不完全数据集的差分隐私保护决策树研究

沈思倩[1] 毛宇光[1,2] 江冠儒[3]
(南京航空航天大学计算机科学与技术学院 南京 211106)[1]
(南京大学计算机软件新技术国家重点实验室 南京 210093)[2]
(卡尔斯鲁厄理工学院计算机系 巴登-符腾堡州 76131)[3]

摘 要 主要研究在对不完全数据集进行决策树分析时,如何加入差分隐私保护技术。首先简单介绍了差分隐私ID3算法和差分隐私随机森林决策树算法;然后针对上述算法存在的缺陷和不足进行了修改,提出指数机制的差分隐私随机森林决策树算法;最后对于不完全数据集提出了一种新的WP(Weight Partition)缺失值处理方法,能够在不需要插值的情况下,使决策树分析算法既能满足差分隐私保护,也能拥有更高的预测准确率和适应性。实验证明,无论是Laplace机制还是指数机制,无论是ID3算法还是随机森林决策树算法,都能适用于所提方法。

关键词 差分隐私保护,不完全数据集,ID3算法,随机森林决策树
中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.06.023

Method of Constructing Differential Privacy Decision Tree Classifier with Incomplete Data Sets

SHEN Si-qian[1] MAO Yu-guang[1,2] JIANG Guan-ru[3]
(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)[1]
(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)[2]
(Informatics, Karlsruhe Institute of Technology, Baden-Württemberg 76131, Germany)[3]

Abstract We mainly studied the problem of constructing differential privacy decision tree classifier with incomplete data sets. We first introduced the differential privacy ID3 decision tree algorithm and differentially private random decision tree algorithm. Then we considered the weakness of the algorithms talked above, and created a new differentially private random decision tree algorithm with exponential mechanism. Finally, an approach for decision tree classifier with incomplete data sets was proposed, which yields better prediction while maintaining good privacy without inserting values, called WP(Weight Partition). And the experimental results show that our approach is suitable for either differential privacy ID3 decision trees or differentially private random decision trees, either laplace or exponential mechanism.

Keywords Differential privacy, Incomplete data sets, ID3 decision tree algorithm, Random decision tree algorithm

1 引言

当今为信息爆炸的时代,在利用各种新技术将生活中丰富的数据搜集存储起来以便于研究的同时,个人隐私数据的泄露成为了当今社会关注的问题。于是,隐私保护问题引起了人们的高度重视,尤其在数据挖掘领域,当数据拥有者将数据提交给数据挖掘者时,必然会给数据集中的隐私信息带来泄露的风险。

微软研究院的德沃克(Dwork)于2006年提出了一种新型的隐私保护方法,即差分隐私保护(Differential-Privacy)[1]。该方法通过添加噪声机制使数据失真来保护数据的隐私性,同时在一定程度上保留数据的可用性。

目前,差分隐私保护在机器学习和数据挖掘领域的应用是研究热点。数据挖掘通过对看似杂乱无章的信息进行有效提取分析来得到价值连城的决策信息。在数据挖掘领域,决策树是一种有效的归纳推理算法,它根据一系列规则对数据进行分类分析。在进行决策树分析的过程中,利用差分隐私保护算法进行隐私保护是十分重要的[2]。

在真实世界的数据集中,数据缺失是经常发生的,而以往的带差分隐私保护的分类决策树算法往往不适用于不完全数据集。所以本文针对真实的不完全数据集,寻找一种合适的带差分隐私保护的分类决策树算法,在保证数据隐私性的同时,尽量提高数据的可用性。

2 相关工作

数据的隐私保护问题最早由统计学家Dalenius于20世

到稿日期:2016-05-22 返修日期:2016-07-18
沈思倩(1991-),女,硕士,主要研究方向为数据库安全、差分隐私保护,E-mail:ssqshelia@outlook;毛宇光(1962-),男,博士,副教授,硕士生导师,主要研究方向为数据库系统及理论、数据挖掘与数据仓库;江冠儒(1991-),男,硕士生,主要研究方向为信息安全、隐私保护。

纪 70 年代末提出。自此之后出现了很多隐私保护模型，如 K-匿名[3]和 l-多样[4]是比较有代表性的模型，但是它们都缺少严格的攻击模型，存在较大的缺陷。直到微软研究院的德沃克(Dwork)于 2006 年提出了差分隐私保护的概念，很好地弥补了之前的隐私保护方法的漏洞。差分隐私保护假设攻击者拥有最大的知识背景，即除目标记录外的所有其他信息。在最大知识背景假设下，差分隐私保护拥有极强的隐私保护能力和严谨的数学定义，有着其他隐私保护模型没有的优势，自诞生起就得到了广泛的应用[5]。

当前学术界的研究重点主要有差分隐私保护下的计数查询、数据合成、机器学习与数据挖掘以及子图计数查询等[6-13]。

在带差分隐私保护的分类决策树算法研究方面，SuLQ 框架于 2005 年被提出，随后实现了差分隐私保护的 SuLQ-based ID3 算法[14]，其基本思想是在每次计算属性的信息增益时，加入噪声的计数值并最终生成决策树。但准确率(隐私预算小于 1)相较于无隐私保护的 ID3 算法，大约降低了 30%。

McSherry 等人于 2009 年开发了一套隐私保护框架 PINQ[15]。Friedman 和 Schuster 于 2010 年基于 PINQ 平台对 ID3 算法进行了改进[16]。但由于计算信息增益的计数查询需要根据各个属性单独查询，导致每个查询的预算很小，无法显著降低噪声，因此 Friedman 和 Schuster 又提出了指数机制下的差分隐私保护算法。在此机制下，只需要一次查询即可实现某一个分裂点的划分，大大减少了隐私预算的消耗，有效地降低了噪声。

Jagannathan 等人于 2009 年提出了差分隐私保护随机森林决策树构建方法[17]。与传统决策树相比，随机森林决策树首先通过随机选择分类属性来构建一个决策树，此过程与数据集中的记录完全无关；然后再把数据集的记录输入这个决策树并分配到相应叶节点中；最后统计各叶节点中的记录数量。一个随机决策树分类器由多个这样的决策树构成，它们共同评估一个记录的分类结果。

本文在研究已有的带差分隐私保护的 ID3 算法和随机森林决策树算法[18]的基础上，提出了指数机制的差分隐私随机森林决策树算法。同时针对缺失数据集，提出了一种全新的 WP(Weight Partition)缺失值处理手段，能够在不需要插值的情况下使决策树算法既能满足差分隐私保护又能进一步提高预测的准确率。无论是 Laplace 机制还是指数机制，无论是 ID3 算法还是随机森林决策树算法，都能够适用于本方法。

3 基础知识

3.1 差分隐私保护

3.1.1 差分隐私保护的定义

差分隐私保护[19-21]是一种基于数据库访问的隐私保护机制，防止在对数据库进行统计查询时可能产生的隐私泄露。它通过添加噪声使得不论从数据库中删除或添加一条记录都几乎不会影响查询结果，从而保护了数据库中每一条记录的隐私信息。定义 1 给出了差分隐私保护的数学表达。

定义 1 假设 D 和 D' 是两个相邻(差别至多为一条记录)的数据集，随机算法为 M，R 是随机算法 M 输出的结果，如果满足 $\frac{Prob(M(D)=R)}{Prob(M(D')=R)} \leqslant A=e^{\varepsilon}$，则认为随机算法 M 提供了 ε-的差分隐私保护。其中，e^{ε} 一般接近于 1，称为隐私预算，其值很小，接近于 0。

ε 越小，查询结果的分布越接近，隐私保护性也就越高。那么随机算法 M 需要在原有查询结果的基础上添加的噪声数量，取决于数据集 D 中某一条记录的改变对查询结果的最大影响，即定义 2 中的敏感度函数。

定义 2 给定 D 和 D' 只相差一条记录的两个数据集。假设 $F(D)=X$ 是一个确定的、没有去除隐私信息的查询函数，作用在数据集 D 上，返回结果 X。敏感度函数 $\Delta F=\max \| F(D)-F(D') \|_{L1}$，指从一个数据集中加入或删除一条记录所能造成的查询结果的最大差别总和。

3.1.2 噪声机制

为了在最坏情况下使两个相邻的数据集能够给出一个分布相似的结果，需要通过加噪声来覆盖此敏感性缺口。常用的噪声机制有 Laplace 机制和指数机制[22]。

定理 1(Laplace 机制) 随机算法 $M(D)=F(D)+Y$，噪声 $Y\sim(Laplace(\Delta F/\varepsilon))$ 服从 Laplace 分布，则该算法满足 ε-的差分隐私保护。

Laplace 机制仅适用于查询函数返回值为实数的情况。但在多数实际应用中，查询函数的返回值为实体对象(如一种方案或一种选择)。针对该情况，一般采用指数机制。

定理 2(指数机制) 随机算法 $M(D,q)=\{$返回值是 r 的概率 $\propto \exp(\frac{\varepsilon q(D,r)}{2\Delta q})\}$，其中，$q(D,r)$ 是可用性度量函数，Δq 是可用性函数的敏感度函数 $q(D,r)$，返回值 r 是一个实体对象。

3.1.3 差分隐私保护的性质

通常，一个复杂的隐私保护问题往往需要多次应用差分隐私保护算法才能得以解决。在此情况下，为了保证整个过程的隐私控制在给定的预算 ε 内，需要合理地将全部预算分配到整个算法的各个步骤中。此时，可以利用差分隐私的两个基本性质。

性质 1 设有算法 $M_1,M_2,\cdots,M_n$，其隐私保护预算分别为 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$，那么对于同一数据集 D，这些算法构成的组合算法 $M(M_1(D),M_2(D),\cdots,M_n(D))$ 提供 $(\sum_i^n \varepsilon_i)$-差分隐私保护。

该性质表明，一个差分隐私保护算法序列构成的组合算法提供的隐私保护水平为全部隐私保护水平的总和。因此该性质也称为序列组合性。

性质 2 设有算法 $M_1,M_2,\cdots,M_n$，其隐私保护预算分别为 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$，那么对于不相交的数据集 $D_1,D_2,\cdots,D_n$，这些算法构成的组合算法 $M(M_1(D),M_2(D),\cdots,M_n(D))$ 提供 $(\max \varepsilon_i)$-差分隐私保护。

该性质表明，如果一个差分隐私保护算法序列中所有算法处理的数据集彼此不相交，那么该算法序列构成的组合算

法提供的隐私保护水平取决于算法序列中的隐私保护水平最差者。该性质也称为并行组合性。

3.2 不完全数据集

如果某一个数据集中某一条或者几条数据的属性值未知，那么即称该数据集为不完全数据集。分类决策树对不完全数据集的做法主要为：如果缺失值较少，可以直接丢弃相应的缺失数据；如果缺失值比例较大，则需要采取一定的插值方法使之尽可能地保存原有信息。

4 指数机制的差分隐私保护随机森林决策树算法

4.1 决策树构建时差分隐私预算的分析

通常 ID3 算法和随机森林决策树算法是以提高决策的准确性为目标的，其本身并不能够提供对数据集的保护。如果将未经过隐私处理的决策树直接发布出去，训练集中的数据隐私就会受到威胁。

带差分隐私的决策树算法是在决策树生成的过程中加入一定量的噪声，以满足事先设定好的隐私预算。如在 ID3 算法中，当划分属性计算每一个属性的信息增益（InfoGain）时，总的隐私预算 B 已知，树的深度为 d，总的属性数为 $a(a<d)$，根据树的深度和属性数分配给每一次差分隐私保护噪声 $B/(d*a)$；但同时又必须保证决策树的可用性，即在测试集检验中准确率不至于降低得太过明显。在随机森林决策树算法中，假设树的个数为 n，决策属性值的个数为 c，每一颗树分配的预算为 $B/(c*n)$，相对于 Laplace 机制，ID3 算法有所减少，与之前的两种方法不同，此隐私预算只与树的个数和决策属性值个数有关。

而在大多数决策树算法（如 ID3）中，几乎都是采用某一种评估算法对整个数据集进行属性的划分并生成一棵决策树。对于比较复杂的数据集，容易造成过拟合和准确率的下降。但是随机森林决策树在属性划分过程中并不会使用数据集中的信息，很好地保留了数据隐私，可以将原本有限的总隐私预算更多地分配给被保护的数据信息。

4.2 决策树构建时隐私保护机制的分析和指数机制的引入

在带差分隐私保护的决策树算法中，其具体的保护机制一般有两种：拉普拉斯（Laplace）机制和指数机制。前人在 ID3 决策树算法中实现了 Laplace 保护机制和指数保护机制，在随机决策树算法中实现了 Laplace 机制，本文则将指数机制引入随机决策树，使之能够在相同隐私预算的情况下，进一步提高预测的准确率。

Laplace 机制的随机森林决策树是当 n 棵随机决策树生成之后，在每一棵树上遍历一遍训练集 S 所有的记录。这样每一个叶节点拥有相同属性的部分记录，统计记录个数并根据个数计算树与树的相对权重。当需要预测的记录到来时，选择相应的叶节点，将预测结果的权重相加并加入相应的 Laplace 噪声，得到最终结果。

本文在此基础上提出了指数机制随机森林决策树算法。在需要预测的记录到来时，首先以每棵树叶节点的权重为可用性函数，根据指数机制选出权重最高的叶节点，同时在最终预测结果时加入指数机制的噪声，得到最终的预测结果。

5 带 WP 缺失值处理的差分隐私分类决策树算法

在现实中，数据集并不都是完整数据集，差分隐私保护在处理缺失数据时往往带有缺陷，即基础的差分隐私保护无法处理记录为空值的数据。简单地说，当对存在数据缺失的数据集进行查分隐私保护时，需要对原数据进行 Marginalisation（删除）。但是一旦缺失的数据过多，或者缺失的数据是有偏的，将对差分隐私保护后的数据可用性产生较大影响。因此，有必要针对缺失的数据集进行特殊的差分隐私处理。

无论是简单的 ID3 算法还是随机决策树，本文采用的 WP 缺失值处理方法均能够对其适用，具体过程主要由以下两部分组成。

5.1 ID3 决策树的构建

首先根据是否缺失将源数据集划分为完整数据集 $S_{complete}$ 和缺失数据集 $S_{missing}$。在 ID3 算法中每次划分属性，并计算每个属性信息增益的过程中，选择某一个属性 A，统计其在 $S_{missing}$ 数据集中的缺失记录的数量 $|S_{missing}|$，由于属性值的缺失并不能够给攻击者提供任何信息，因此 $|S_{missing}|$ 将不再加入噪声，将信息增益公式转化为：

$$infoGain(S,A)=Entropy(S)-\sum_{Axal\in Value(A)}\frac{|S_{Axal}|}{|S|}Entropy(S_{Axal})-\frac{|S_{missing}|}{|S|}Entropy(S_{missing})$$

然后仍按正常的 ID3 算法完成差分隐私保护决策树的构造。

5.2 WP 缺失值处理

本文设置每一条数据记录（包括带有缺失值的记录）的权重 $w=1.0$。根据类似测试集的方法将缺失集中的每一条记录 $d(d\in S_{missing})$ 自顶向下遍历一遍，如果遇到对 d 而言节点的属性是缺失的，则根据当前节点属性值的数量 T_a 计算新的权重 $w_i=\frac{w}{T_a}$，保证 $\sum w_i=1.0$，d 分裂成 T_a 条记录后依次继续对每一个分支进行遍历，一直到叶节点。将该记录和当前的权重 w 添加到各自叶节点的数据集中。

最后每一个叶节点的权重值由两部分组成：完整数据集 D_{oc}（其每一条数据权重都是 1.0）和缺失数据集 D_{om}。每个叶节点所得到的权重为 w_i，由之前差分隐私保护方法对决策树的保护过程得知，在叶节点加入 Laplace 或者指数噪声即可实现对该决策树的差分隐私保护。至此，带有差分隐私保护以及 WP 缺失值处理的决策树构造完毕。

在处理随机森林决策树时，首先构建完随机森林，同样按照上述方法完成 WP 缺失值的处理。

决策树构建的具体算法如算法 1 所示，时间复杂度为 $O(n^2)$。带 WP 缺失值处理的差分隐私分类决策树算法能够有效地处理带有缺失值的数据集，且其时间复杂度也尚可接受。

算法 1 SuLQ-based ID3 with missing data

1. $\forall d\in\mathcal{S}$: $w_d=1.0$
2. procedure BUILDTREE($\mathcal{S}$;$\mathcal{A}$;C; d)
3. if d>depth$_{max}$ or $\mathcal{A}=\emptyset$ then

4. $\mathcal{S}_c$=PARTITION($\mathcal{S}$; ∀c∈C: r_C=c)
5. ∀c∈C: Wc=DPWEIGHT(S_c)
6. return a leaf labeled with W_c
7. end if
8. for every attribute A∈$\mathcal{A}$ do
9. $\mathcal{S}_a$=PARTITION($\mathcal{S}$; ∀a∈A: r_A=a)
10. ∀a∈A: $\mathcal{S}_a$=PARTITION($\mathcal{S}_a$; ∀c∈C: r_C=c)
11. $\mathcal{S}_{missing}$=PARTITION($\mathcal{S}$; r_A=missing)
12. $\mathcal{S}_{missing}$=PARTITION($\mathcal{S}_{missing}$; ∀c∈C: r_C=c)
13. N_c=DPWEIGHT($\mathcal{S}_a$)
14. N_a^c=DPWEIGHT($\mathcal{S}_a$)
15. $N_{missing}$=DPWEIGHT($\mathcal{S}_{missing}$)
16. $N_{missing}^c$=DPWEIGHT($\mathcal{S}_{missing,c}$)
17. $\overline{\text{infoGain}_A}=\sum_{a\in A}\sum_{c\in C}\frac{N_a^c}{N_c}\cdot\log\frac{N_a^c}{N_c}+\sum_{c\in C}\frac{N_{missing}^c}{N_{missing}}\cdot\log\frac{N_{missing}^c}{N_{missing}}$
18. end for
19. $A_{split}=\text{argmax}_A\ \overline{\text{infoGain}_A}$
20. $\mathcal{S}_i$=PARTITION($\mathcal{S}$; ∀i∈A_{split}: $r_{A_{split}}$ = i)
21. ∀i∈A_{split}: Subtree_i=BUILDTREE($\mathcal{S}_i$, $\mathcal{A}\backslash A_{split}$, C, d+1)
22. return a tree with root node A_{split}
23. end procedure

5.3 满足 ε-差分隐私保护的证明

本小节给出了该方法符合 ε-差分隐私保护的证明。

在叶节点中，有：

$$\text{Var}\left(\frac{\sum_{w_i\in D_{ac}\text{ and }c_i=c}w_i}{\sum_{w_i\in D_{ac}}w_i}\right)<\text{Var}\left(\frac{\sum_{w_i\in D_{ac}\text{ and }c_i=c}w_i+\sum_{w_i\in D_{am}\text{ and }c_i=c}w_i}{\sum_{w_i\in D_{ac}}w_i+\sum_{w_i\in D_{am}}w_i}\right)$$

假定有两个数据集 D_1 和 D_2，两者只相差一条记录。V_1，V_2 是根据 D_1，D_2 生成的某一叶节点的所有记录。假定 V_1，V_2 也只相差一条记录，需证明对任意一棵树 R，满足 $\frac{P(A(D_1)=R)}{P(A(D_2)=R)}\leqslant e^{\varepsilon}$。已知针对某一个函数 $f(D)$，加入相应的 Laplace 噪声 $Lap(S(f)/\varepsilon)$能够满足 ε-差分隐私保护。

定义加入隐私后的函数为 $\lambda(A(D_1))$，由于每一个叶节点在计算时都加入了 $Lap(1/\varepsilon)$噪声，那么 $\frac{P(\lambda(A(D_1))=V)}{P(\lambda(A(D_2))=V)}\leqslant e^{\varepsilon}$ 也能够被满足，因此该算法满足 ε-差分隐私保护。

6 实验评估

6.1 实验准备

本文实验采用 UCI Machine Learning Repository 中的 Adult 数据集进行测试，该数据集由美国人口普查数据构成，包含 48840 条记录（训练数据量为 32560，测试数据量为 16280）、15 个属性，本文选取其中 7 个属性作为研究对象，共有 6 个离散属性{workclass, education, relationship, race, sex, native-country}和 1 个决策属性{salary}。决策属性 salary 的值共有“>50K”和“<50K”两种，如果不进行数据挖掘，则直接预测的准确率为 50%。

实验从数据集的大小、数据集的缺失率、隐私保护的机制和预算等几方面进行分析，将本文所提出的 WP 缺失处理算法与对照组进行比较。

实验的硬件环境为 Intel Core™ i3-3220 @ 3.30GHz，1.99GB内存；操作系统为 Microsoft Windows 7 Home；算法在借助 Weka 3.6 平台用 Java 实现。

6.2 实验结果

完整数据集的差分隐私保护 ID3 算法的实验结果如图 1 所示。

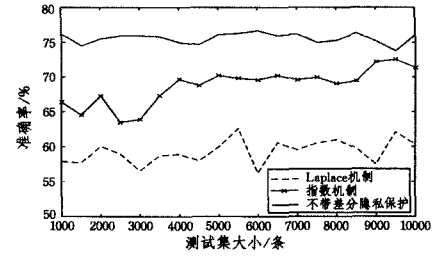

图 1 完整数据集的差分隐私保护 ID3 算法的实验结果

由图 1 可以看出，在 ID3 算法下，Laplace 机制的准确率为 55%～60%，准确率与数据集的大小关系不明显；指数机制的准确率为 65%～73%且逐渐上升；两者的比较指数机制更占有优势，Laplace 机制相比正常数据集有 20%～30%的差距。

完整数据集的差分隐私随机森林决策树算法的实验结果如图 2 所示。

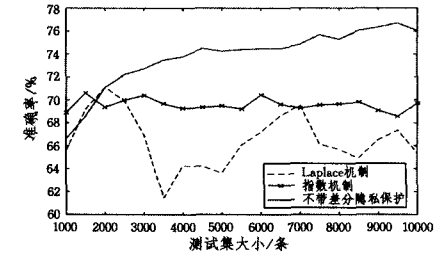

图 2 完整数据集的差分隐私随机森林决策树算法的实验结果

由图 2 可以看出，在随机森林决策树算法下，Laplace 机制的准确率平均值在 66%左右；指数机制的准确率在 71%左右；两者比较，在同等隐私预算的情况下，指数机制的效果更好。

缺失率为 10%的数据集下的差分隐私（Laplace 机制）ID3 和随机森林决策树算法的实验结果如图 3 所示。

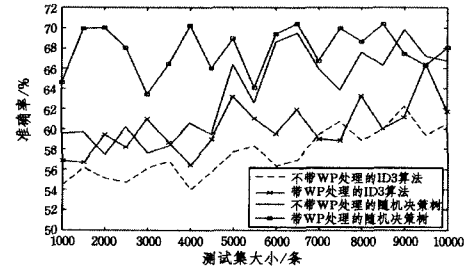

图 3 缺失率为 10%的缺失数据集（Laplace 机制）的实验结果

由图 3 可以看出，在缺失率为 10%且采用 Laplace 机制的情况下，不带缺失值处理的 ID3 算法整体预测准确率较低；带 WP 缺失值处理的 ID3 算法较前者有一定的提升，平均能够提升 4%～5%；不带缺失值处理的随机森林决策树算法相

比 ID3 算法准确率提升的幅度更大；带 WP 缺失值处理的随机森林决策树算法能够一直维持在 66%～70%之间；相互比较可知，带缺失处理后的算法相较于未处理的算法有一定幅度的提升。

缺失率为 10%的数据集下的差分隐私(指数机制)ID3 和随机森林决策树算法的实验结果如图 4 所示。

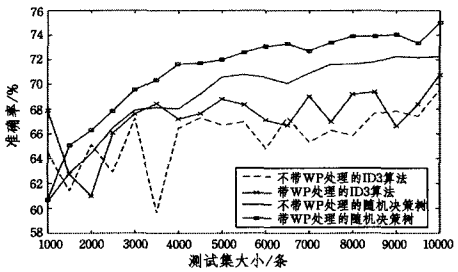

图 4　缺失率为 10%的缺失数据集(指数机制)的实验结果

由图 4 可以看出，在缺失率为 10%且采用指数机制的情况下，不带缺失值处理的 ID3 算法较图 3 的 Laplace 机制有更好的表现，不过波动幅度也较大；在缺失率较低的情况下，带 WP 缺失值处理的随机森林决策树算法与未经过处理的算法相比有了明显的提升。

不同缺失率(5%～70%)数据集的差分隐私(指数机制)ID3 和随机森林决策树算法的实验结果如图 5 所示。

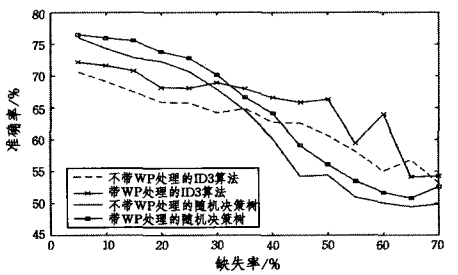

图 5　不同缺失率(5%～70%)的指数机制的实验结果

由图 5 可以看出，在不同缺失率(5%～70%)、数据量为 20000 且采用指数机制的情况下，不带缺失值处理的 ID3 算法的准确率维持在 55%以上并逐渐降低；带 WP 缺失值处理的 ID3 算法相对较好；指数机制下的随机决策树算法具有相对光滑的准确率曲线；不带缺失值处理的随机森林决策树算法相比带 WP 缺失值处理的随机森林决策树算法的预测准确率有相当一部分的降低。

不同隐私预算(0.01～2)的差分隐私(指数机制)ID3 和随机森林决策树算法的实验结果如图 6 所示。

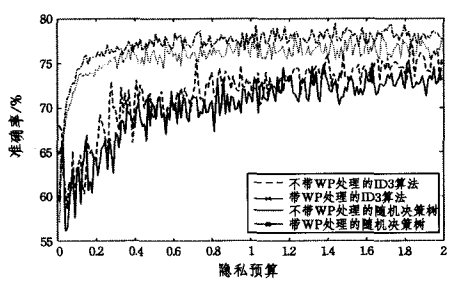

图 6　不同隐私预算(0.01～2)的差分隐私(指数机制)实验结果

由图 6 可以看出，在不同隐私预算(0.01～2)、缺失率为 10%，训练集数据量仍为 20000 且采用指数机制的情况下，不带缺失值处理的 ID3 算法大约从 55%上升到 72%；带 WP 缺失值处理的 ID3 算法为 59%～75%；不带缺失值处理的随机森林决策树算法为 68%～76%；带 WP 缺失值处理的随机森林决策树算法为 68%～79%；比较后可知 WP 缺失值处理后的算法无论是 ID3 算法还是随机决策树算法均有一定的提升。

总体来说，Laplace 机制的波动较大，指数机制的波动较小，更具有可靠性。随着缺失率的增加，经过 WP 缺失处理的算法与未经过缺失处理的算法相比优势越来越大，直到由于缺失率过大导致两者的准确率都无法被接受。随机森林决策树算法相对于 ID3 算法也具有相对较高的准确率。

结束语　本文所述的带 WP 缺失值处理的差分隐私分类决策树方法对于差分隐私保护下的两种决策树算法都具有良好的适应性，能够比较明显地提升决策树算法的准确率，尤其是在随机决策树和指数机制下效果更好。随着缺失率的增加，准确率提升得更加明显。

在未来，希望将本文所采用的方法进行改进，使其能够适用于任意一种决策树算法，同时能够优化一些细节使之在满足同样差分隐私保护的前提下能够拥有更高的预测准确率。

参 考 文 献

[1] DWORK C, DANOS V, KASHEFI K, et al. Automata, Languages and Programming[J]. Verlag Lecture Notes in Computer Science, 1980, 27(3): 282-298.

[2] CLIFTON C, KANTARCIOGLU M, VAIDYA J. Defining Privacy for Data Mining[C]// Proceedings of the National Science Foundation Workshop on Next Generation Data Mining. 2012: 126-133.

[3] SWEENEY L. k-anonymity: A Model for Protecting Privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[4] MACHANAVAJJHALA A, KIFER D, GEHRKE J, et al. l-Diversity: Privacy Beyond k-Anonymity[C]// Proceeding of the 22nd International Conference on Data Engineering. 2006: 24.

[5] 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. (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.

[6] HAY M, RASTOGI V, MIKLAU G, et al. Boosting the Accuracy of Differentially Private Histograms Through Consistency [J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 1021-1032.

[7] XIAO X, WANG G, GEHRKE J. Differential Privacy via Wavelet Transforms[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(8): 1200-1214.

[8] CORMODE G, PROCOPIUC C, SRIVASTAVA D, et al. Differentially Private Spatial Decompositions[C]// 2012 IEEE 28th International Conference on Data Engineering (ICDE). 2012: 20-31.

(下转第 149 页)

[11] HOU S J,ZHANG Y J,LIU G H. Spatial K-Anonymity Reciprocal Algorithm Based on Locality-sensitive Hashing Partition [J]. Computer Science,2013,40(8):115-118. (in Chinese)
侯士江,张玉江,刘国华. 基于位置敏感哈希分割的空间 K-匿名共匿算法[J]. 计算机科学,2013,40(8):115-118.

[12] DATAR M,IMMORLICA N,INDYK P,et al. Locality-sensitive hashing scheme based on p-stable distributions[C]// Twentieth Symposium on Computational Geometry. 2004:253-262.

[13] KULIS B,GRAUMAN K. Kernelized Locality-Sensitive Hashing[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2012,34(6):1092-1104.

[14] FISICHELLA M,DENG F,NEJDL W. Efficient Incremental Near Duplicate Detection Based on Locality Sensitive Hashing [M]// Database and Expert Systems Applications. Springer Berlin Heidelberg,2010:152-166.

[15] LI H M,HAO W N,CHEN G. Collaborative filtering recommendation algorithm based on exact Euclidean locality-sensitive hashing[J]. Journal of Computer Applications,2014,34(12):3481-3486. (in Chinese)
李红梅,郝文宁,陈刚. 基于精确欧氏局部敏感哈希的协同过滤推荐算法[J]. 计算机应用,2014,34(12):3481-3486.

[16] LI H M,HAO W N,CHEN G. Collaborative Filtering Recommendation Algorithm Based on Improved Locality-sensitive Hashing[J]. Computer Science,2015,42(10):256-261. (in Chinese)
李红梅,郝文宁,陈刚. 基于改进 LSH 的协同过滤推荐算法[J]. 计算机科学,2015,42(10):256-261.

[17] LIU Z,LIU T,GIBBON D C,et al. Effective and scalable video copy detection[C]// ACM Sigmm International Conference on Multimedia Information Retrieval. Mir 2010,Philadelphia,Pennsylvania,Vsa,March. DBLP,2010:119-128.

[18] SARAVANAN K,SENTHILKUMAR A. Security Enhancement in Distributed Networks Using Link-Based Mapping Scheme for Network Intrusion Detection with Enhanced Bloom Filter [J]. Wireless Personal Communications,2015,84(2):821-839.

[19] MALHI A,BATRA S. Privacy-preserving authentication framework using bloom filter for secure vehicular communications [J]. International Journal of Information Security,2015,13(1):1-21.

[20] HUO Z,XIAO L,ZHONG Q,et al. MBFS:a parallel metadata search method based on Bloomfilters using MapReduce for large-scale file systems[J]. The Journal of Supercomputing,2016:12(8):3006-3032.

[21] BHUSHAN M,SINGH M,YADAV S K. Big data query optimization by using Locality Sensitive Bloom Filter[C]// International Conference on Computing for Sustainable Global Development. IEEE,2015:70-71.

[22] UCI Machine Learning Repository[OL]. http://archive.ics.uci.edu/ml.

(上接第 143 页)

[9] LI C,HAY M,RASTOGI V,et al. Optimizing Linear Counting Queries under Differential Privacy[C]// Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. 2010:123-134.

[10] XIONG P,ZHU T,NIU W,et al. A Differentially Private Algorithm for Location Data Release[J]. Knowledge and Information Systems,2016,47(3):647-669.

[11] XIAO X,BENDER G,HAY M,et al. iReduct:Differential Privacy with Reduced Relative Errors[C]// Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. 2011:229-240.

[12] CHAUDHURI K,MONTELEONI C,SARWATE A D. Differentially Private Empirical Risk Minimization[J]. The Journal of Machine Learning Research,2011,12:1069-1109.

[13] ZHANG J,XIAO X,YANG Y,et al. PrivGene:Differentially Private Model Fitting using Genetic Algorithms[C]// Proceedings of the 2013 ACM SIGMOD International Conference on Mana-gement of Data. 2013:665-676.

[14] BLUM A,DWORK C,MCSHERRY F,et al. Practical privacy:the SuLQ Framework[C]// Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. 2005:128-138.

[15] MCSHERRY F D. Privacy Integrated Queries:an Extensible Platform for Privacy-preserving Data Analysis[C]// Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. 2009:19-30.

[16] FRIEDMAN A,SCHUSTER A. Data mining with Differential Privacy[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010:493-502.

[17] JAGANNATHAN G,PILLAIPAKKAMNATT K,WRIGHT R N. A Practical Differentially Private Random Decision Tree Classifier[C]// IEEE International Conference on Data Mining Workshops,2009(ICDMW'09). 2009:114-121.

[18] XIONG P,ZHU T Q,JING D W. Different Private Data Publishing Algorithm for Building Decision Tree[J]. Application Research Computers,2014,31(10):3108-3112. (in Chinese)
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]. 计算机应用研究,2014,31(10):3108-3112.

[19] DWORK C. The promise of Differential Privacy:A Tutorial on Algorithmic Techniques[C]// Foundations of Computer Science (FOCS). 2011:1-2.

[20] DWORK C. Differential Privacy in New Settings[C]// SODA. 2010:174-183.

[21] DWORK C. Differential Privacy:A survey of Results[M]// Theory and Applications of Models of Computation. Springer,2008:1-19.

[22] MCSHERRY F,TALWAR K. Mechanism Design via Differential Privacy[C]// 48th Annual IEEE Symposium on Foundations of Computer Science,2007(FOCS'07). 2007:94-103.