调用GainRatio方法计算属性的信息增益率并进行比较，将具有最大值的属性作为测试属性，主程序实现步骤如下：

输入：训练集D；

输出：决策树。

Step 1：节点0作为起点；

Step 2：对未标记属性值的集合调用Mapper任务进行计数，Reducer过程完成对计数值的加噪处理，生成中间文件temporary_0用于存储每个属性的样本数；

Step 3: 利用temporary_0返回值作为参数计算每个属性的信息增益率，指数机制返回信息增益率最大的属性；

Step 4: 信息增益率最大的属性标记节点0；

Step 5: if(剩余样本== 0)

递归停止；

else下一节点作为起点执行Step 2～Step 4。

Mapper类主要负责计数任务。Mapper获取数据集的属性数以及属性的值域，每个属性的取值记为*attr_value*，并将(*attr_value*,1)发送给Reducer类。Mapper的主要步骤如下：

输入：训练集D；

输出：(*attr_value*,1)。

Step 1: 将读入的一行数据依据空格拆分成属性值*attr_value*和类标签*class_label*；

Step 2: 计算属性数之和；

Step 3: 输出(*attr_value*,1)。

Reducer类主要执行计数任务和对计数结果进行加噪。加噪时首先获取当前节点的决策树高度，根据节点高度分配隐私预算，调用GenLaplaceNoise函数产生服从Laplace分布的随机数。Reducer实现步骤如下：

输入：(*attr_value*,1)；

输出：加噪后的结果*n_sum*。

Step 1: 获取(*attr_value*,1)；

Step 2: 按行读取HDFS中的文件，查找到属性值为*attr_value*样本时，*sum*加1；

Step 3: 获取当前节点树高*current_height*；

Step 4: 依据当前高度对应的隐私预算，调用算法GenLaplaceNoise生成随机值*mu*，与*sum*相加得到加噪结果*n_sum*；

Step 5: 返回*n_sum*。

3 实验结果与分析

3.1 安全性分析

本文方案将整体隐私预算ε平均分为两部分$\varepsilon_1=\varepsilon/2$和$\varepsilon_2=\varepsilon/2$。其中，$\varepsilon_2$用于指数机制进行细分方案的选择，$\varepsilon_1$用于在Reducer类中对计数结果添加Laplace噪声。递归过程中每轮消耗的隐私预算为等差分配，整个过程消耗的隐私预算小于等于ε，即$\varepsilon=\varepsilon_1+\varepsilon_2$，本文方案满足$\varepsilon$–差分隐私保护。

由差分隐私理论可知，为构建决策树分配的隐私预算ε可表征方案的隐私保护水平，可将ε作为方案的安全性指标。方案中ε值越小，则隐私保护水平越高，数据的安全性越高。

3.2 性能指标

本文方案满足ε–差分隐私保护可确保数据安全性，并应在一定程度上可确保分类数据的有效性。本文方案采用分类准确率η作为有效性指标。根据训练集产生的决策树规则，沿根节点到叶子节点的路径，对训练集每一个样本进行测试，判断依据决策树规则得到的类标签与样本的真实类标签是否一致，若一致则正确数加一。依次遍历所有测试集样本，总正确数除以样本总数即可得到分类准确率η。

3.3 结果分析

实验环境配置为虚拟机上运行的Linux操作系统的Ubuntu 32位16.04LTS，RAM 4.00 GB，使用Eclispe4.7.3为集成开发环境，以Hadoop2.8.3为大数据平台，采用Java语言实现算法。

采用UCI数据库的car数据集进行测试。car数据集最初为DEX演示开发的简单分层决策模型，用于预测车型的受欢迎程度。该数据集的类标签共4种，包括unacc、acc、good、v-good，分别表示无法接受、可接受、好、非常好。car数据集共有6种属性，如表1所示。

表 1 car数据集属性分布

Tab. 1 Car dataset attribute distribution

属性	属性值
buying	v-high、high、med、low
maint	v-high、high、med、low
doors	2、3、4、5-more
persons	2、4、more
lug boot	small、med、high
safety	low、med、high

将传统隐私预算分配和等差隐私预算分配下的决策树C4.5算法分别在MapReduce框架上进行实现，将car作为测试数据集，隐私预算分别取1.0、0.9、0.8、0.7、0.6、0.5。

将两种差分隐私预算分配方案得到的分类结果进行统计对比。当数据集car未加噪声时，分类准确率η=73.379 6%。对于car数据集给定上述6种不同的隐私预算进行实现，部分分类结果如图4、5所示。

上述实验可得到car数据集在不同隐私预算的分类结果，计算两种隐私预算分配方式的分类准确率，如表2和图6所示。

由实验结果可以得到以下结论：

1）决策树的分类准确率与给定的隐私预算是正相关的。给定的隐私预算越大，生成的Laplace噪声越小分类准确率越高，但隐私保护水平有所下降。

2）在给定相同的隐私预算时，决策树构建过程中等差隐私预算分配比传统隐私预算分配方案的准确率高。这表明本文方案的隐私预算分配方式在一定程度上可弥补传统方案的缺陷，提高分类准确率。

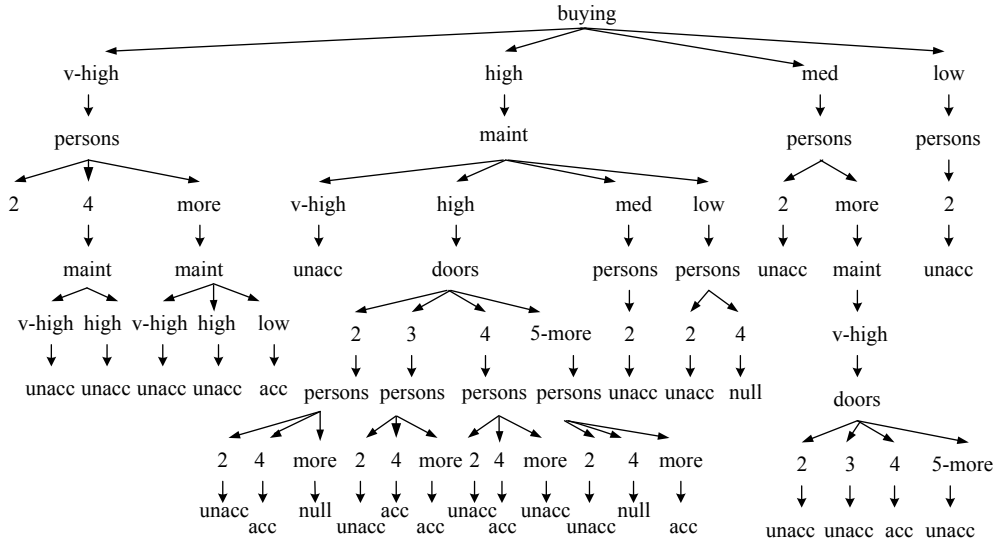

图 4　传统隐私预算分配结果(ε=0.8)

Fig. 4　Results of traditional privacy budget allocation (ε=0.8)

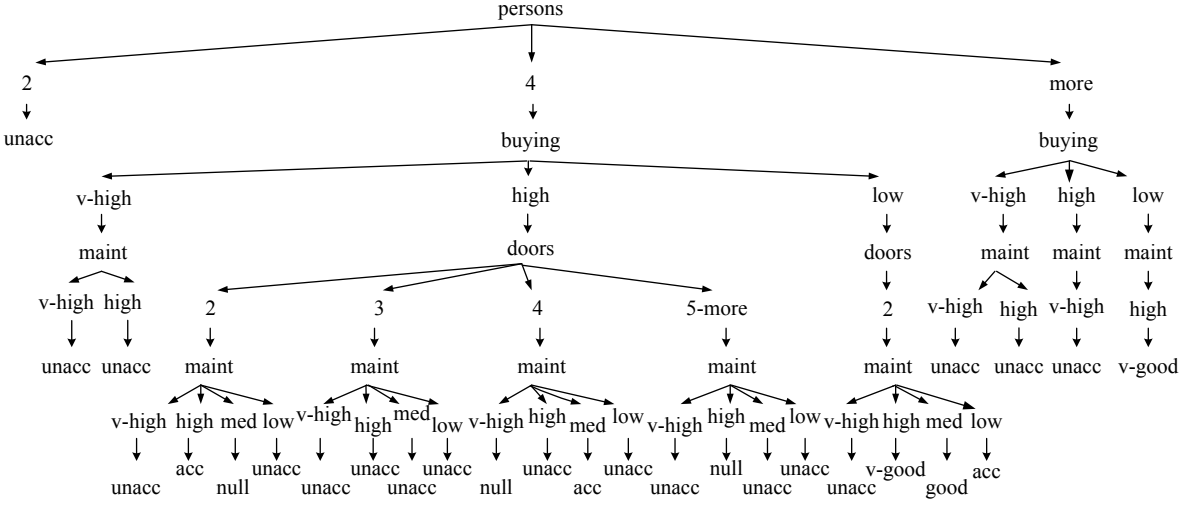

图 5　等差隐私预算分配结果(ε=0.8)

Fig. 5　Results of equal difference privacy budget allocation (ε=0.8)

表 2　两种隐私预算分方案的分类准确率对比

Tab. 2　Comparison of classification accuracy on two privacy budget allocation schemes

隐私预算ε	分类准确率/%	
	等差分配	传统分配
0.5	54.05	42.71
0.6	56.83	50.58
0.7	63.19	52.31
0.8	66.13	54.89
0.9	67.48	57.87
1.0	69.68	62.62

3)对于car测试数据集, 当隐私预算为0.7或0.8时可较好地兼顾数据集的安全性和有效性。

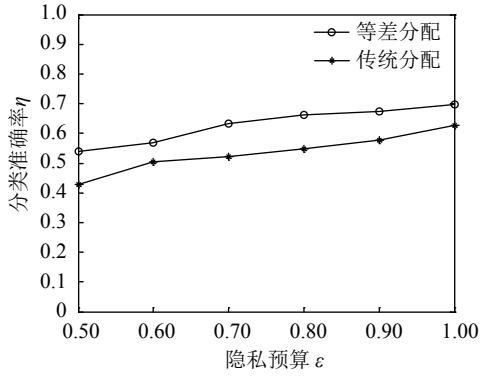

图 6　两种隐私预算分配方案下分类准确率与隐私预算的关系

Fig. 6　Relationship between classification accuracy and privacy budget on two privacy budget allocation schemes

4 结 论

作者将传统隐私预算分配方案与主流决策树构建算法C4.5进行结合，提出基于等差隐私预算分配的大数据决策树C4.5方法。应用差分隐私指数机制对最佳细分方案进行选择，并用Lapalce机制为数据添加随机噪声，可确保决策树每一个分支以及节点数据的安全性。利用大数据平台Hadoop对方案进行实验分析，可以发现本文方案在可接受的分类准确率降低的情况下满足ε–差分隐私保护，同时在相同隐私预算下本文方案与传统隐私预算分配方案相比在一定程度上改善了分类准确率。后续研究中可以寻求新的细分方案、隐私预算分配方案以最大化改善分类准确率。

参考文献:

[1] Cao Hua.Research on decision tree algorithm for privacy preserving[D].Lanzhou:Lanzhou University of Technology,2008.[曹华.保护隐私的决策树算法的研究[D].兰州:兰州理工大学,2008.]

[2] Shao Minhui.A review of typical decision tree algorithm[J].Computer Knowledge and Technology,2018,14(8):175–177.[邵旻晖.决策树典型算法研究综述[J].电脑知识与技术,2018,14(8):175–177.]

[3] Du Wenliang,Zhan Zhijun.Using randomized response techniques for privacy-preserving data mining[C]//Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining.New York:ACM,2003:505–510.

[4] Szűcs G.Random response forest for privacy-preserving classification[J].Journal of Computational Engineering,2013,2013(309):1–6.

[5] Tai R K H,Ma J P K,Zhao Yongjun,et al.Privacy-preserving decision trees evaluation via linear functions[C]//Proceedings of the 22nd European Symposium on Research in Computer Security.Berlin:Springer,2017:494–512.

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

[7] Chen Yu,Li Lingjuan.A decision tree-based privacy preserving classification mining algorithm for data streams[J].Computer Technology and Development,2017,27(7):111–119.[陈煜,李玲娟.一种基于决策树的隐私保护数据流分类算法[J].计算机技术与发展,2017,27(7):111–119.]

[8] Dwork C.The promise of differential privacy:A tutorial on algorithmic techniques[C]//Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science.Washington:IEEE,2011:1–2.

[9] 朱明.数据挖掘导论[M].安徽:中国科学技术大学出版社,2012:44–69.

[10] Gao Shang,Wang Changbao.Personal credit scoring based on decision tree C5.0 algorithm[C]//Proceedings of the 7th International Conference on Education,Management,Computer and Society.Pairs:Atlantis,2017:1729–1734.

[11] Chen Yang,Yu Shoujian.Research on differential privaty for decision tree release technology[J].Computer and Modernization,2017,10(3):59–64.[陈杨,于守健.基于差分隐私的决策树发布技术研究[J].计算机与现代化,2017,10(3):59–64.]

[12] Xiong Ping,Zhu Tianqing,Jin Dawei.Differential private data publishing algorithm for building decision tree[J].Application Research of Computers,2014,31(10):3108–3112.[熊平,朱天清,金大卫.一种面向决策树构建的差分隐私保护算法[J].计算机研究与应用,2014,31(10):3108–3112.]

[13] McSherry F,Talwar K.Mechanism design via differential privacy[C]//Proceedings of the 48th Annual Symposium on Foundations of Computer Science.Washington:IEEE,2007:94–103.

[14] Lam C,韩冀中.Hadoop实战[M].北京:人民邮电出版社,2011:17–134.

[15] Rashid M M,Gondal I,Kamruzzaman J.Dependable large scale behavioral patterns mining from sensor data using Hadoop platform[J].Information Sciences,2017,379:128–145.

[16] Zhu Tianqing,Xiong Ping,Xiang Yang,et al.An effective differentially 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:IEEE,2013:388–395.

（编辑 赵 婧）

引用格式: Shang Tao,Zhao Zheng,Shu Wangwei,et al.Big data decision tree algorithm based on equal-arrival privacy budget allocation[J].Advanced Engineering Sciences,2019,51(2):130–136.[尚涛,赵铮,舒王伟,等.基于等差隐私预算分配的大数据决策树算法[J].工程科学与技术,2019,51(2):130–136.]