面向医疗集值数据的差分隐私保护技术研究

王美珊，姚兰，高福祥，徐军灿

引用本文

王美珊，姚兰，高福祥，徐军灿. 面向医疗集值数据的差分隐私保护技术研究[J]. 计算机科学, 2022, 49(4): 362-368.

WANG Mei-shan, YAO Lan, GAO Fu-xiang, XU Jun-can. Study on Differential Privacy Protection for Medical Set-Valued Data[J]. Computer Science, 2022, 49(4): 362-368.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300032

面向医疗集值数据的差分隐私保护技术研究

王美珊　姚　兰　高福祥　徐军灿
东北大学计算机科学与工程学院　沈阳 110169
(641234923@qq.com)

摘　要　信息技术和医疗健康信息化的不断发展使医疗数据大规模涌现,为数据分析、数据挖掘、智能诊断等更深层次的应用提供了条件。医疗数据集庞大且涉及大量病人隐私,如何在使用医疗数据的同时保护病人隐私极具挑战性。目前应用于医疗领域的隐私保护技术主要以匿名化技术为主,但当攻击者具有强大的背景知识时,此类方法无法兼顾数据集的隐私性和可用性。因此提出了一种优化分类树算法,并改进了 Diffpart 分区算法,以数据间关联性为前提,挑选出医疗集值数据集中的适当数据,利用差分隐私保护技术进行加噪处理,满足差分隐私干扰并支持统计查询。最后在 24 万余条真实医疗数据集上进行测试。实验结果表明,所提算法满足差分隐私分布,并且相比 Diffpart 算法具备更高的隐私性和效用。

关键词:集值数据;医疗大数据;差分隐私;隐私保护;数据可用性
中图法分类号　TP309.2

Study on Differential Privacy Protection for Medical Set-Valued Data

WANG Mei-shan, YAO Lan, GAO Fu-xiang and XU Jun-can
School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China

Abstract　Electronic medical data surges along with the constant development of information technologies and medical care digitalization. It provides foundations for further application on data analysis, data mining and intelligent diagnosis. The fact that medical data are massive and involve a lot of patient privacy. How to protect patient privacy while using medical data is challenging. The predominant principle for the solutions is anonymity. It is not competent in confidentiality or availability when attackers possess strong background knowledge. This paper proposes an optimized classification tree and an improved Diffpart. In our design, association of data is introduced to sift set-valued data for DP based perturbation, which satisfies the utility and supports statistic query. Then test is conducted with 240000 practical medical data and the results show that the proposed algorithm holds DP distribution and outperforms Diffpart in privacy and utility.

Keywords　Set-Valued data, Medical big data, Differential privacy, Privacy protection, Data utility

1　引言

医疗技术的进步和互联网技术的发展使得医疗数据信息化的规模与日俱增,各类庞大的医疗数据集已经成为重要资源。科研人员对医疗大数据的分析结果可用于疾病的预防与控制、新药物的研发以及对疾病的治疗。但是人们获益于医疗大数据的同时,病人个人信息易遭到泄露,如病人的身份证号、手机号以及病症等,这些信息会随着数据集的发布或共享被他人获取甚至恶意使用。为防止病人隐私信息泄露,医疗信息的隐私保护非常必要,且对医疗大数据的获取、分析及发展有着重要作用。

医疗大数据的获取来源有很多,以电子病历为例,包括患者的姓名、住址等基本个人信息,同时还包括一次住院期间的各个时段的体征数据和诊断结果。由于每个病人在各个时段对应多条体征数据且病人基数过大,数据记录量极其庞大。

医疗大数据中主要采用匿名技术进行隐私保护,例如 k-匿名模型[1]。它通过删除或者隐藏数据集的标识属性以达到隐私保护的目的。虽然此方式可以在一定程度上保护病人的隐私,但一些攻击案例[2-3]表明,这种简单的操作远不足以保证隐私信息的安全[4]。

医疗数据的特殊性在于,原始的医疗大数据以标识病人身份属性(如住院号)及临床记录属性值为基础。检测指标的多样性以及检测次数频繁导致医疗数据集庞大、复杂且冗余。此外,人的部分体征具有关联性使体征数据的变化也具有相关性,例如,在一般情况下,当人的血压升高时,其心率也会随之加快。这些特点均符合集值数据的特征。集值数据由一个标识和一个数据组合构成,符合医疗数据中病人唯一的身份标识对应一组临床体征数据模式。通过将多条记录数据转化为一条集值数据的形式存储,可以有效减少医疗大数据集中的数据量,同时有利于对联系较密切的数据进行相同程度的干扰。

本文首先将大量的医疗数据转化成集值数据集,提出了

到稿日期:2021-03-02　返修日期:2021-08-07
通信作者:高福祥(gaofuxiang@mail.neu.edu.cn)

一种分类树算法并改进了 Diffpart 算法进行数据分区与筛选,通过减少需要加噪的数据量来提高数据的可用性;再对挑选出的数据添加 Laplace 噪声并证明其满足 ε-差分隐私。本文第 2 节介绍了与集值数据相关的已有工作;第 3 节给出了问题定义及算法的概要设计;第 4 节讨论了算法设计的详细原理;第 5 节设计实验并分析了算法的可用性及效用;最后对全文进行总结。

2 相关工作

差分隐私是 Dwork[5] 在 2006 年针对统计数据库的隐私泄露问题提出的一种新的隐私定义。在此定义下,对数据集的计算处理结果对具体某个记录的变化不敏感,单条记录在数据集中或者不在数据集中,对计算结果的影响微乎其微。因此,一条记录因其加入到数据集中所产生的隐私泄露风险被控制在极小的、可接受的范围内,攻击者无法通过观察计算结果而获取准确的个体信息。目前,一些企业已经开展了相关的工程实践。Google 利用本地化差分隐私保护技术每天从 Chrome 浏览器采集超过 1 400 万用户的行为统计数据。

Xiao 等[6] 基于小波分析提出一种 ε-差分隐私发布策略,为区间查询提供准确的结果;Hay 等[7] 针对图的度分布估计问题,提出一种有效的差分隐私算法;Mcsherry 等[8] 为用户行为提供推荐的同时满足差分隐私约束;Chen 等[9] 首次提出事务数据的差分隐私发布机制。

在实践中,通常使用拉普拉斯机制[10](Laplace Machanism)和指数机制[11](Exponential Mechanism)来实现差分隐私保护。其中,拉普拉斯机制用于对数值型结果的保护,指数机制用于对离散型结果的保护。

Abadi 等[12] 通过将数据分为不同适当大小的组并加以计算,然后重新计算了应用于算法的隐私预算等参数,提高了算法的效率和数据集的可用性。Cai 等[13] 研究了统计查询结果的准确性和敏感信息保护程度之间的平衡关系,通过改变算法中使用的参数在结果满足差分隐私的条件下提高了统计结果的准确性。

Beaulieus-Jones 等[14] 利用机器学习技术提出了一种基于 DP-SGD 训练 AC-GANs 的模型,称为 DP-GANs,其使用深度神经网络在差分隐私技术下生成合成的数据,为数据的隐私保护提供解决方案并应用于医疗领域。

针对应用于聚类中的差分隐私保护,Blum[15] 在 2005 年提出了应用差分隐私保护技术的 k-means 聚类方法和获取 ε-差分隐私保护的主要思想。2011 年,Dwork 等[16] 对算法进行了更进一步的研究和分析,提出了 k-means 差分隐私保护算法中每个查询函数和整个查询序列敏感度的计算方法。在国内,Li 等[17] 在 2013 年提出了一种新的满足 ε-差分隐私保护的 IDP k-means 聚类方法,解决了差分隐私 k-means 聚类方法的聚类结果可用性差的问题。

Song 等[18] 提出的随机 k 匿名方法考虑了数据规模较大的因素。首先将原始的大数据集划分为几个不相交的子数据集,然后在各个子数据集中形成各自的等价类,通过减少在整个数据集中寻找等价类的计算代价和匿名数据集中的信息损失来提高数据的可用性。

Shi 等[19] 提出一种基于分类树的差分隐私保护下的动态集值型数据发布的算法。该算法首先根据数据集中项的全集构造关系矩阵,挑选关系最紧密的项集构造分类树;然后设定一个边界值来限制数据的增量更新,并将新增的记录添加到分类树的根节点中,按照初始分类树的分配法迭代分配每个记录;最后根据拉普拉斯机制向叶子节点中加入噪声,保证整个算法满足差分隐私的要求。该算法优化了分类树,使建立的分类树模型有少量叶子节点产生,减少了噪声的添加。

Li 等[20] 提出一种基于差分隐私的数据查询分级控制策略。当查询用户提交查询请求时,根据查询者的权限、信誉值和数据隐私属性计算查询安全信任度并量化分级,对不同信任等级的查询返回结果添加服从不同分布特性的 Laplace 噪声以保护数据隐私,并引入可用性评估模块,在保护隐私的同时对数据的可用性进行分析。

Dong 等[21] 基于差分隐私技术不需要考虑攻击方所具备的任何相关背景知识等特性,将差分隐私技术应用于推荐系统中,分析了差分隐私与推荐算法相结合的应用情况,并讨论了差分隐私技术结合不同种推荐算法的使用场景。

在医疗方面,Chen 等[22] 提出了一种建立在 (α,k)-匿名数据基础上的支持数据动态更新的算法——(α,k)-UPDATE。该算法通过对语义贴近度的计算,在 (α,k)-匿名数据集中选择最贴近的等价类,再进行相应的更新操作。更新后的匿名数据集满足 (α,k)-匿名约束,可有效地保护患者的隐私信息。Hmood 等对电子病历按照疾病种类进行聚类操作,再通过 LKC-privacy 技术对聚类内的具体的元组进行加噪处理,使数据的可用性有所提升。

但是,当具有很强背景知识的新型攻击出现时,以匿名为基础的隐私保护模型需要不断调整和完善,才可以避免攻击,而且其隐私保护的有效性等性能也很难证明。相比较而言,差分隐私技术充分考虑了强大的攻击者模型,提供了一种严格的、可证明的隐私保护性能保证。它既不依赖于背景知识假设,也不会对数据集的可用性产生较大的影响,很好地解决了匿名化隐私保护模型存在的问题。

基于前人的工作,本文引入差分隐私技术,并结合医疗数据庞大且病人身份信息冗余的特点,在数据发布之前对医疗数据集进行处理,将其转为集值数据形式,利用本文提出的分类树算法及改进的 Diffpart 分区算法进行数据分区与筛选,通过减少需要加噪的数据量来提高数据的可用性;再对挑选出的数据添加 Laplace 噪声,使数据既满足差分隐私扰动,又具备较高的可用性,对数据集进行统计查询时,达到在个人层面保护病人隐私的目的。

3 问题描述

3.1 集值数据

集值数据由一个代号和一个元素集合组成,可表示为 $\{id,\{item_1,item_2,\cdots,item_n\}\}$。每个记录中元素集合的大小不确定,任意一个或几个元素的组合都有可能是敏感属性,有泄露隐私信息的风险。例如在顾客购买的商品记录中,每个顾客对应一个物品集合,形如{客人 1,{商品 1,商品 2}},{客人 2,{商品 1,商品 3,商品 4}}。类似地,在医疗数据方面,

每个病人的住院号对应一组诊疗数据，病人个人情况不同导致诊疗数据集的大小存在差异。为了消除这种差异，将医疗数据归一化，医疗数据集可以以集值数据的形式进行存储和运算。

3.2 差分隐私及其特性

对于给定的两个至多相差一条记录的数据集 D_1 和 D_2，A 为一随机算法，$range(A)$ 表示算法 A 的所有输出构成的集合，S 为 $range(A)$ 的子集。若算法 A 满足：

$$\Pr[A(D_1)\in S]\leqslant e^{\varepsilon}\times\Pr[A(D_2)\in S] \tag{1}$$

则算法 A 具有 ε-差分隐私性。ε 为隐私保护预算，代表了算法的隐私保护水平，ε 取值越小，则保护水平越高。

差分隐私定义了极其严格的攻击模型，不关心攻击者拥有多少背景知识，通过对统计结果或原数据集添加噪声的方式达到隐私保护效果。添加的噪声不会影响数据集的大小，同时对于海量数据集，添加少量噪声就会达到很好的隐私保护效果。当数据集的操作为统计查询时，拉普拉斯机制是最优的加噪方法。

3.3 噪声机制

3.3.1 拉普拉斯机制

对于任意一个函数 $f:D\rightarrow R_d$，若算法 Y 满足等式(2)，则 Y 满足 ε-差分隐私。

$$Y(D)=f(D)+Lap\left(\frac{\Delta f}{\varepsilon}\right) \tag{2}$$

其中，函数 $Lap\left(\frac{\Delta f}{\varepsilon}\right)$ 表示 Laplace 密度函数。

$$\Delta f=\max_{D_1,D_2}|f(D_1)-f(D_2)| \tag{3}$$

为函数 $f(D)$ 的查询敏感度。

3.3.2 指数机制

给定打分函数 $q:(D\times O)\rightarrow R$，若算法 k 满足等式(4)，则 k 满足 ε-差分隐私。

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

若算法 k 以正比于 $\exp(\cdot)$ 的概率输出 r，那么 $k(D,\mu)$ 满足 ε-差分隐私。

3.4 差分隐私组合特性

差分隐私保护技术有序列组合性和并行组合性两个重要的组合性质。

序列组合性[23]：给定数据库 D 与 n 个随机算法 $A_1,\cdots,A_n$，且 $A_i(1\leqslant i\leqslant n)$ 满足 ε-差分隐私，则 $(A_1,\cdots,A_n)$ 在 D 上的序列组合满足 ε-差分隐私，即 $\varepsilon=\sum\varepsilon_i$。

并行组合性[23]：设 D 为一个数据库，被划分成 n 个不相交的子集，$D=\{D_1,\cdots,D_n\}$，设 A 为任一个随机算法满足 ε-差分隐私，则算法 A 在 $D=\{D_1,\cdots,D_n\}$ 上的系列操作满足 ε-差分隐私。

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 $D=\{D_1,\cdots,D_n\}$，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。

并行组合特性为差分隐私应用于集值数据集的隐私性验证提供了理论支持。

3.5 基本思想

由于实验数据集庞大，为了更有效地选择数据集以及保证数据加噪后具有适当的隐私性与可用性，本文首先对数据进行预处理，即将医疗大数据集转化为集值数据集；再将属性类别向上泛化形成属性集，通过建立二维数组的形式统计属性间的关联性，建立一棵结构紧密的分类树以指导数据分区；通过改进 Diffpart 分区算法减少项目组的交集，从而减少叶子节点数量；最后对分区后的叶子节点中的数据进行挑选和加噪，并验证算法的可行性及效用。

4 算法设计

4.1 隐私性验证

对于差分隐私算法，其隐私性是关键指标，表现为验证扰动后的数据集是否满足 ε-差分隐私，即满足下列不等式：

$$\Pr[f(D_1)\in S_m]\leqslant e^{\varepsilon}\Pr[f(D_2)\in S_m] \tag{5}$$

其中，D_1 为原始数据集，D_2 为扰动后的数据集，f 为查询算法，P_m 为 M 所有可能的输出构成的集合，S_m 为 P_m 的任何子集。

在医疗数据集中，因为数据集 D_1 为原始数据集，查询函数输出的值即为真实的值，所以当查询操作的输出值是对某一个特定值的判断时，有两种情况，即概率为 0 或 1。在式(5)中，$\Pr[f(D_1)]$ 为 0 时，$\Pr[f(D_2)]$ 为大于或等于 0 且小于或等于 1 的数，则不等式右侧恒大于或等于 0，由此可得不等式恒成立。

假设数据集 D_2 中的数据对应一个 $m\times n$ 阶矩阵，即 m 个病案号及与之对应的集值数据；n 为一条集值数据中属性的个数(不包括 ID 和时间属性)。假设加噪函数只作用于一条记录，即改变 n 个数值。当 $\Pr[f(D_1)]$ 为 1，说明 D_2 中加噪的数据没有影响查询结果，这时 $\Pr[f(D_2)\in S_m]$ 的值表示为 $\Pr[f(D_2)\in S_m]=1-\left(\frac{1}{2\lambda}e^{-|x|\lambda}\right)^n$。差分隐私的可行性推导如下。

假设加噪后的数据集满足 ε-差分隐私，则此数据集满足式(5)。若要证明式(5)，即证：

$$\begin{aligned}&\frac{\Pr[f(D_1)\in S_m]}{\Pr[f(D_2)\in S_m]}\leqslant e^{\varepsilon}\\&\frac{1}{\Pr[f(D_2)\in S_m]}\leqslant e^{\varepsilon}\end{aligned} \tag{6}$$

对式(6)不等式两边同时取对数，得：

$$\ln\left(\frac{1}{\Pr[f(D_2)\in S_m]}\right)\leqslant\varepsilon$$

$$\ln 1-\ln(\Pr[f(D_2)\in S_m])\leqslant\varepsilon$$

$$-\ln(\Pr[f(D_2)\in S_m])\leqslant\varepsilon$$

$$-\ln\left(1-\left(\frac{1}{2\lambda}e^{-|x|\lambda}\right)^n\right)\leqslant\varepsilon$$

$$-\ln\left(1-\frac{e^{-|x|\lambda n}}{2^n\lambda^n}\right)\leqslant\varepsilon$$

$$-\ln\left(\frac{2^n\lambda^n-e^{-|x|\lambda n}}{2^n\lambda^n}\right)\leqslant\varepsilon$$

$$\ln^{2^n\lambda^n}-\ln^{2^n\lambda^n-e^{-|x|\lambda n}}\leqslant\varepsilon$$

$$\frac{\ln^{2^n\lambda^n}-\ln^{2^n\lambda^n-e^{-|x|\lambda n}}}{e^{-|x|\lambda n}}\leqslant\frac{\varepsilon}{e^{-|x|\lambda n}}$$

不等式左侧可看作 $\ln(x)$ 的导数形式，则不等式的转化如下：

$$\frac{1}{2^n\lambda^n}\leqslant\frac{\varepsilon}{e^{-|x|\lambda n}}$$

$$\frac{e^{-|x|\lambda n}}{2^n\lambda^n}\leqslant\varepsilon$$

ε-差分隐私中，$\lambda=\frac{\Delta f}{\varepsilon}$。在上述推导中，$\Delta f$ 最大的情况为 n 项均改变，即

$$\max(\Delta f)=n,\lambda=\frac{n}{\varepsilon} \tag{7}$$

将式(7)代入上述不等式中，得：

$$\frac{e^{-|x|\lambda n}}{2^n\lambda^n}\leqslant\frac{n}{\lambda}$$

$$\frac{e^{-|x|\lambda n}}{2^n\lambda^{n-1}}\leqslant n \tag{8}$$

式(8)中，不等式左侧的分子 $0<e^{-|x|\lambda n}\leqslant 1$，分母中的 $\lambda\geqslant 1$，所以分母 $2^n\lambda^{n-1}\geqslant 2$，即不等式左侧恒大于 0 且小于 1。不等式右侧 n 为任一条集值数据的属性个数 n，有 $n\geqslant 1$。综上，不等式(8)恒成立，说明本文提出的方法符合 ε-差分隐私的条件。

当选择一个矩阵中的多个行元素进行加噪时，可视为将数据集分成多个子集。对它们的加噪操作等价于分别对子数据集施加拉普拉斯噪声。根据差分隐私的并行组合性，这种方法均符合 ε-差分隐私，因此加噪后的数据集仍然满足 ε-差分隐私，本文算法隐私性即证。

4.2 结构紧密的分类树算法

本文使用的医疗数据集数据量庞大，若加噪算法产生的噪声量过大，则无法保证数据效用。本文提出的结构紧密的分类树算法较 Diffpart 算法中使用的随机分类树算法有效减少了加噪的数据量。数据利用率反映了在差分隐私保护技术处理后的数据集上进行数据挖掘时的效用情况。特别地，本文使用相对标准偏差来衡量数据的利用率，其定义为：

相对标准偏差(RSD)＝标准偏差(SD)/计算结果的算术平均值(X)＊100%。

为了降低敏感度，首先计算属性间的皮尔森相关系数，将绝对值在 0～0.2 之间的极弱相关或不相关属性删除。将数据集中的属性向上泛化，减少属性种类，形成属性集。例如，将“体温”“脉搏”等属性泛化为“基础特征集”。统计每条集值数据中泛化后各个属性的个数，建立一个二维数组，统计每两个属性同时出现的次数。二维数组如图 1 所示。

$$\begin{bmatrix}
-1 & 53 & 716 & 32 & 2104 & 11 & 258 & 34 & 5 & 16 & 2 & 14 & 4 & 98 & 0\\
53 & -1 & 48 & 2 & 51 & 0 & 8 & 1 & 0 & 0 & 0 & 0 & 1 & 2 & 0\\
716 & 48 & -1 & 26 & 691 & 11 & 61 & 17 & 5 & 1 & 2 & 3 & 2 & 4 & 0\\
32 & 2 & 26 & -1 & 31 & 0 & 7 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0\\
2104 & 51 & 691 & 31 & -1 & 11 & 242 & 30 & 5 & 9 & 2 & 11 & 4 & 91 & 0\\
11 & 0 & 11 & 0 & 11 & -1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
258 & 8 & 61 & 7 & 242 & 2 & -1 & 13 & 4 & 0 & 1 & 5 & 1 & 34 & 0\\
34 & 1 & 17 & 0 & 30 & 0 & 13 & -1 & 0 & 0 & 0 & 0 & 1 & 9 & 0\\
5 & 0 & 5 & 0 & 5 & 0 & 4 & 0 & -1 & 0 & 0 & 1 & 0 & 1 & 0\\
16 & 0 & 1 & 0 & 9 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0\\
2 & 0 & 2 & 2 & 2 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0\\
14 & 0 & 3 & 0 & 11 & 0 & 5 & 0 & 1 & 0 & 0 & -1 & 0 & 9 & 0\\
4 & 1 & 2 & 0 & 4 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 0\\
98 & 2 & 4 & 0 & 91 & 0 & 34 & 9 & 1 & 0 & 0 & 9 & 1 & -1 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1
\end{bmatrix}$$

图 1 属性间关联性二维数组

Fig. 1 Two-dimensional array of correlations between attributes

由图 1 可知，数组中数值越大，两个属性间联系越紧密。利用二维数组建立一棵结构紧密的分类树：数值最大的元素横坐标对应属性为根节点，将横坐标放入 first 列表，以 first 列表为横坐标找到二维数组中的最大值，放入 first 列表，将最小值放 second 列表，third 列表、forth 列表的数据分配方式同 first 列表、second 列表，以 third 列表为横坐标，将最大值放入 third 列表，最小值放入 forth 列表。

建立的分类树如图 2、图 3 所示，first 列表是一组与根节点关系最密切的节点集合，second 列表中各节点和 third 列表的第一个节点与根节点关系最不密切，forth 列表与 third 列表关系最不密切，同时每个列表本身都是一组关系紧密的节点组合，这保证了列表内数据高聚合、列表间数据低耦合。合并 first 列表和 forth 列表作为 D1、合并 second 列表和 third 列表作为 D2，将两组合并的列表作为首次分区，并按照分类树路径指导下述分区。分类树算法的流程如算法 1 所示。

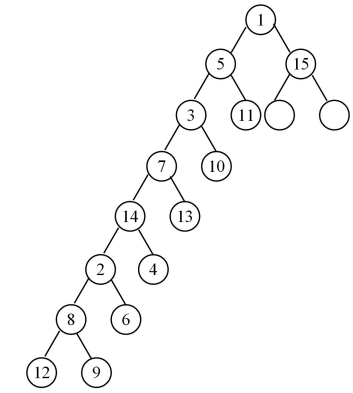

图 2 结构紧密的分类树

Fig. 2 A tightly structured classification tree

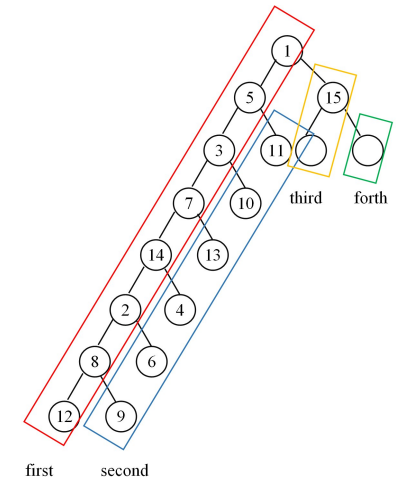

图 3 first，second，third，forth 列表分配

Fig. 3 first，second，third and fourth list allocation

算法 1 分类树算法

输入：二维数组 count，属性列表 l=[T_1，…，T_{15}]，首次分区 D

输出：分类树 T

1. Initialize tree T，flag=0
2. while len(l)>0 do：
3. if flag==0：
4. contrTree(first)
5. contrTree(second)
6. else

```
7.      contrTree(third)
8.      contrTree(forth)
9. return T
```

建立分类树时，总是先建立左子树，再建立右子树。使用 *contrTree*()函数将选出的属性 T 分配到分类树对应的子节点位置，在 count 中找到最大值作为左子树，最小值作为右子树。*contrTree*(*first*)的首次使用用于完成根节点的分配。

4.3 改进 Diffpart 算法

Diffpart 算法采用自顶向下的方法发布集值型数据集，首先利用扇出值来生成分类树，然后根据分类树将数据集中所有记录泛化到层次分割的根节点，再从层次分割的根节点开始迭代，生成不同的子分割和叶子分割，最后对分割后的叶子节点添加拉普拉斯噪声，得到隐私保护后的数据集。但是当数据量较大时，算法添加的噪声量过大，不仅会增加算法的运行时间，还会大大降低数据集的可用性。由于医疗数据集十分庞大，Diffpart 算法无法满足实际应用需求，因此本文对 Diffpart 算法进行改造。

将 4.2 节得到的 D_1，D_2 作为首次分区应用于 Diffpart 算法，并应用分类树路径指导 Diffpart 分区。此改进算法保证了子分区内数据高聚合、子分区间数据低耦合，通过增加伪空节点[22]（子分割中记录数为 1 的不可再分的非叶子分割节点）的数目来减少叶子节点的数目，进一步减少了添加的噪声量，并有效降低了分区的深度。改进的 Diffpart 算法如算法 2 所示。

算法 2 改进的 Diffpart 算法

```
输入：分类树 T
输出：需加噪节点列表 p[]，对应集值数据 ID 列表 S[]
1.  输入首次分区列表
2.  子分区进栈
3.  while not stack.empty():
4. stack.pop()
5.      if '{}' in D:
6.  dealPart()
7.      else
8. p.append(),s.append()
9. return p,s
```

根据算法 2，若数据项中有“{}”，说明此分区可继续划分。在 *dealPart*()函数中，若数据量大于 1，则将分区中项集二分生成新的子分区，并将子分区入栈，否则说明此分区中只有一条记录，即使继续分区也不会成为加噪的对象节点，所以将分区中对应的节点视为伪空节点舍弃。算法 2 得到的集值数据 ID 列表即为需要加噪的数据列表。

4.4 基于差分隐私的医疗集值数据集隐私保护算法

为进一步减少添加的噪声量，同时降低数据集的随机性，对选出的集值数据中的奇数条属性对应的值添加拉普拉斯噪声，若选取的属性值为非数值类型，则执行跳过操作。将加噪后的数据保存到数据库中，即得到了加噪后的数据集。此时数据集上的统计查询操作在保护病人隐私的前提下在统计意义上可用。

5 实验与分析

5.1 实验测试

为测试算法的隐私性及效用，本文设计了如下实验，并将所提算法与 Diffpart 算法进行比较，验证其在隐私性能、运行效率等方面的性能。实验使用 Python 语言实现，编程环境为 Pycharm，实验环境为 Microsoft Windows 7(64 位)操作系统，处理器为 Intel Core(TM) i5-4210U CPU @2.40 GHz，数据库使用 MYSQL 语言。

为验证算法的实际效果，本文在真实的医疗数据集上进行实验，数据集包含 24 万余条医疗数据记录，均为病人住院期间的各项指标数据。对于查询操作，首先设计一个查询函数，即“查询数据集中发烧的人数”。对加噪后的数据集进行大量重复性查询实验，分析实验结果是否呈 Laplace 分布。图 4 为不同实验次数下查询结果的统计柱状图。

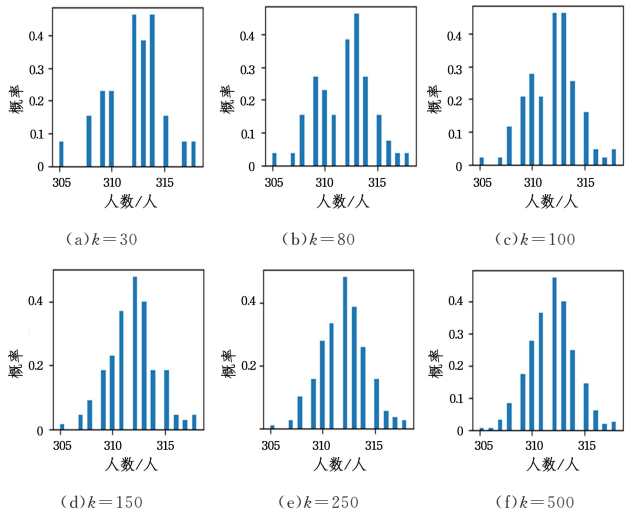

图 4 查询结果统计图

Fig. 4 Statistics of query results

根据图 4 可知，查询结果均呈 Laplace 分布，随着实验次数的增加，数据分布更加拟合拉普拉斯分布模型，为前文的理论分析提供了实验支持。

其次，在不同数据量下对比本文算法和 Diffpart 算法的分区时间，如图 5 所示。

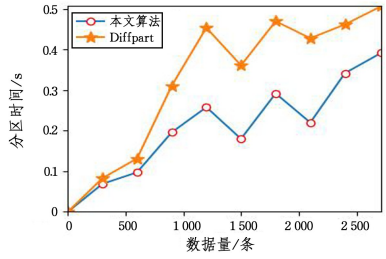

图 5 分区时间对比图

Fig. 5 Comparison of partition time

在不同数据量下本算法的分区时间均少于 Diffpart 算法，这是由于数据分区时本算法保证了各子分区中数据高聚合、分区间数据低耦合，使得分区深度降低，从而有效减少了算法的分区时间，同时因为通过增加伪空节点的数目来减少

分区过程中的叶子节点，因此需要加噪的集值数据量减少，从而减少了加噪时间，如图 6 所示。

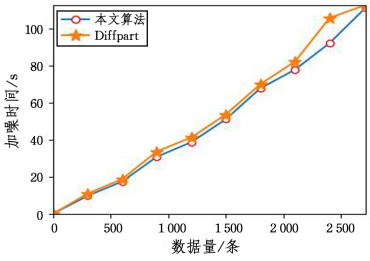

图 6　加噪时间对比图

Fig. 6　Comparison of noise adding time

最后，评估算法在不同隐私预算下的性能。图 7 给出了本文算法和 Diffpart 算法在不同隐私预算下，查询函数在不同数据量的数据集上对应的平均相对误差。从图 7 可以看出，隐私预算越小，查询结果的平均相对误差越大，数据集的隐私性就越强。随着数据集中数据量的增加，平均相对误差增大，这是因为数据集包含了更多噪声扰动后的数据。本文算法的平均相对误差均小于 Diffpart 算法，由此可见分类树的不同导致了添加的噪声量不同，进而影响平均相对误差。本文算法通过减少需要加噪的节点来减少噪声量，减小了查询结果的平均相对误差，从而增加数据集的可用性。

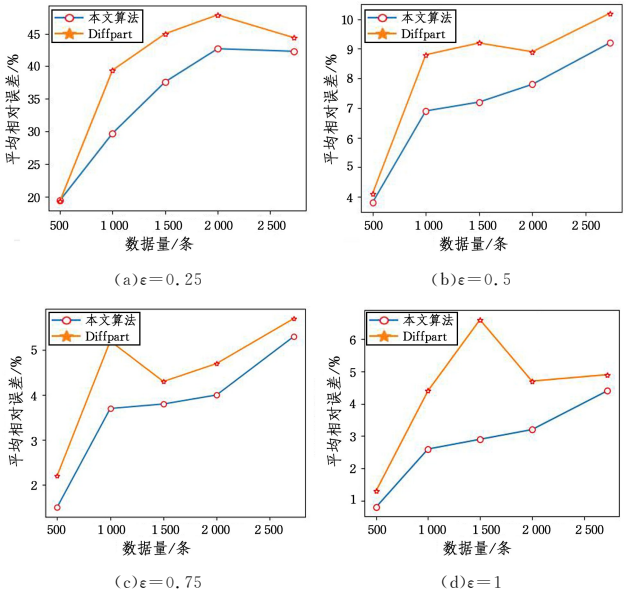

图 7　不同 ε 下平均相对误差

Fig. 7　Mean relative error under different ε

5.2　算法效用比较

本文利用相对标准偏差作为衡量算法效用的标准。将本文算法和 Diffpart 算法在不同数据量下分别进行 100 次实验，计算查询结果的相对标准偏差，结果如图 8 所示。在不同数据量下，本文算法的相对标准误差均小于 Diffpart 算法，即本文算法的隐私性优于 Diffpart 算法，这是因为本算法减少了添加的噪声量，使数据集的扰动程度降低，减少了数据集的标准误差，数据集的可用性得到提高。然后在全数据集上进行多次重复性实验，对比本文算法在不同实验次数下的相对标准偏差。

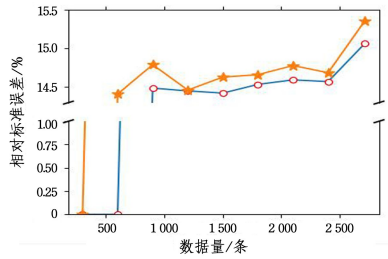

图 8　两种算法的 RSD 对比图

Fig. 8　RSD comparison of two algorithms

图 9 中，随着实验次数的增加，查询结果相对标准误差降低，这是由于查询结果随实验次数的增加越来越聚集在真实值附近，使得相对标准偏差降低。

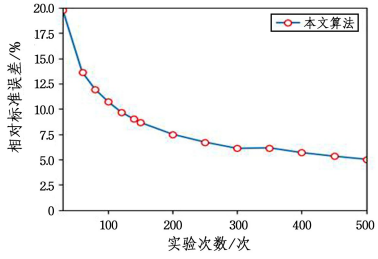

图 9　本文算法的 RSD 图

Fig. 9　RSD diagram of the proposed algorithm

结束语　与传统的在查询函数上加噪声的方法不同，本文通过将医疗大数据转化为集值数据的形式，提出了一个结构紧密的分类树算法，并改进了 Diffpart 算法及加噪算法，对海量医疗数据中多条数据加噪声后这些医疗数据仍然满足 ε-差分隐私，得到一个满足差分隐私扰动的数据集。该算法在支持统计查询的同时，不但能够保护患者的隐私，还最大限度地在统计层面上满足了数据的可用性。对比实验显示，本文提出的方法在数据加噪时间及数据可用性方面均具有优越性。今后将研究算法在多重查询条件下的性能及效用。

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] SAMARATI P. Protecting respondents' identities in microdata release[J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(6): 1010-1027.

[3] NARAYANAN A, SHMATIKOV V. Robust de-anonymization of large sparse datasrts[C]// Proceedings of the 2008 IEEE Symposium on Security and Privacy. Oakland, USA, 2008: 111-125.

[4] 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.

[5] DWORK C. Differential privacy: A survey of results[C]// Proceedings of the 5th International Conference on Theory and Applications of Models of Computation. Xi'an, China, 2008: 1-19.

[6] XIAO X, WANG G, GEHREKE J. Differential privacy via wavelet transforms [C]// Proceedings of the IEEE 26th Interna-

tional Conference on Data Engineering. Piscataway, NJ: IEEE, 2010: 225-236.

[7] HAY M, LI C, MIKLAU G, et al. Accurate estimation of the degree distribution of private networks[C]// Proceedings of the 9th IEEE International Conference on Data Mining. Piscataway, NJ: IEEE, 2009: 169-178.

[8] MCSHERRY F, MIRONOV I. Differentially private recommender systems: building privacy into the net [C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009: 627-636.

[9] CHEN R, MOHAMMED N, FUNG B C M, et al. Publishing set-valued data via differential privacy [J]. Proceedings of the VLDB Endowment, 2011, 4(11): 1087-1098.

[10] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis [C]// Proceedings of the 3rd Conference on Theory of Cryptography. New York, USA, 2006: 265-284.

[11] MCSHERRY F, TALWAR K. Mechanism design via differential privacy[C]// Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science. Providence, Rhode Island, USA, 2007: 94-103.

[12] ABADI M, GOODFELLOW I. Deep learning with differential privacy[C]// ACM Sigsac Conference on Computer and Communications Security. ACM, 2016: 308-318.

[13] CAI T T, WANG Y, ZHANG L. The cost of privacy: optimal rates of convergence for parameter estimation with differential privacy[J]. arXiv: 1902. 04495, 2019.

[14] BEAULIEU-JONES B K, WU Z S, WILLIAMS C, et al. Privacy-preserving generative deep neural networks support clinical data sharing[J]. BioRxiv, 2017, 159756.

[15] 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. 2005: 128-138.

[16] DWORK C, NAOR M, PITASSI T, et al. Pan-private streaming algorithms[C]// Proceedings of the 1st Symposium on Innovations in Computer Science. 2010.

[17] LI Y, HAO Z F, WEN W, et al. Research on differential privacy preserving K-means clustering [J]. Computer Science, 2013, 40(3): 287-290.

[18] SONG F G, MA T H, TIAN Y, et al. A new method of privacy protection: random k-anonymous [J]. IEEE Access, 2019, 7: 75434-75445.

[19] SHI X J, HU Y L. Proprietary protection of dynamic set-valued data release based on classification tree[J]. Computer Science, 2017, 44(5): 120-124, 165.

[20] LI S Y, JI X S, YOU W, et al. A data query hierarchical control strategy based on differential privacy [J]. Computer Science, 2019, 46(11): 130-136.

[21] DONG X M, WANG R, ZOU X K. Survey on Privacy Protection Solutions for Recommended Applications[J]. Computer Science, 2021, 48(9): 21-35.

[22] CHEN H Y, WANG J H, HU Z P, et al. Dynamic update privacy protection algorithm for medical data publishing[J]. Computer Science, 2019, 46(1): 206-211.

[23] MCSHERRY F. Privacy integrated queries: An ex- tensible platform for privacy-preserving data analysis[J]. Communications of the ACM, 2010, 53(9): 89-97.

WANG Mei-shan, born in 1996, postgraduate. Her main research interests include privacy protection and so on.

GAO Fu-xiang, born in 1961, Ph.D, professor. His main research interests include computer network security, embedded computer networks.

(责任编辑:李亚辉)