差分隐私保护及其应用

熊 平[1),2)] 朱天清[2),4)] 王晓峰[3)]

[1)](中南财经政法大学信息与安全工程学院 武汉 430073)

[2)](澳大利亚迪肯大学信息技术学院 墨尔本 澳大利亚 3125)

[3)](中国科学院计算技术研究所无线传感网络实验室 北京 100190)

[4)](武汉轻工大学数学与计算机学院 武汉 430023)

摘 要 数据发布与数据挖掘中的隐私保护问题是目前信息安全领域的一个研究热点.作为一种严格的和可证明的隐私定义,差分隐私近年来受到了极大关注并被广泛研究.文中分析了差分隐私保护模型相对于传统安全模型的优势,对差分隐私基础理论及其在数据发布与数据挖掘中的应用研究进行综述.在数据发布方面,介绍了各种交互式和非交互式的差分隐私保护发布方法,并着重从精确度和样本复杂度的角度对这些方法进行了比较.在数据挖掘方面,阐述了差分隐私保护数据挖掘算法在接口模式和完全访问模式下的实现方式,并对这些算法的执行性能进行了分析.最后,介绍了差分隐私保护在其它领域的应用,并展望未来的研究方向.

关键词 差分隐私;数据发布;数据挖掘;机器学习;统计查询;隐私保护

中图法分类号 TP391 **DOI 号** 10.3724/SP.J.1016.2014.00101

A Survey on Differential Privacy and Applications

XIONG Ping[1),2)] ZHU Tian-Qing[2),4)] WANG Xiao-Feng[3)]

[1)](*School of Information and Security Engineering, Zhongnan University of Economics and Law, Wuhan* 430073)

[2)](*School of Information Technology, Deakin University, Melbourne* 3125, *Australia*)

[3)](*Wireless Sensor Network Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[4)](*School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan* 430023)

Abstract Privacy preserving in data release and mining is a hot topic in the information security field currently. As a new privacy notion, differential privacy (DP) has grown in popularity recently due to its rigid and provable privacy guarantee. After analyzing the advantage of differential privacy model relative to the traditional ones, this paper surveys the theory of differential privacy and its application on two aspects, privacy preserving data release (PPDR) and privacy preserving data mining (PPDM). In PPDR, we introduce the DP-based data release methodologies in interactive/non-interactive settings and compare them in terms of accuracy and sample complexity. In PPDM, we mainly summarize the implementation of DP in various data mining algorithms with interface-based/fully access-based modes as well as evaluating the performance of the algorithms. We finally review other applications of DP in various fields and discuss the future research directions.

Keywords differential privacy; data release; data mining; machine learning; statistical query; privacy preserving

---

收稿日期:2013-04-16;最终修改稿收到日期:2013-11-20.本课题得到国家自然科学基金(61202211,61304067)、教育部人文社科研究青年基金(12YJC630078)、中央高校基本科研业务费专项资金(31541311302,31541111305)资助.**熊 平**,男,1974 年生,博士,副教授,主要研究方向为信息安全、机器学习、数据挖掘.E-mail:pingxiong@znufe.edu.cn.**朱天清**,女,1979 年生,博士研究生,讲师,主要研究方向为隐私保护与网络安全.**王晓峰**,男,1978 年生,博士,助理研究员,主要研究方向为人工智能、数据挖掘、无线传感网络.

1 引 言

随着信息技术应用的不断普及和深入,各种信息系统存储并积累了丰富的数据,例如医疗机构建立的患者诊断数据集,电子商务企业收集的客户在线交易数据集等.对这些数据集进行分析可以使人们获得更多关于真实世界的知识.因此,对于研究机构、信息咨询组织以及政府决策部门来说,数据是非常重要的基础资源.这种需求极大地促进了数据的发布、共享与分析.

然而,数据集里通常包含着许多个人的隐私信息,如医疗诊断结果、个人消费习惯以及其它能够体现个人特征的数据,这些信息会随着数据集的发布和共享而被泄露.虽然删除数据集的标识符属性(如姓名、ID 号等)能够在一定程度上保护个人隐私,但一些攻击案例[1-4]表明,这种简单的操作远不足以保证隐私信息的安全.

数据的隐私保护问题最早由统计学家 Dalenius 在 20 世纪 70 年代末提出.他认为,保护数据库中的隐私信息,就是要使任何用户(包括合法用户和潜在的攻击者)在访问数据库的过程中无法获取关于任意个体的确切信息[5].虽然这一定义具有理论上的指导意义,但显然它是主观的和模糊的.以这一定义为目标,学者们在后续的研究中提出了许多量化指标更明确、可操作性强的隐私保护模型和方法.

从已有的研究来看,k-anonymity[6]及其扩展模型在隐私保护领域影响深远且被广泛研究.这些模型的基本思想是将数据集里与攻击者背景知识相关的属性定义为准标识符,通过对记录的准标识符值进行泛化、压缩处理,使得所有记录被划分到若干个等价类(Equivalence Group),每个等价类中的记录具有相同的准标识符值,从而实现将一个记录隐藏在一组记录中.因此,这类模型也被称为基于分组的隐私保护模型.

然而后续研究表明,这些模型存在两个主要缺陷.其一,这些模型并不能提供足够的安全保障,它们总是因新型攻击的出现而需要不断完善.例如为了抵制"一致性"攻击,l-diversity[7]、t-closeness[8]、(α,k)-anonymity[9]、M-invariance[10]等模型相继被提出;文献[11]提出了 m-confidentiality 模型以抵制"最小性"攻击.许多新型的攻击方式都对基于分组的隐私保护模型形成了挑战,例如"合成式"攻击[12]、"前景知识"攻击[13]、"deFinetti"攻击[14]等.出现这一局面的根本原因在于,基于分组的隐私保护模型的安全性与攻击者所掌握的背景知识相关,而所有可能的背景知识很难被充分定义.所以,一个与背景知识无关的隐私保护模型才可能抵抗任何新型的攻击.第二个缺陷是这些早期的隐私保护模型无法提供一种有效且严格的方法来证明其隐私保护水平,因此当模型参数改变时,无法对隐私保护水平进行定量分析.这个缺点削弱了隐私保护处理结果的可靠性.因此,研究人员试图寻求一种新的、鲁棒性足够好的隐私保护模型,能够在攻击者拥有最大背景知识的条件下抵抗各种形式的攻击.差分隐私(Differential Privacy,DP)[15]的提出使得实现这种设想成为可能.

差分隐私是 Dwork 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[16].在此定义下,对数据集的计算处理结果对于具体某个记录的变化是不敏感的,单个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.所以,一个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.

差分隐私能够解决传统隐私保护模型的两个缺陷.首先,差分隐私保护模型假设攻击者能够获得除目标记录外所有其它记录的信息,这些信息的总和可以理解为攻击者所能掌握的最大背景知识.在这一最大背景知识假设下,差分隐私保护无需考虑攻击者所拥有的任何可能的背景知识,因为这些背景知识不可能提供比最大背景知识更丰富的信息.其次,它建立在坚实的数学基础之上,对隐私保护进行了严格的定义并提供了量化评估方法,使得不同参数处理下的数据集所提供的隐私保护水平具有可比较性.因此,差分隐私理论迅速被业界认可,并逐渐成为隐私保护领域的一个研究热点.近几年来,差分隐私和其它领域研究的结合使得大量新的成果不断涌现.本文在总结已有研究成果的基础上,对差分隐私的理论发展及其在数据发布与数据挖掘领域的应用进行综述.

本文第 2 节介绍差分隐私保护模型的相关定义与基础理论;第 3 节介绍差分隐私保护数据发布在交互式和非交互式环境下的实现方法,并对这些方法进行分析和比较;第 4 节阐述差分隐私保护数据挖掘在接口模式和完全访问模式上的差异,并着重介绍差分隐私保护在各种挖掘算法中的实现;第 5 节简介差分隐私保护在其它领域的应用;最后在第

6 节讨论差分隐私保护研究所面临的挑战和未来的发展方向.

2　差分隐私保护模型

差分隐私保护模型的思想源自于一个很朴素的观察：当数据集 D 中包含个体 Alice 时，设对 D 进行任意查询操作 f（例如计数、求和、平均值、中位数或其它范围查询等）所得到的结果为 $f(D)$，如果将 Alice 的信息从 D 中删除后进行查询得到的结果仍然为 $f(D)$，则可以认为，Alice 的信息并没有因为被包含在数据集 D 中而产生额外的风险. 差分隐私保护就是要保证任一个体在数据集中或者不在数据集中时，对最终发布的查询结果几乎没有影响. 具体地说，设有两个几乎完全相同的数据集（两者的区别仅在于一个记录不同），分别对这两个数据集进行查询访问，同一查询在两个数据集上产生同一结果的概率的比值接近于 1.

差分隐私保护模型最初被应用在统计数据库安全领域，旨在发布统计信息时保护数据库中个体的隐私信息，之后被广泛应用于隐私保护数据发布（Privacy Preserving Data Release，PPDR）与隐私保护数据挖掘（Privacy Preserving Data Mining，PPDM）等领域. 已有的研究表明，差分隐私保护方法既可以应用于交互式的统计查询，也可以应用在各种非交互式的信息发布场合.

2.1　差分隐私的定义与相关概念

2.1.1　基本定义

对于一个有限域 Z，$z\in Z$ 为 Z 中的元素，从 Z 中抽样所得 z 的集合组成数据集 D，其样本量为 n，属性的个数为维度 d.

对数据集 D 的各种映射函数被定义为查询（Query），用 $F=\{f_1,f_2,\cdots\}$ 来表示一组查询，算法 M 对查询 F 的结果进行处理，使之满足隐私保护的条件，此过程称为隐私保护机制.

设数据集 D 和 D'，具有相同的属性结构，两者的对称差记作 $D\Delta D'$，$|D\Delta D'|$ 表示 $D\Delta D'$ 中记录的数量. 若 $|D\Delta D'|=1$，则称 D 和 D' 为邻近数据集（Adjacent Dataset）.

定义 1[17].　差分隐私. 设有随机算法 M，P_M 为 M 所有可能的输出构成的集合. 对于任意两个邻近数据集 D 和 D' 以及 P_M 的任何子集 S_M，若算法 M 满足

$$Pr[M(D)\in S_M]\leq \exp(\varepsilon)\times Pr[M(D')\in S_M] \quad (1)$$

则称算法 M 提供 ε-差分隐私保护，其中参数 ε 称为隐私保护预算[18].

如图 1 所示，算法 M 通过对输出结果的随机化来提供隐私保护，同时通过参数 ε 来保证在数据集中删除任一记录时，算法输出同一结果的概率不发生显著变化.

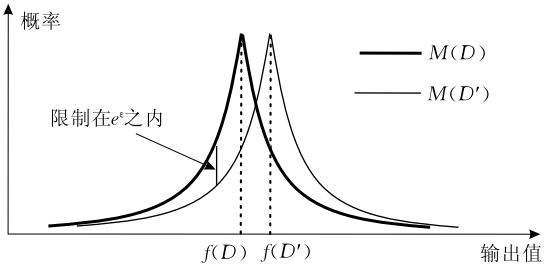


图 1　随机算法在邻近数据集上的输出概率

例如，表 1 显示了一个医疗数据集 D，其中的每个记录表示某个人是否患有癌症（1 表示是，0 表示否）. 数据集为用户提供统计查询服务（例如计数查询），但不能泄露具体记录的值. 设用户输入参数 i，调用查询函数 $f(i)=count(i)$ 来得到数据集前 i 行中满足“诊断结果”=1 的记录数量，并将函数值反馈给用户. 假设攻击者欲推测 Alice 是否患有癌症，并且知道 Alice 在数据集的第 5 行，那么可以用 $count(5)-count(4)$ 来推出正确的结果.

表 1　医疗数据集示例

姓名	诊断结果
Tom	0
Jack	1
Henry	1
Diego	0
Alice	1

但是，如果 f 是一个提供 ε-差分隐私保护的查询函数，例如 $f(i)=count(i)+noise$，其中 $noise$ 是服从某种随机分布的噪声. 假设 $f(5)$ 可能的输出来自集合 $\{2,2.5,3\}$，那么，$f(4)$ 也将以几乎完全相同的概率输出 $\{2,2.5,3\}$ 中的任一可能的值，因此攻击者无法通过 $f(5)-f(4)$ 来得到想要的结果. 这种针对统计输出的随机化方式使得攻击者无法得到查询结果间的差异，从而能保证数据集中每个个体的安全.

2.1.2　相关概念

（1）隐私保护预算

从定义 1 可以看出，隐私保护预算 ε 用来控制算法 M 在两个邻近数据集上获得相同输出的概率

比值,它事实上体现了 M 所能够提供的隐私保护水平.在实际应用中,ε 通常取很小的值,例如 0.01,0.1,或者 ln2,ln3 等.ε 越小,表示隐私保护水平越高.当 ε 等于 0 时,保护水平达到最高,此时对于任意邻近数据集,算法都将输出两个概率分布完全相同的结果,这些结果也不能反映任何关于数据集的有用的信息.因此,ε 的取值要结合具体需求来达到输出结果的安全性与可用性的平衡.

(2)敏感度

差分隐私保护可以通过在查询函数的返回值中加入适量的干扰噪声来实现.加入噪声过多会影响结果的可用性,过少则无法提供足够的安全保障.敏感度是决定加入噪声量大小的关键参数,它指删除数据集中任一记录对查询结果造成的最大改变.在差分隐私保护方法中定义了两种敏感度,即全局敏感度(Global Sensitivity)和局部敏感度(Local Sensitivity).

定义 2[15]. 全局敏感度.设有函数 $f:D\to R^d$,输入为一数据集,输出为一 d 维实数向量.对于任意的邻近数据集 D 和 D',

$$GS_f=\max_{D,D'}\|f(D)-f(D')\|_1 \tag{2}$$

称为函数 f 的全局敏感度.

其中,$\|f(D)-f(D')\|_1$ 是 $f(D)$ 和 $f(D')$ 之间的 1-阶范数距离.

函数的全局敏感度由函数本身决定,不同的函数会有不同的全局敏感度.一些函数具有较小的全局敏感度(例如计数函数,其全局敏感度为 1),因此只需加入少量噪声即可掩盖因一个记录被删除对查询结果所产生的影响,实现差分隐私保护.但对于某些函数而言,例如求平均值、求中位数等函数,则往往具有较大的全局敏感度.以求中位数函数为例,设函数为 $f(D)=\text{median}(x_1,x_2,\cdots,x_n)$,其中 $x_i(i=1,\cdots,n)$ 是区间 $[a,b]$ 中的一个实数.不妨设 n 为奇数,且数据已被排序,那么函数的返回值即为第 $m=(n+1)/2$ 个数.在某种极端的情况下,设 $x_1=x_2=\cdots=x_m=a$ 且 $x_{m+1}=x_{m+2}=\cdots=x_n=b$,那么从中删除一个数就可能使函数的返回值由 a 变为 b,因此函数的全局敏感度为 $b-a$,这可能是一个很大的值.

当全局敏感度较大时,必须在函数输出中添加足够大的噪声才能保证隐私安全,导致数据可用性较差.针对这个问题,Nissim 等人定义了局部敏感度以及与其计算相关的其它概念.

定义 3[19]. 局部敏感度.设有函数 $f:D\to R^d$,输入为数据集 D,输出为一 d 维实数向量.对于给定数据集 D 和它的任意邻近数据集 D',则

$$LS_f(D)=\max_{D'}\|f(D)-f(D')\|_1 \tag{3}$$

称为函数 f 在 D 上的局部敏感度.

局部敏感度由函数 f 及给定数据集 D 中的具体数据共同决定.由于利用了数据集的数据分布特征,局部敏感度通常要比全局敏感度小得多.以前文的求中位数函数为例,其局部敏感度为 $\max(x_m-x_{m-1},x_{m+1}-x_m)$.另外,局部敏感度与全局敏感度之间的关系可以表示为

$$GS_f=\max_{D}(LS_f(D)) \tag{4}$$

但是,由于局部敏感度在一定程度上体现了数据集的数据分布特征,如果直接应用局部敏感度来计算噪声量则会泄露数据集中的敏感信息.因此,局部敏感度的平滑上界(Smooth Upper Bound)被用来与局部敏感度一起确定噪声量的大小.

定义 4[19]. 平滑上界.给定数据集 D 及其任意邻近数据集 D',函数 f 的局部敏感度为 $LS_f(D)$.对于 $\beta>0$,若函数 $S:D\to R$ 满足 $S(D)\geq LS_f(D)$ 且 $S(D)\leq e^{\beta}S(D')$,则称 S 为函数 f 的局部敏感度的 β-平滑上界.

所有满足这一定义的函数都可被定义为平滑上界,将局部敏感度代入此函数中则可得到平滑敏感度(Smooth Sensitivity),进而用于计算噪声大小.平滑上界与局部敏感度的关系如图 2 所示.Nissim 在文中给出了一个平滑上界的例子,并以此生成了平滑敏感度.

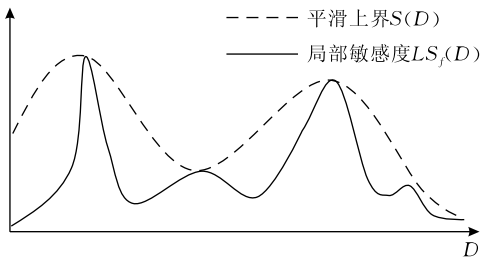


图 2 局部敏感度的平滑上界

定义 5[19]. 平滑敏感度.给定数据集 D 及 D',函数 $S_{f,\beta}(D)=\max_{D'}(LS_f(D')\times e^{-\beta|D\Delta D'|})$ 称为函数 f 的 β-平滑敏感度,其中 $\beta>0$.

由于绝大部分关于差分隐私保护的研究均针对计数查询、求和查询等敏感度较小的函数,因此,若无特殊说明,本文中敏感度均指全局敏感度.

2.2 差分隐私保护算法的组合性质

一个复杂的隐私保护问题通常需要多次应用差

分隐私保护算法才能得以解决.在这种情况下,为了保证整个过程的隐私保护水平控制在给定的预算 ε 之内,需要合理地将全部预算分配到整个算法的各个步骤中.这时可以利用隐私保护算法的两个组合性质,如图 3 所示.

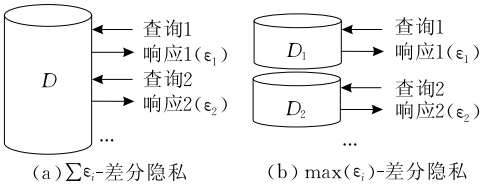


图 3　差分隐私保护算法的组合性质

性质 1[20].　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,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.

该性质表明,一个差分隐私保护算法序列构成的组合算法,其提供的隐私保护水平为全部预算的总和.该性质也称为"序列组合性".

性质 2[20].　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 $D_1,D_2,\cdots,D_n$,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.

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保护水平最差者,即预算最大者.该性质也称为"并行组合性".

2.3　实现机制

在实践中为了使一个算法满足差分隐私保护的要求,对不同的问题有不同的实现方法,这些实现方法称为"机制".Laplace 机制[21](Laplace Mechanism)与指数机制[22](Exponential Mechanism)是两种最基础的差分隐私保护实现机制.其中,Laplace 机制适用于对数值型结果的保护,指数机制则适用于非数值型结果.

2.3.1　Laplace 机制

Laplace 机制通过向确切的查询结果中加入服从 Laplace 分布的随机噪声来实现 ε-差分隐私保护.记位置参数为 0、尺度参数为 b 的 Laplace 分布为 $Lap(b)$,那么其概率密度函数为

$$p(x)=\frac{1}{2b}\exp\left(-\frac{|x|}{b}\right) \tag{5}$$

定义 6[21].　Laplace 机制.给定数据集 D,设有函数 $f:D\to R^d$,其敏感度为 Δf,那么随机算法 $M(D)=f(D)+Y$ 提供 ε-差分隐私保护,其中 $Y\sim Lap(\Delta f/\varepsilon)$ 为随机噪声,服从尺度参数为 $\Delta f/\varepsilon$ 的 Laplace 分布.

从不同参数的 Laplace 分布(如图 4)可以看出,ε 越小,引入的噪声越大.

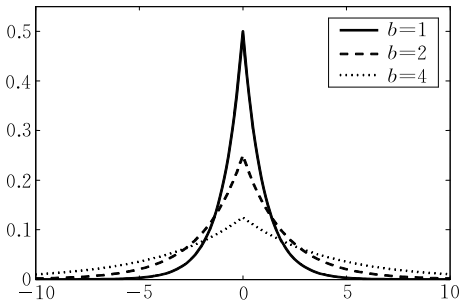


图 4　Laplace 概率密度函数

2.3.2　指数机制

由于 Laplace 机制仅适用于数值型查询结果,而在许多实际应用中,查询结果为实体对象(例如一种方案或一种选择).对此,McSherry 等人提出了指数机制.

设查询函数的输出域为 $Range$,域中的每个值 $r\in Range$ 为一实体对象.在指数机制下,函数 $q(D,r)\to R$ 称为输出值 r 的可用性函数,用来评估输出值 r 的优劣程度.

定义 7[22].　指数机制.设随机算法 M 输入为数据集 D,输出为一实体对象 $r\in Range$,$q(D,r)$ 为可用性函数,Δq 为函数 $q(D,r)$ 的敏感度.若算法 M 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,那么算法 M 提供 ε-差分隐私保护.

以下是一个指数机制的应用实例.假如拟举办一场体育比赛,可供选择的项目来自集合{足球,排球,篮球,网球},参与者们为此进行了投票,现要从中确定一个项目,并保证整个决策过程满足 ε-差分隐私保护要求.以得票数量为可用性函数,显然 $\Delta q=1$.那么按照指数机制,在给定的隐私保护预算 ε 下,可以计算出各种项目的输出概率,如表 2 所示.

表 2　指数机制应用示例

项目	可用性 $\Delta q=1$	概率		
		$\varepsilon=0$	$\varepsilon=0.1$	$\varepsilon=1$
足球	30	0.25	0.424	0.924
排球	25	0.25	0.330	0.075
篮球	8	0.25	0.141	1.5E-05
网球	2	0.25	0.105	7.7E-07

可以看出,在 ε 较大时(如 $\varepsilon=1$),可用性最好的选项被输出的概率被放大.当 ε 较小时,各选项在可用性上的差异则被平抑,其被输出的概率也随着 ε 的减小而趋于相等.

2.4 主要研究方向

由于理论上的可证明性和应用上的通用性,差分隐私保护方法得到了业内学者们的认可.近年来的一系列研究[23-30]使其在理论上不断成熟.目前差分隐私保护的理论与应用研究主要集中在两个方向,即隐私保护数据发布与隐私保护数据挖掘.

2.4.1 隐私保护数据发布

隐私保护数据发布研究的问题是如何在满足差分隐私的条件下保证发布数据或查询结果的精确性,研究内容主要集中在发布机制和算法复杂度的调整上,研究方法主要是基于计算理论和学习理论的定量分析.

差分隐私保护数据发布根据实现环境不同可分为两种,即交互式数据发布和非交互式数据发布[21],如图 5 所示.

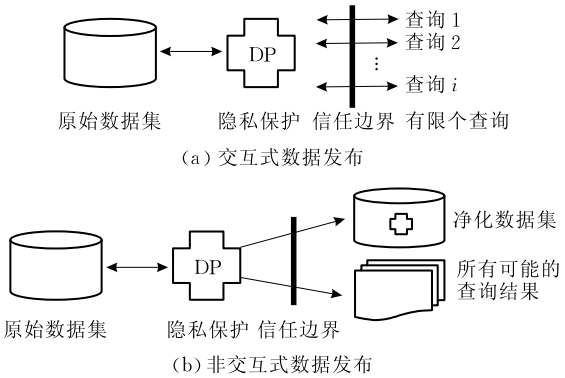


图 5 PPDR 的实现环境

在交互式环境下,用户向数据管理者提出查询请求,数据管理者根据查询请求对数据集进行操作并将结果进行必要的干扰后反馈给用户,用户不能看到数据集全貌,从而保护数据集中的个体隐私.

在非交互式环境下,数据管理者针对所有可能的查询,在满足差分隐私的条件下一次性发布所有查询的结果.或者,数据管理者发布一个原始数据集的"净化"版本,这是一个不精确的数据集,用户可对该版本的数据集自行进行所需的查询操作.

2.4.2 隐私保护数据挖掘

隐私保护数据挖掘研究的问题是如何在保证数据集隐私安全的前提下获取性能最优的数据挖掘模型.其研究通常面向数据挖掘领域的具体算法,通过对已有算法的调整和对挖掘结果的性能评估,来寻求数据安全性和模型可用性的平衡.

差分隐私保护数据挖掘有两种实现模式,即接口(Interface)模式和完全访问(Fully Access)模式[31],如图 6 所示.

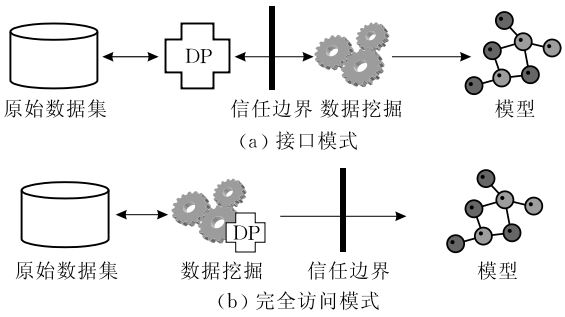


图 6 PPDM 的实现模式

在接口模式下,数据挖掘者被视为不可信的.数据管理者不会发布原始数据集,而只是对外提供访问接口,并在接口上实施差分隐私保护.数据挖掘者只能通过接口获取进行数据挖掘所需的统计类信息,其查询数目受隐私保护预算的限制.在这种模式下,隐私保护的功能完全由接口来提供,数据挖掘者无需关心任何隐私保护需求,也无需掌握任何有关隐私保护的知识,其进行数据挖掘所采用的各种算法也无需因隐私保护做任何修改.

在完全访问模式下,数据挖掘者被认为是可信的,能够直接访问数据集并执行挖掘算法.但他们必须具备隐私保护的领域知识以对传统的数据挖掘算法进行必要的修改,使得这些算法能够满足差分隐私保护的要求,从而保证最终发布的模型不会泄露数据集中的隐私信息.完全访问模式对查询数量没有限制,因此数据挖掘者在设计算法时具有更大的灵活性.

3 基于差分隐私保护的数据发布

基于差分隐私保护的数据发布是差分隐私研究中的核心内容.本节针对交互式和非交互式两种不同的实现环境,介绍差分隐私在数据发布中的应用方法.

3.1 交互式数据发布

交互式数据发布问题可表述为:给定数据集 D 和查询集合 F,需寻求一种数据发布机制,使其能够在满足差分隐私保护的条件下逐个回答 F 中的查询,直到耗尽全部隐私保护预算.发布机制的性能通常由精确度来衡量.交互式数据发布即是要在满足一定精确度的条件下,以给定的隐私保护预算回答尽可能多的查询.

对于 F 中任意查询 f,设定一个足够小的实数

$\delta<1$，查询结果的精确度 α 应满足

$$Pr_{f\in F}[|f(D)-M(f(D))|\leq\alpha]\geq 1-\delta \quad (6)$$

其中 $f(D)$ 为查询 f 在 D 上的结果，$M(f(D))$ 为随机算法 M 对 $f(D)$ 的干扰结果. α 越小，精确度越高[32].

为了达到这个精确度所需的样本量为样本复杂度. 精确度与样本复杂度之间的关系有两种不同的表达方式. (1) 给定样本量 n，则发布机制的精确度为 $\alpha(n)$，当 $n\to\infty$ 时，$\alpha(n)$ 趋于 0；(2) 给定 α，达到这一精确度所需的样本量为 $n_0(\alpha)$，则发布机制在 $n\geq n_0(\alpha)$ 时保证了精确度 α，其中当 $\alpha\to 0$ 时 $n_0(\alpha)$ 趋于无穷大. 交互式数据发布采用精确度和相应的样本复杂度来评估发布机制的性能.

交互式隐私保护数据发布的研究主要集中在发布机制和基于直方图的发布方法上. 两者的区别在于，前者直接对数据集进行操作来响应查询，而后者先根据数据集建立直方图分布，然后根据直方图分布来响应查询.

3.1.1　交互式发布机制研究

最早用于交互式数据发布的差分隐私保护机制是 Dwork 等人提出的 Laplace 机制. 在该机制下，根据查询函数的敏感度和隐私保护预算 ε 产生服从 Laplace 分布的噪声，并添加到每个查询结果中. 这种简单的机制能够处理各种类型的查询，但缺点是查询的数量有限，与数据集记录数为次线性关系. 另外，在干扰针对连续属性的查询结果时会产生较大的噪声.

之后，Roth 和 Roughgarden[33] 提出了中位数机制(Median). 相对于 Laplace 机制，中位数机制能够在相同预算下提供更多数量的查询. 在该机制下，查询被分为“难查询”和“易查询”两类. 其中，“易查询”的结果可以根据“难查询”的结果来确定，因此“易查询”就无需消耗任何预算. 他们的研究证明，给定域 Z 和 k 个查询，“难查询”的数量级为 $O((\log_2 k)(\log_2|Z|))$，其它均为“易查询”. “难查询”的结果通过独立的 Laplace 噪声进行干扰，而“易查询”的结果则用之前查询结果的中位数来确定. 中位数机制的缺点则在于其算法的时间复杂度会随着数据集容量的增长呈指数增长，同时，其样本复杂度也是超多项式的.

Hardt 等人[34] 提出了另一种有效的机制，即 PMW(Private Multiplicative Weights). 该机制的理论来源是机器学习中的加权多数算法(Weighted Majority Algorithm)，该算法用于通过投票机制来构建一个复合算法. 与之相似，PMW 也采用了一种投票机制来减少隐私保护预算的消耗，使得该机制能够在给定的隐私保护预算下，回答更多的查询. 具体方式为，PMW 把数据集在数据域上的分布视作一个直方图，首先将每个频数设为相同，然后等待查询，每个查询的结果加上 Laplace 噪声后会和上一次查询结果相比较，若差异小于设定的阈值，则发布上一次查询结果的值，此步骤不耗费隐私保护预算. 只有当差异大于此阈值时，才会发布新的查询结果，并调整直方图中相应频数的值. 由于很多查询并不耗费隐私保护预算，所以这个机制能比普通 Laplace 机制回答更多的查询. 在精确度方面，对于 k 个查询，每个结果的误差为 $O(\sqrt{(\log k)/n})$. 此方法在提高查询数量的基础上较好地保证了精确性，但缺点是只能处理计数类型的查询. 因此，针对复合线性查询的噪声复杂度问题，Hardt 等人[35] 又提出了 K-norm 机制，将差分隐私保护的噪声复杂度作为高维凸体的几何属性来研究. 结果表明，该机制下每次查询的噪声量为

$$O(\min\{k/\varepsilon,\sqrt{k\log(n/k)}/\varepsilon\}) \quad (7)$$

在 $k\ll n$ 时小于 Laplace 机制下的噪声量为 $O(\min\{k/\varepsilon,\sqrt{n}/\varepsilon\})$. 但由于该机制包含在高维凸体上的随机采样操作，导致了较高的计算复杂度.

Gupta 等人[36] 提出一种更为通用的迭代数据集生成架构(IDC Framework)，并进一步证明之前的 Median 和 PMW 机制都是此架构的特例. 在该架构下，对于一个查询集合 F，首先在定义域空间中任意选择一个数据集，作为初始数据集假设，然后用此数据集来回答所有的查询，若发现有某个查询结果和真实结果之间的差别大于预定义的阈值时，则根据此查询结果来更新数据集假设，使更新后的数据集能在阈值范围内回答此查询. 这个迭代过程循环进行，直到所有的查询结果和真实结果的差异不再大于阈值. 由于迭代次数会少于总查询数量，所以耗费的隐私保护预算会比普通机制更少，从而降低噪声总量.

以上几种交互式发布机制的比较如表 3 所示.

表 3　交互式发布机制比较

机制	查询数量级	计算效率	精确度
Median[33]	n 的指数	高效	$\frac{n^{2/3}(\log k)(\log\lvert Z\rvert)^{1/3}}{\varepsilon^{1/3}}$
PMW[34]	n 的指数	非高效	$\frac{n^{1/2}(\log k)(\log\lvert Z\rvert)^{1/4}}{\varepsilon}$
K-norm[35]	n 的次线性	非高效	$\frac{\sqrt{k}}{\varepsilon}\left(\log\left(\frac{\lvert Z\rvert}{k}\right)\right)^{1/2}$
IDC[36]	n 的指数	非高效	$\frac{n^{1/4}(\log k)^{1/2}(\log\lvert Z\rvert)^{1/4}}{\varepsilon^{1/2}}$

另外,还有一些针对特殊查询的发布机制,如布尔连接查询(Boolean Conjunction Query)、半空间范围查询(Halfspace Range Query)等.

Gupta 等人[37]研究了在数据集只提供统计查询的前提下,回答布尔连接查询集合 F 所需的样本复杂度问题.他们证明这种查询所需的样本复杂度等于不可知学习的复杂度,进而把查询集合 F 转化为使用次模(Submodular)函数描述的问题,并给出了此机制下布尔连接查询的样本复杂度为

$$n=\frac{1}{\varepsilon}d^{O(\log(1/\delta)/\alpha^2)} \tag{8}$$

其中 d 为数据集维度.

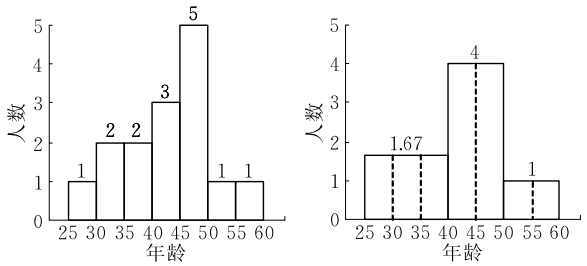
半空间区域查询实际上是一种高维空间里的范围查询.它用于回答"当定义好一个超平面时,有多少个点会落在超平面上方的空间里".Muthukrishnan 等人[38]对半空间查询的发布机制进行了研究,他们将高维空间的范围查询分解成多个范围,并应用差异理论(Discrepancy Theory)来确定范围的个数和每个范围的大小,然后分别对这些范围添加 Laplace 噪声来保证差分隐私,并达到最优的隐私性和可用性平衡.在该机制下,半空间查询的均方误差下界为 $\Omega(n^{1-1/d})$.

交互式发布机制的研究方法是应用计算理论来分析各种发布机制在保证差分隐私的前提下查询结果的精确度,这些研究的结论是差分隐私在数据挖掘、机器学习等领域应用研究的理论基础.

3.1.2 基于直方图的发布

直方图是一种直观的数据分布表示形式,其结果可作为其它统计查询或线性查询的依据.在差分隐私保护条件下,删除数据集的一条记录只会影响直方图中一个数据格(Bin)的结果,因此计算计数查询以及其他相关查询的敏感度是比较容易的,这使得根据直方图来为各种查询提供差分隐私保护的响应是可行的.

在形成直方图时,需要根据属性(或属性组合)的 w 个不同等级将数据集划分为 w 个数据格,然后分别统计每个数据格的频数.为了提供差分隐私保护,一种简单的方法是向 w 个数据格的频数分别添加独立的 Laplace 噪声,也称为基于数据格的方法.因为频数统计函数的敏感度仅为 1,所以这种方法对于 w 较小时(例如二维直方图)是有效的.但对于多维直方图而言,多个变量的组合可以形成大量的数据格,这时会使一些范围计数查询(Range Counting Query)的结果因累积噪声过大而失效.例如,图 7(a)中显示了一个按照年龄段对人数进行统计的直方图,如果为每个频数加入噪声 Y,则总体噪声为 $7Y$.为了降低噪声,一种有效的方法是将所有数据格合并为若干个分区(Partition),每个分区的频数为其中全部频数的平均值,如图 7(b)所示,然后为每个分区频数加入噪声,总体噪声为 $3Y$.但是,分区后数据格的频数相对于分区前的频数产生了一定的误差.因此,如何结合数据分布的特征寻求合理的分区方案,在减小 w 的同时尽可能降低频数误差,成为直方图发布的主要研究内容.



(a) 原始直方图 (b) 分区直方图

图 7 直方图分区方案示例

显然,如果一个分区方案能够将直方图划分为尽可能少的分区且每个分区中数据格频数彼此接近,就能降低噪声并减少查询结果的误差.因此,Xiao 等人[39]提出了一种基于 k-d 树的直方图发布算法.算法首先根据给定的数据集及其 k 个属性产生原始直方图,得到所有数据格及其频数,用 $\varepsilon/2$ 的预算为所有频数加入 Laplace 噪声.然后以这些干扰后的频数作为 k 维空间的数据点,采用 k-d 树算法对空间进行递归划分,在每一次迭代中,首先计算当前分区 P_0 中频数分布的紧密程度为

$$L(P_0)=\sum_{cell_i\in D_0}|count(cell_i)-a_0| \tag{9}$$

其中 $cell_i$ 是 P_0 中的一个数据格,$count(cell_i)$ 为其频数,a_0 是 P_0 中的所有数据格频数的平均值.如果 $L(P_0)$ 大于预定义的阈值,则将该分区划分为两个子分区.k-d 树算法最终将空间划分为若干个分区,这个划分的结果实际上代表了一种分区方案.最后,以这个分区方案对原始直方图进行分区,并以 $\varepsilon/2$ 的预算向所有分区的实际频数加入噪声并发布.在 Xiao 等人的后续研究里,这一算法被命名为 DPCube[40].与基于数据格的方法相比,当参数(频数分布紧密度阈值、空间分割次数)的取值适当时,DPCube 算法在查询数量和查询误差等方面具有更好的性能.

同样针对寻求最优分区方案问题，Xu[41]等人采用平方误差和（SSE）来衡量一种分区方案的优劣程度：

$$SSE(P)=\sum_{j=1}^{l}\sum_{cell_i\in P_j}(b_j-count(cell_i))^2 \quad (10)$$

其中 $P=\{P_1,P_2,\cdots,P_l\}$ 表示一种分区方案，由 l 个分区构成，b_j 为 $P_j(j=1,\cdots,l)$ 的估计频数，$cell_i$ 为 P_j 中任一数据格，$count(cell_i)$ 为 $cell_i$ 的频数. 显然，SSE 越小，查询精确度就越高. 基于以上原理，他们提出了两种实现算法，即 NoiseFirst 和 StructureFirst 算法. 其中 NoiseFirst 算法基于 Laplace 机制，StructureFirst 算法则基于指数机制. 实验结果表明，NoiseFirst 算法所产生的直方图用于短范围查询（Short Range Query）时精确度更高，而 StructureFirst 算法产生的直方图更适用于长范围查询（Long Range Query）.

基于隐私保护的直方图发布所面临的另一个问题是范围查询结果的一致性问题. 例如查询序列 $F=\{f_1,f_2,f_3\}$ 的原始结果为 $E=\{e_1,e_2,e_3\}$，查询结果存在的约束关系为 $e_3=e_1+e_2$，向 E 中加入独立的 Laplace 噪声后输出为 $\widetilde{E}=\{\tilde{e}_1,\tilde{e}_2,\tilde{e}_3\}$，就有可能使结果违背原有的约束关系. 对此，Hay 等人[42]提出了一种对差分隐私保护结果进行后处理（Post-processing）的方法. 该方法用一棵树来表示一个查询序列，树中每个节点表示一个范围查询，查询之间的约束关系则表现为节点与其子节点之间的连接关系. 设 $\tilde{e}[v]$ 为节点 v 的差分隐私保护输出，采用线性组合方法在约束关系下计算最接近 $\tilde{e}[v]$ 的无偏估计 $\bar{e}[v]$，所有节点的估计值构成最终的发布序列 $\bar{E}$. 理论证明和实验结果均表明，$\bar{E}$ 不但在差分隐私保护机制下保证了一致性，而且由于利用了查询间的约束关系有效降低了查询误差.

由于直方图结构能够降低查询敏感度，所以成为目前差分隐私保护中常用的数据结构，在其基础上进一步进行发布机制的研究是目前差分隐私研究的重要内容之一.

3.2　非交互式数据发布

非交互式数据发布问题可表述为：给定数据集 D 和查询集合 F，需寻求一个数据发布机制，使其能够在满足差分隐私保护的条件下一次性回答 F 中所有的查询.

早期的差分隐私保护研究认为数据发布很难在非交互式环境下实现隐私保护. Dinur 等人[43]曾提出，一个数据集如果精确回答了超过次线性个查询，那么用户就能够以很高的概率还原出原始数据集. 因此，如果要在非交互式环境下回答查询，或者发布一个被净化处理的数据集，必须在发布的内容中加入大量噪声，但这会极大地破坏其可用性. 所以，早期的研究大多集中在交互式环境下的查询数据发布. 但随着应用要求的提高，单纯的交互式环境在查询数量和应用方式上存在许多局限，从而促进了非交互式环境下数据发布的隐私保护研究.

非交互式数据发布的研究主要集中在批查询、列联表发布、基于分组的发布方法以及净化数据集（Sanitized Dataset）发布方法上.

3.2.1　批查询（Batch Query）

数据管理者针对所有可能的查询，一次性对外发布所有查询的结果，这种模式称为批查询. 在批查询模式下，由于各查询之间彼此相关，删除数据集中任一记录有可能会导致多个查询结果发生改变. 因此，基于差分隐私保护的批查询函数具有比单一查询高得多的敏感度.

例如，图 8 为一个根据原始数据集生成的频数统计变量集 $U=\{x_1,x_2,x_3,x_4\}$ 及其范围查询集合 F，$F=\{f_1,\cdots,f_{10}\}$ 为所有可能的范围查询. 显然，删除原始数据集中任一记录，最多可以使 6 个查询结果发生改变，根据定义 2，F 查询集合敏感度为 6.

等级	计数	统计变量
90～100	12	x_1
80～89	24	x_2
70～79	3	x_3
60～69	8	x_4

f_1	x_1	+	x_2	+	x_3	+	x_4
f_2	x_1	+	x_2	+	x_3		
f_3			x_2	+	x_3	+	x_4
f_4	x_1	+	x_2				
f_5			x_2	+	x_3		
f_6					x_3	+	x_4
f_7	x_1						
f_8			x_2				
f_9					x_3		
f_{10}							x_4

图 8　批查询示例

当 U 中元素的数量为 n 时，如果直接采用 Laplace 机制为每个范围查询加入独立的噪声（记为 Lap1 方法），则查询的总敏感度为 $O(n^2)$，每个查询误差的方差为 $O(n^4/\varepsilon^2)$. 或者，先为每个单一查询加入噪声，然后根据干扰后的单一查询计算其它范围查询的结果（记为 Lap2 方法），这种方法虽然敏感度仅为 1，但由于噪声的累加，在最坏情况下，查询误差的方差为 $O(n/\varepsilon^2)$.

以上两种方法都是对 Laplace 机制的直接应

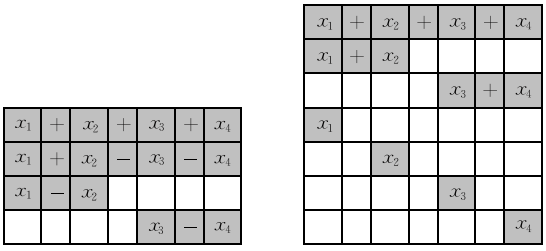
用,由于噪声过大而难以满足实际应用要求.

目前,批查询主要通过映射和变换查询集合来降低总敏感度,从而降低噪声量.比较有代表性的有Xiao等人[44]提出的小波变换方法(Privelet)、Hay等人[42]提出的层次查询方法等.

Privelet方法通过Haar小波变换将频数统计矩阵$\boldsymbol{U}$映射到小波系数矩阵$\boldsymbol{U}_C$,然后为$\boldsymbol{U}_C$中的每个小波系数添加Laplace噪声.由于在计算小波系数时提交的查询的总敏感度为$(\log n+1)$,因此这里加入的噪声服从$Lap((\log n+1)/\varepsilon)$分布,另外查询集合中的任一查询均由相应的$(\log n+1)$个小波系数的线性组合来表示.所以,最终每个查询的噪声的方差为$O((\log n)^3/\varepsilon^2)$,当$n$较大时(例如$n>1024$)小于Dwork的方法中加入的噪声量($O(n/\varepsilon^2)$).另外,Xiao等人[44]还提出了针对多维数据集的小波变换方法,其平均噪声方差为$O((\log n)^{3d}/\varepsilon^2)$,$d$为数据集维度.

Hay等人提出的层次查询方法和小波变换类似,也能将敏感度降低到$\log n+1$,每个查询所需噪声的方差也为$O((\log n)^3/\varepsilon^2)$,但并未提及多维数据集的查询方式.

以图8中的频数统计变量为例,小波变换和层次查询方法所提交的查询如图9所示.



(a) 小波变换提交查询 (b) 层次查询提交查询

图9 频数统计矩阵转换示例

表4列出了以上几种方法中F的总敏感度以及加在每个查询上的噪声方差.

表4 批查询的敏感度及噪声方差

批查询方法	总敏感度	噪声方差
Lap1[21]	$O(n^2)$	$O(n^4/\varepsilon^2)$
Lap2[21]	1	$O(n/\varepsilon^2)$
Privelet[44]	$\log n+1$	$O((\log n)^{3d}/\varepsilon^2)$
Hierarchical ranges[42]	$\log n+1$	$O((\log n)^3/\varepsilon^2)$

类似的研究还有Li等人[45]提出了基于负载矩阵分解的方法;Cormode等人[46]提出的基于Quadtree的查询,用于实现二维数据集的查询;以及Barak提出的傅里叶变换方法[47]等.

3.2.2 列联表发布方法

列联表是对数据集中的记录按照变量进行分类时所列出的频数表,它是非交互式数据发布的一种特殊形式.例如数据集D包含n个记录,由k个布尔变量组成,列联表就是对数据集按照2^k个可能的组合值进行统计计数所形成的表格.事实上,在数据分析研究中发布的内容通常并非列联表本身,而是按照多个变量的组合值进行统计所得的计数,也称为边缘频数(Marginal).研究表明,虽然边缘频数只是关于数据集的某些记录的统计数据,在某些条件下却可能会泄露相关记录的隐私信息[48-51].

差分隐私保护在列联表发布中的应用主要有两种方法.其一是向列联表的每个单元格中加入噪声[52],用户可以根据被干扰的频数计算每个边缘频数.这种方法能够维持所有边缘频数之间的一致性,但边缘频数的累积噪声也会较大.另一种方法是先计算出拟发布的边缘频数,然后对它们加入噪声后再发布.由于每个边缘频数只添加了一次噪声,所以数据可用性更好.但因为每个边缘频数都是独立地添加噪声,因此所发布的各边缘频数之间可能会违背数据上的一致性.

针对在实现列联表隐私保护过程中数据准确性与一致性之间的矛盾,Barak等人[47]提出了一个综合考虑准确性与一致性的整体解决方案,将列联表的数据进行傅里叶转换,就能对边际频数进行无冗余的编码.加入傅里叶域的噪声不会破坏一致性,因为任一组傅里叶系数都对应着一个一致性列联表,计算低阶的边际频数只需要很少的傅里叶系数,因此加入到边际频数的噪声就会很小.然而,Fienberg等人[53]以及Yang等人[54]的研究表明,文献[47]提出的方法并不适用于稀疏数据集和非布尔变量构成的数据集,在这些条件下的列联表发布问题还需要进一步的研究.

3.2.3 基于分组的数据发布方法

基于分组的数据发布方法将早期的匿名泛化技术应用到非交互式环境下来实现差分隐私保护.

k-anonymity及其衍生模型(如l-diversity模型和t-closeness模型)属于典型的基于分组的隐私保护方法.研究表明,这些模型对攻击者所掌握的知识假定过少[55],并不能提供足够的安全保障.但一些学者也认为差分隐私保护对攻击者所掌握的知识假定过多(假定攻击者掌握了数据集中除攻击目标之外所有其它个体的信息),对安全的要求过于严

格. Li 等人[56]研究了 k-anonymity 与差分隐私保护模型各自的优缺点，他们认为，通过对原始数据集进行随机抽样，可以增加数据集的不确定性，相当于减少了差分隐私保护中对攻击者所掌握的知识假定，因此提出了一种将 k-anonymity 模型和差分隐私保护模型结合起来的新方法并称之为“安全 k-匿名”模型. 该方法首先确定一个随机抽样函数 A_m，并对数据集 D 进行抽样，得到 $S_D=A_m(D)$，最后从 S_D 中删除出现次数小于 k 的记录. 理论证明，当函数 A_m 满足 ε-差分隐私保护条件时，安全 k-匿名模型能够提供 (ε,δ)-差分隐私保护[23]. 该研究对实现 (ε,δ)-差分隐私保护提供了一个新的思路，但其缺点在于仅仅提出了一个理论框架，对于如何实现一个满足 ε-差分隐私保护条件的抽样函数等问题并未给出具体解决方案.

基于分组的方法也被用于面向数据挖掘算法的差分隐私保护问题. Mohammed 等人[57]采用“自顶向下、逐步细分”的策略，提出了一种针对决策树分析的差分隐私保护数据发布算法 DiffGen. 该算法首先将数据集完全泛化，然后进入细分迭代循环. 一个逐步细分的决策树示例如图 10 所示，DiffGen 算法要保证决策树的生成过程满足差分隐私保护的要求.

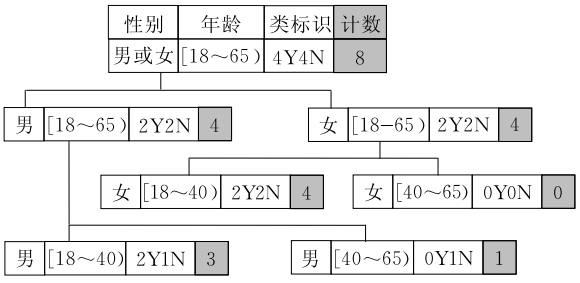


图 10　决策树逐步细分示例

设一轮迭代中有 y_1 个离散属性值和 y_2 个连续属性值. 按照细分层次树，y_1 个离散属性值产生 y_1 个细分方案. y_2 个连续属性值则产生了 y_2 个细分方案集 $N_1,N_2,\cdots,N_{y_2}$. 对每个连续属性值的细分方案集 $N_j,j=1,2,\cdots,y_2$，应用指数机制从 N_j 中选择一种方案，总共得到 y_2 种方案，然后将它们与 y_1 个离散属性的细分方案放在一起，采用指数机制从这 y_1+y_2 种方案中选择一种. 循环这一过程，直到达到预定的细分次数.

虽然 DiffGen 算法是满足 ε-差分隐私保护要求的，但其缺点在于没有充分利用给定的预算，导致加入不必要的冗余噪声. 因为算法在每次迭代中都为连续属性的处理分配了一定的隐私保护预算，但事实上只有在上一次迭代时选择连续属性细分方案的前提下，本次迭代才需要重新处理连续属性. 因此，除非每次迭代均选择连续属性细分方案，整个预算 ε 才能够得到完全充分的利用.

针对这个问题，Zhu 等人[58]对 DiffGen 算法进行了改进，提出了一种新的决策树构建算法. 该算法在每次细分迭代中，将所有连续属性细分方案乘以相应的权重后和离散属性细分方案一起构成候选方案集，然后调用指数机制来选择细分方案. 该算法减少了调用指数机制的次数，从而提高了隐私保护预算的利用率，使得在给定的隐私保护预算下，数据集能够更大程度地精确化，进而提高分类模型的准确率.

基于分组的数据发布延续了 k-anonymity 模型的思想，并用差分隐私的要求来控制分组及匿名处理的整个过程. 从已有的研究来看，隐私保护预算在算法中的合理分配与充分利用是基于分组的数据发布需要继续研究的问题.

3.2.4　净化数据集发布方法

净化数据集是对原始数据集进行隐私保护处理后发布给用户的数据集. 直接发布一个满足差分隐私的净化数据集来让用户进行任意的查询，一直被认为是一个十分困难的问题. 因为这会引入非常大的噪声，并覆盖整个数据集的原始数据[28]. 但随着差分隐私保护理论研究的深入，研究者发现如果将学习理论引入差分隐私保护，可以在一定程度上解决非交互式数据发布中精确度较低的问题.

从学习理论的角度来看，对一个数据集进行统计分析的目的是为了获得数据集某个群体的信息. 如果将一个数据集 D 视为从某个分布 χ 随机抽取的样本集合，而且我们只关心对给定类别 C 的布尔值预测问题，那么当数据集的样本复杂度达到 $O(VC(C)/\alpha^2)$ 时，对于任意查询 $f\in C$，D 中正例的比例与 χ 中正例的比例误差在 $\pm\alpha$ 之内. 这是学习理论的一个重要结论[59]. 如果能将这个结论引入差分隐私，则有可能找到 D 的净化数据集 $\hat{D}$，对于某一类别 C 的所有查询，其精确度为 α.

Kasiviswanathan[60]和 Blum[61]等人分别就这个问题的两个方面进行了研究. 其中 Kasiviswanathan 等人关注的是在差分隐私的条件下学习理论是否成立. Blum 则研究在差分隐私学习成立的前提下，

是否能在定义域 Z 中搜索到满足条件的净化数据集 $\hat{D}$.

首先,Kasiviswanathan 等人[60]结合差分隐私保护与学习理论,提出了隐私保护学习(Private Learning)的概念,从计算学习理论的角度对差分隐私保护条件下学习的可行性进行了必要的分析,即在差分隐私保护条件下,成功地运行一个学习算法需要多少样本复杂度和计算复杂度.他们在概率近似正确(PAC)框架[62]下对差分隐私保护下的 PAC 学习进行了定义.

定义 8[60]. 差分隐私 PAC 学习. 设概念类别 C 定义在实例集合 X 上,如果存在一个满足 ε-差分隐私保护的算法 A,使用假设空间 H,对所有概念 $c\in C$ 和所有 X 上的分布 χ 以及精确度和可用性参数 $\alpha,\delta\in(0,1/2)$,算法 A 输出假设 $h\in H$,使得

$$Pr[err(h)\leq\alpha]\geq 1-\delta \tag{11}$$

成立,$err(h)$ 为 h 对分布 χ 上的样例的分类错误率,则称 C 在差分隐私保护条件下是可 PAC 学习的,其所需样本数量是 $1/\varepsilon$、$1/\alpha$ 和 $\log(1/\delta)$ 的多项式函数,计算时间则是 $1/\alpha$ 和 $\log(1/\delta)$ 的多项式函数.

对于 PAC 扩展后的不可知学习[63]则在相同条件下输出 h 的概率满足

$$Pr[err(h)\leq OPT+\alpha]\geq 1-\delta \tag{12}$$

其中 $OPT=\min_{c\in C}\{err(c)\}$ 为 C 中概念的最小分类错误率.在以上定义的基础上,一个通用的基于差分隐私保护的不可知学习器 $A_q^{\varepsilon}(z)$ 被提出.函数 $A_q^{\varepsilon}(z)$ 采用指数机制以正比于 $\exp(\varepsilon q(z,h)/2)$ 的概率输出($h\in H$),其中 $q(z,h)=-m_h$ 为可用性函数,m_h 为 h 对数据集 z 的误分类样例数.可以证明,$A_q^{\varepsilon}(z)$ 是满足差分隐私保护要求的.同时,概念类别 C 使用假设空间 $H=C$ 和学习器 $A_q^{\varepsilon}(z)$ 是不可知学习的,其样本复杂度为

$$O((\ln|H|+\ln(1/\delta))\cdot\max\{1/(\varepsilon\alpha),\alpha^2\}) \tag{13}$$

对于无限假设空间,则样本复杂度为

$$O((VC(C)\cdot\ln|X|+\ln(1/\delta))\cdot\max\{1/(\varepsilon\alpha),\alpha^2\}) \tag{14}$$

其中 $VC(C)$ 为 C 的 VC 维.

Kasiviswanathan 的结论表明,如果一个概念类别在无隐私保护要求和多项式样本复杂度下是可学习的,那么它在差分隐私保护条件下也是可学习的.这个结论验证了在学习理论中引入差分隐私的可行性,使之能被应用于更广泛的领域.

另一方面,Blum 等人[61]的研究证明,给定一个由离散属性构成的域,对于任意具有多项式 VC 维的概念类别,如果不考虑计算复杂度,利用指数机制遍历整个数据域,最终寻找出满足差分隐私保护的数据集是可行的,且用此数据集能以指定的精确度回答所有关于这个概念类别的查询.同时,他们给出了满足精确度 α 时数据集的最小样本复杂度为

$$O\left(\frac{dVC(C)\log(1/\alpha)}{\alpha^3\varepsilon}+\frac{\log(1/\delta)}{\alpha\varepsilon}\right) \tag{15}$$

其中,d 为数据集维度,δ 为可用性参数.

这两个研究为净化数据集的发布提供了理论依据和相关实现方法.但是这一方法具有较高的计算复杂度,为数据域大小 $|Z|$ 和 $|C|$ 的超多项式函数,在实际中很难应用.围绕这个问题,学者们提出了一些解决方法.例如,Dwork 等人[29]将所有概念类别划分为若干子集,然后采用一种递归的方式来为各个子集构建发布数据集,最终形成可发布的总数据集.类似的研究还包括 Hardt 等人[64]提出的基于阈值学习的隐私数据发布算法、Dwork 等人[65]提出的 Boosting 算法等.

总之,学习理论拓展了非交互式发布的研究领域,证明了在保证精确度的情况下发布针对某类查询的净化数据集是可行的.但研究的难点在于如何降低计算复杂度,如何处理数值型数据以及如何提供更广泛的查询类型等问题上.

3.3 PPDR 小结

数据发布一直是差分隐私研究的核心内容,其研究进展直接影响到差分隐私在其它相关领域的应用.本节对基于差分隐私保护的数据发布方法进行分类并分析了各自的特点(如表 5 所示).可以看出,虽然目前差分隐私在数据发布上获取了很大进展,但仍然有些关键问题需要进一步解决:

(1)高敏感度查询问题.差分隐私针对低敏感度查询的效果较好,例如敏感度为 1 的计数查询,其噪声方差仅为 $2/\varepsilon^2$.但实际应用中,会遇到很多高敏感度查询,例如查询最大值,其敏感度可能远远大于 1.这时加入的噪声往往会覆盖原有数据,造成数据可用性急剧下降.

(2)计算复杂度问题.大部分数据发布机制的计算复杂度都是非高效(Inefficient)的,超过 n 的多项式阶,甚至达到指数阶.高复杂度限制了差分隐私在实际应用中的效率,成为目前需要解决的问题之一.

表 5　基于差分隐私保护的数据发布方法分类比较

方式	方法	方法描述	典型机制或方法	优点	缺点
交互式	交互式查询	对原始数据集查询产生结果，加噪后发布	Laplace[21]，Exponential[22]，Median[33]，PMW[34]，K-norm[35]，IDC[36]	容易实现，可满足所有查询类型	噪声较大，查询次数有限
	直方图发布	由原始数据集产生加噪后的直方图，根据直方图响应查询	KD-tree[39]，DBCube[40]，NoiseFirst[41]，StructureFirst[41]，P-HPartition[66]，Constrained Inference[42]	敏感度小，分析简单，噪声可以控制在较小范围内	查询类型受限制，查询次数有限
非交互式	批查询发布	数据管理者针对所有可能的查询，一次性对外发布所有查询的结果	Privelet[44]，Hierarchical ranges[42]，Matrix mechanism[45]，Low-rank mechanism[67]，Quad-tree[46]，Adaptive Mechanism[68]	容易实现，可满足所有查询类型	噪声较大，但可以采用不同机制降低到一定程度
	列联表发布	对数据集中的记录按 k 个属性的排列组合产生 k 维频数表，加噪后发布	Fourier basis[47]，Non-uniform strategy[69]，Correlated Row[52]	可满足大部分查询类型	高维度列联表噪声大
	分组发布	对原始数据集进行泛化处理并发布	Safe k-anonymization[56] DiffGen[57]，DT-Diff[58]	结合泛化和差分隐私方法，容易实现	隐私保护预算分配主观性大
	净化数据集发布	对原始数据集加入噪声后产生净化数集并对外发布	Exponential Searching[60]，Boosting[65] Recursive[28]，Threshold Learning[64] Learning theory approach[61]	可满足多种查询类型，查询次数可达到 n 的指数阶	时间复杂度高，实现困难，噪声大

4　基于差分隐私保护的数据挖掘

基于差分隐私保护的数据挖掘是差分隐私研究的另一个重要方向，本节根据差分隐私在 PPDM 中的不同应用模式，分别介绍接口模式和完全访问模式下的各种差分隐私保护数据挖掘算法，并对其性能进行分析和比较.

4.1　接口模式下的数据挖掘

在接口模式下的数据挖掘研究中，有两个被广泛应用的接口框架，即 SuLQ 和 PINQ，两者都采用了 Laplace 机制作为实现差分隐私的主要方式. SuLQ 框架由 Blum 等人[70]提出. 首先，他们将最简单的单属性布尔查询定义为查询原语(Primitive)[71]，然后以查询原语为基本操作来组合查询集合，可以实现更加复杂的查询函数，只要查询的数量是记录数量的次线性函数，即可以较小的噪声实现足够的隐私保护. 在此基础上，Blum 等对 SuLQ 原语做了两个方面的扩展后形成了 SuLQ 框架：(1)将处理的数据类型从布尔型数据扩展到连续型数据；(2)以 SuLQ 原语为基本算子，设计提供隐私保护功能的复杂算法，如 k-means 算法、ID3 分类器以及统计查询学习模型等.

PINQ[20]是由 McSherry 等人开发的一套为数据提供差分隐私保护的框架，它基于 LINQ 查询语言并提供一系列便于二次开发的应用程序接口，其中定义的 Partition 算子允许在查询中对数据集进行分割. 由于 Partition 算子可将数据集分割成不相交的子集，因此可以利用差分隐私保护算法的并行组合性，提高隐私保护预算的利用率.

4.1.1　接口模式下的分类算法

分类算法[72]旨在根据训练数据集建立分类器模型，用以推测新记录的类标识. ID3[73]是最经典的分类算法之一，它以信息增益为标准对训练数据集进行迭代划分，从而建立一棵决策树. SuLQ 框架提出了实现差分隐私保护的 SuLQ-based ID3 算法[70]. 其基本思想是在每次计算属性的信息增益时，使用加入噪声的计数值，最终生成相应的决策树. 此方法虽然可以保证差分隐私，但缺点是噪声过大. 从文献[31]中对模拟数据集的实验结果来看，在隐私保护预算小于 1 的情况下，SuLQ-based ID3 算法相对于无隐私保护功能的 ID3 算法，其预测准确率大约降低了 30%.

Friedman 和 Schuster 基于 PINQ 平台对 SuLQ-based ID3 算法进行了改进[31]，利用其中的 Partition 算子将数据集分割成不相交的子集，然后再实现 ID3 算法. 虽然在计算信息增益的过程中应用 Partition 算子避免了不必要的预算消耗，但由于为计算信息增益而进行的计数查询必须为每个属性单独执行，因此整个预算必须事先分配给每一次计数查询，这导致每个查询的预算相对很小，所以无法显著降低 SuLQ-based ID3 所引入的噪声.

针对这个问题，Friedman 和 Schuster 进一步在 ID3 算法中应用了指数机制来实现差分隐私保护，提出了 DiffPID3 算法[31]. 由于在指数机制下，只需一个查询即可实现一次对全部属性的评估，决策树的一次分裂只需消耗一次预算，因此每个查询所分配的预算较大，有效降低了噪声. 另外，通过将离散

属性的处理扩展到连续属性，Friedman 和 Schuster 还提出了 DiffP-C4.5 算法[31]. 在实际数据集上的测试表明，DiffPID3 算法和 DiffP-C4.5 算法的分类准确率较 SuLQ-based ID3 算法有极大的提高，在样本量足够大和 ε 等于 1 的条件下均能获得大于 80% 的分类准确率. 但是，DiffP-C4.5 算法的缺点在于，在每一次迭代中必须先用指数机制对所有连续属性选择分裂点，然后将所得结果与全部离散属性一起再次通过指数机制选择最终的分裂方案，由于每次迭代需要两次调用指数机制，因此消耗了过多的隐私保护预算.

对以上几种算法的归纳如表 6 所示. 可以看出，直接采用 Laplace 机制需要较大的噪声干扰. 在算法中应用指数机制则可以提高隐私保护预算的利用率，有效降低噪声.

表 6 基于接口的分类算法

分类算法	实现机制	噪声	数据类型
SuLQ-based ID3[70]	Laplace 机制	高	离散
PINQ-based ID3[31]	Laplace 机制	高	离散
DiffPID3[31]	Laplace 及指数机制	低	离散
DiffP-C4.5[31]	Laplace 及指数机制	低	离散或连续

4.1.2 接口模式下的聚类算法

作为一种无监督学习方法，聚类算法将无类标识的记录划分到若干个簇中，使得簇内记录相似度高，而簇间相似度低. 设聚类算法的输入为一数据集，输出为 k 个聚类，基于差分隐私保护的聚类算法的目标则是在从数据集中删除任一记录时，保证每个聚类的质心以及记录数量所发生的改变不泄露隐私信息.

在 SuLQ 框架里，Blum 等人给出了提供差分隐私保护的 k-means 算法. 由于在计算每个记录与质心的距离时会泄露隐私，因此在 SuLQ 框架下通过发布聚类质心和记录数量的估计值来满足隐私保护的要求. 但根据全局敏感度的定义，查询聚类质心的函数敏感度为聚类的最大直径，而以此敏感度计算出的噪声量较大，降低了聚类结果的可用性.

为了解决这个问题，Nissim 等人[19]认为，在一个给定的聚类中移动一个或者几个点不会显著改变质心的位置，所以可以使用局部敏感度来有效降低噪声. 为了计算一个复杂函数 f 的局部敏感度和相应的平滑边界，他们提出了一种抽样-聚合框架(Sample-aggregate Framework). 首先从原始数据集 D 中随机抽样产生 m 个子集，分别代入函数 f 得到 m 个中间结果. 然后选择一个局部敏感度较低并且容易计算的聚合函数 g，对 m 个中间结果进行聚合运算，得到一个关于 $f(D)$ 的期望值 $\bar{f}(D)$，最后根据函数 g 的平滑敏感度向 $\bar{f}(D)$ 中添加噪声，得到最终的发布数据.

但是基于抽样-聚合框架的算法在应用中存在一定的局限性，即对所有随机抽样生成的数据集进行聚类且输出的结果具有某种程度上的一致性时，算法才可能输出令人满意的结果. 因此，Feldman 等人[74]提出了一种基于核心集(Coreset)的方法，可用于基于差分隐私保护的 k-median 和 k-means 聚类分析. 给定一个 d 维空间里的点集 G，样本容量为 n，可以计算出一个带权重且容量为 $O(k\alpha^{-d}\log n)$ 的点集 $S_G \subseteq G$，使得用 S_G 替代 G 来做 k-median/means 聚类分析能够得到一个 $(1+\alpha)$-近似结果[75]，S_G 称为 G 的核心集. Feldman 等人给出了一个满足 ε-差分隐私保护要求的算法 A，能够以至少 $1-\delta$ 的概率输出核心集 $S_C = A(G)$. 给定对集合 G 的 k-median/means 聚类分析查询集合 F，对于任意查询 $f \in F$，可使

$$(1-\alpha)f(G)-\beta \leq f(S_C) \leq (1+\alpha)f(G)+\beta \quad (16)$$

成立. 其中 β 为预定义的边界值，S_C 称为 G 的 Private 核心集. 由于 S_C 是满足差分隐私保护要求的，因此基于 S_C 的聚类分析查询也是满足差分隐私保护的，且能够满足预定义的聚类分析准确性.

另外，Dwork[17]从预算分配的角度对差分隐私保护 k-means 算法进行了完善，针对两种不同的情形给出了分配预算的方法. 设数据集维度为 d，则查询函数的敏感度为 $d+1$. 若算法迭代次数指定为 u，则每次迭代的预算为 ε/u，添加的噪声服从 $Lap((d+1)u/\varepsilon)$ 分布；若算法迭代次数不定，则第一次分配的预算为 $\varepsilon/2$，之后每次迭代的预算为上一次迭代预算的 1/2.

4.2 完全访问模式下的数据挖掘

4.2.1 完全访问模式下的分类算法

Jagannathan 等人应用随机决策树算法[76]，提出了完全访问模式下的差分隐私保护随机决策树分类器构建方法[77]. 与传统决策树的构建方法不同，随机决策树首先通过随机选择分类属性来构建一个决策树架构，这个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，并将不符合预定义规则的叶节点剪枝. 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. 显然，删除数据集中的一个记录

会使决策树的某个叶节点发生改变，甚至在剪枝过程中使得一个子树被删除. 为了使随机决策树模型满足差分隐私保护的要求，首先去掉构建模型过程中的剪枝步骤，然后取出全部叶节点中所有可能的类标签的计数值，形成向量 $\boldsymbol{V}$，它由 $N\times T$ 个整数构成，其中 N 为叶节点数量，T 为类标签值的数量. 由于向量 $\boldsymbol{V}$ 的全局敏感度为 1，因此向 $\boldsymbol{V}$ 中加入 $Lap(1/\varepsilon)$噪声即可达到差分隐私保护的要求(如果分类器包含 k 棵树，则构建每棵数所分配的预算为 ε/k)，有效降低了噪声量. 该算法在 3 个实际数据集(UCI 中的 Nursery、Congressional Voting Records 和 Mushroom)上的测试表明，在不同预算条件下($\varepsilon=0.5,0.75,1$)，所构建的决策树平均分类准确率均超过了 85%.

另外 Chaudhuri 等人[78]提出了一种满足差分隐私保护的 logistic 回归算法. 首先，他们证明了如果直接利用 Laplace 机制，在输出的回归模型上加入噪声，其分类准确度会随着正则化参数的减小而降低. 所以他们提出了一个新的算法，先将噪声加在目标函数的参数中，然后通过标准的 logistic 回归算法进行参数估计来获取模型. 此模型相对于直接利用 Laplace 机制，具有更高的分类准确率. 由此可以发现，正则化参数和隐私保护之间存在着联系，当正则化参数增大时，回归函数的敏感度降低，所需噪声随之减少.

4.2.2　完全访问模式下的频繁项集挖掘

频繁项集挖掘是数据挖掘领域的一项重要技术，可用于关联规则挖掘、用户行为预测以及相关性分析. 给定数据集 D，其中的每个记录称为一个事务，每个事务由一些项(Item)构成，这些项来自于项空间 I. 项集由若干个项组成，是 I 的子集. 如果一个事务中包含了某项集 I_0，则称该事务支持项集 I_0. 支持项集 I_0 的事务数量占全部事务的比例称作项集 I_0 的支持度，支持度大于预定义阈值的项集称作频繁项集. 频繁项集挖掘算法的输出结果即为所有的频繁项集以及它们的支持度.

为了使频繁项集挖掘算法能够应用于包含隐私信息的数据集，Bhaskar 等人[79]对传统算法进行了改进，提出了两个基于“截断支持度”的频繁项集挖掘算法：基于指数机制的 FIM 算法和基于 Laplace 机制的 FIM 算法，用于根据给定数据集寻求长度为 l 的最频繁的 K 个项集.

首先，两个算法具有相同的预处理过程. 利用传统挖掘算法得到所有 l-项集，根据它们的真实支持度 s 计算出各自的“截断支持度” $\hat{s}=\max(s,s_K-\gamma)$. 其中 s_K 为第 K 大的支持度，γ 为用以控制结果可用性的预定义阈值. 算法中采用“截断支持度”的主要目的是为了避免遍历所有项集，降低算法的计算复杂度.

之后，基于指数机制的 FIM 算法以 $\hat{s}$ 为可用性函数 K 次调用指数机制，从预处理结果中不重复地选出 K 个项集；而基于 Laplace 机制的 FIM 算法则对预处理结果中的每个 $\hat{s}$ 值添加 Laplace 噪声，然后从所得结果中选择支持度最大的 K 个项集.

最后，两个算法采用相同的方法，对所选定的 K 个项集，为它们的真实支持度添加 Laplace 噪声后，将项集及干扰后的支持度一同输出.

对算法的隐私性和可用性分析表明，两种算法均能提供 ε-差分隐私保护，也能以 $1-\rho$ 的概率实现两个可用性要求(ρ 为预定义的阈值)，即所有真实支持度大于 $s_K+\gamma$ 的项集都能被输出，且所有被输出的项集的真实支持度都不小于 $s_K-\gamma$.

但这些算法也存在一些缺点. 首先，算法在执行前必须预定义输出项集的长度 l，而不能根据中间结果自适应的确定 l 的值；另外，频繁项集挖掘的难点之一在于数据集的高维度导致的计算复杂度，所以以上算法仅适用于 K 值较小的场合. 当 K 较大(例如 $K>100$)，尤其在项空间 $|I|$ 也较大时，算法的计算性能及输出的准确性均会显著下降.

针对这些问题，Li 等人[80]提出了用于高维数据集的频繁项集挖掘方法——PrivBasis，能够在保证计算性能的前提下实现差分隐私保护. PrivBasis 方法实际上是利用了频繁项集的一个重要性质来实现降维处理，即一个频繁项集的所有子集也都是频繁的. 为了找到最频繁的 K 个项集，PrivBasis 方法希望能够找到一个项集 I_B，使得最频繁的 K 个项集均是 I_B 的子集，然后通过计算并利用 I_B 中 1-项集和 2-项集的支持度，可以重构 I_B 中所有子集的支持度，并添加相应的噪声. 这种方法可以避免遍历所有可能的项集，降低计算代价. 但是当 K 很大时，I_B 也必将很大，导致遍历 I_B 的子集的计算代价也增大. 对此，PrivBasis 将 I_B 拆解为若干子集，提出了 θ-基集(θ-Basis Set)的概念. 一个 θ-基集表示为 $I_B=\{I_{B_1},I_{B_2},\cdots,I_{B_\omega}\}$，其中 $I_{B_i}(i=1,2,\cdots,\omega)$称为一个基(Basis). 任何一个支持度大于 θ 的项集(即一个候选的频繁项集)都是某个 I_{B_i} 的子集. 也就是说 I_B 的子集涵盖了所有支持度大于 θ 的项集. 显然，ω 的

值越小，且每个 I_{B_i} 的长度 $|I_{B_i}|$ 越小（$|I_{B_i}|$ 表示 I_{B_i} 中项的数量），遍历 I_B 所有子集的计算代价就越小. PrivBasis 方法中提出了根据频繁 1-项集和 2-项集来构建 θ-基集 I_B 的算法. 于是可以得到候选项集

$$C(I_B)=\bigcup_{i=1}^{\omega}\{I_X \mid I_X \subseteq I_{B_i}\} \tag{17}$$

其中 I_X 为 I_{B_i} 的任意子集. 最后为候选项集的支持度加入 Laplace 噪声，并选择支持度最大的 K 个项集对外发布. 通过对 5 个实际数据集的实验表明，PrivBasis 方法在频繁项集漏报率以及支持度误差两个方面都低于基于“截断支持度”的方法.

4.3 PPDM 小结

差分隐私在数据挖掘中的应用研究与差分隐私的理论发展密切相关. 本节对基于差分隐私保护的数据挖掘方法进行了归纳和分析，如表 7 所示.

PPDM 中各种算法的有效性不但依赖于算法本身，也受到所采用的差分隐私保护机制的精确度的影响. 所以 PPDR 中面临的高敏感度查询和计算复杂度等问题在 PPDM 中依旧存在. 除此之外，PPDM 需要解决的问题还包括：

(1) 对于许多数据挖掘算法，在接口模式下采用现有的差分隐私保护机制来实现 PPDM 往往需要较大的噪声. 为这些算法设计新的机制以降低噪声是目前需要解决的问题之一.

(2) 在完全访问模式下，为执行不同的数据挖掘任务，必须将各种传统算法改造为满足差分隐私保护要求的算法，如何在一个标准的框架系统内实现对各种算法的差分隐私化，是 PPDM 需要解决的另一个问题.

表 7 基于差分隐私保护的数据挖掘方法分类比较

实现模式	挖掘方法	典型算法	优点	缺点
接口模式	分类	DiffP-C4.5[31], DiffPID3[31], SuLQ-based ID3[70], PINQ-based ID3[31]	容易实现，分类准确率高	需要事先确定迭代次数，隐私保护预算分配困难
	聚类	SuLQ-based k-means[70], Coreset[74] Sample-aggregate Framework[19]	容易实现，方法有效	敏感度高且难以计算；需较大隐私保护预算才能保证精度
完全访问模式	分类/回归	DiffGen[57], Private-RDT [77], Learning guarantees[78], Objective Perturbation[81]	分类准确率高	计算代价高
	频繁项集挖掘	FIM[79], PrivBasis[80] TruncatedDB[82], Diff-FPM[83]	精度较高，挖掘速度快	频繁项集长度有限

5 其它应用

从应用领域来看，差分隐私保护方法还被普遍应用于许多其它场合，例如推荐系统、网络数据分析、搜索日志发布等.

(1) 差分隐私在推荐系统中的应用.

推荐系统帮助用户从大量数据中寻找可能需要的信息. 在许多电子商务网站中，推荐系统用于发现商品项目之间的关系，并向顾客推荐可能消费的项目. 由于推荐系统需要利用大量用户数据进行协同过滤(Collaborative Filtering)，所以数据的隐私保护问题很早就受到人们的关注. Mcsherry 等人[84]最先将差分隐私保护方法引入到推荐系统. 他们假定推荐系统是不可信的，攻击者可以通过分析推荐系统的历史数据来推测用户的隐私信息，因此必须对推荐系统的输入进行干扰. 在分析项目之间的关系时，他们先建立项目相似度协方差矩阵，并向矩阵中加入 Laplace 噪声实施干扰，然后再提交给推荐系统实施常规推荐算法，例如 K 最近邻算法或者因子分解方法.

Machanavajjhala 等人[85]在基于社交网络数据的推荐系统中使用了差分隐私保护方法. 社交网络模型通常用图来表示，图中的节点表示用户，边则表示用户之间的关系并被视为敏感信息. 为了使构建图的过程满足差分隐私保护要求，他们以节点的邻居数为可用性函数并采用指数机制来随机地构造图中的边，最终实现对图中所有边的保护.

Zhu 等人[86]针对 K 最近邻算法所面临的隐私泄露问题提出了一种基于差分隐私保护的邻居协同过滤算法. 该算法通过隐私邻居选择和定义推荐敏感度两个关键的隐私保护步骤，来确保从用户的推荐选择中无法推断出用户的历史记录. 由于在计算噪声的过程中采用了局部敏感度，使得最终的推荐结果保持了较好的可用性，是一种实用的隐私保护推荐算法. 另外，针对基于标签的推荐系统，Zhu 等人[87]提出了一种对用户轮廓(User Profile)进行修改并发布的差分隐私保护算法，能够在一定的精度损失范围内进行标签推荐并保护用户隐私.

(2) 差分隐私在网络踪迹分析中的应用.

网络踪迹分析是通过测量和分析网络流量来获取有用的信息. 网络数据和流量记录往往由一些企业或研究机构共享以供研究分析之用. 但由于这类分析有可能泄露隐私，所以这些网络数据在共享前需要经过净化. 早期的净化方法主要为匿名处理. 但 Mcsherry 等人[88]认为匿名化方法不足以保证网络数据的隐私性，所以将差分隐私的概念引入，并在 PINQ 平台上实现了网络数据统计分析的差分隐私保护方法. 其基本思想是发布网络数据的各项统计数据时，根据每项统计的敏感度，在结果中加入 Laplace 噪声，使网络数据中的单独个体对统计结果不会有影响. 相对于早期的匿名化方法，此方法较好地保证了网络数据的大部分统计特性.

(3) 差分隐私在运输信息保护中的应用.

Chen 等人[89]将差分隐私用于对运输信息的保护. 这里的运输信息是指公共交通系统中乘客的各种乘车及换乘信息，对这些信息的分析可以促进零售业和交通系统内的知识发现. 但由于其中包含了乘客的个人信息，所以在发布和共享之前，需要进行隐私保护处理. 分析运输信息的目的是寻找最频繁的乘车路线，因此本质上这是一个频繁序列挖掘问题. Chen 等人根据数据的特征，采用前缀树(Prefix Tree)来表示运输信息数据集. 树中每个节点表示一个序列以及数据集对该序列的支持计数. 由于这些支持计数中加入了 Laplace 噪声，从而保证了挖掘结果满足了差分隐私保护的要求.

(4) 差分隐私在搜索日志保护中的应用.

文献[90]提出了一种搜索日志(Search Log)发布算法，用于搜索引擎公司在差分隐私保护条件下对外发布高频关键词、查询和点击记录等信息.

已有的应用研究表明，差分隐私作为一种严格的隐私定义，能够为解决现实中的隐私保护问题提供有效的解决方法，具有良好的应用前景.

6　总结与展望

差分隐私是一种严格的和可证明的安全模型. 近年来的研究使得其在理论上不断发展和完善，并在统计学、机器学习、数据挖掘、社交网络等领域得到了初步应用.

本文介绍了差分隐私保护的基础理论，并着重对基于差分隐私保护的数据发布和数据挖掘方法进行了综述. 差分隐私保护数据发布关注的是在给定的隐私保护预算下，发布查询结果的精确性，以及保证此精确性所需要的样本复杂度，其实现方法除了加噪发布外，还有小波分析、域空间搜索、递归发布等等. 差分隐私保护数据挖掘关注的则是在实现隐私保护的前提下，所得挖掘模型的分类/预测准确性，其实现方法主要是将噪声机制或指数机制嵌入到数据挖掘算法中，使得数据挖掘的过程满足差分隐私保护的要求.

当然，差分隐私保护还是一个相对年轻的研究领域，在理论和应用上都还存在一些难点以及新的方向需要进一步深入研究，包括：

(1) 复杂数据的差分隐私保护.

在实际应用中存在许多复杂的数据集，其中的记录之间往往存在某种联系. 然而目前的差分隐私保护方法并未考虑数据之间的联系，因此无法有效地处理这类数据集[91]. 例如在社交网络数据中，每个用户都会和许多其他用户产生联系，因此即便从数据集中删除了某个用户，仍可能从与其他用户的联系中推断出该用户的信息. 在这种情况下，如果采用传统的差分隐私保护方法，同时考虑数据之间的联系，则查询敏感度会很高，从而引入过多噪声.

(2) 连续数据发布的隐私保护.

已有差分隐私研究大多针对静态数据发布问题，但在实际应用中，很多数据集都是动态更新的. 例如在线零售数据、推荐系统信息等.

连续数据发布的差分隐私保护问题主要有两个研究难点：其一是隐私保护预算的分配问题. 现有的机制需要预先定义发布的次数，然后分配隐私保护预算. 当数据持续更新超出这个次数时，预算被耗尽，发布机制失效. 第二个难点是噪声大. 由于每次更新后的数据发布必须包含之前发布时的噪声，因此随着发布次数的增长，累积噪声会迅速增大，导致发布结果的可用性极低. 对此，Chan 等人[92]提出了 p-sum 方法. 该方法实质上是对实时数据进行重新划分，只有在实时数据累积增加到阈值 p 的时候，才加上噪声重新发布新的结果. 但该方法并没有解决隐私保护预算耗尽后的机制失效问题以及累积噪声随发布次数迅速增大的问题.

连续数据发布中的隐私保护问题还需要进一步的深入研究. 我们认为，采用动态的局部敏感度是一种可行的发布方式，即数据更新后的查询敏感度可根据更新前的发布结果重新计算，这样可以通过降低查询敏感度来降低噪声.

(3)差分隐私保护框架系统.

差分隐私保护框架系统是实现隐私保护的基础设施.在其基础上,研究人员可以自行设计更加复杂的差分隐私保护算法,例如在 PINQ 或 SuLQ 的基础上设计满足差分隐私的数据挖掘算法.但此类框架系统的研究难点在于如何实现查询敏感度的自动计算和发布机制的自动优化.查询敏感度的自动计算是一个十分复杂的问题,可能的解决方案是为一类查询提供一个噪声上界,但此方法可能会增加不必要的噪声.发布机制的自动优化也同样具有挑战性,对于复杂的查询,直接使用 Laplace 机制和指数机制效果并不理想,如何根据发布任务让差分隐私保护框架自动选择优化机制还需要更深入的研究.从目前的研究来看,复杂的差分隐私保护算法还只能由领域专家提出并证明其差分隐私保护的性能.

(4)分布式差分隐私计算.

分布式隐私保护是隐私保护领域的一个重要分支,它研究互不信任的多个实体如何对信息进行共享而不泄露自己的隐私信息[93].在具体实现中,各实体将自己的数据集输入一个安全函数,并共享函数输出结果.该方向的研究难点在于两点:(1)如何选择安全函数,使之满足差分隐私的要求;(2)如何设计协议以兼顾差分隐私性和计算复杂度[94].对这两个问题,目前的研究还只是从理论上提出了可行性以及相应计算的误差界限,具体的实现方法还需要进一步的研究.

(5)差分隐私定义的延伸.

差分隐私是一种严格的定义,它假定攻击者具有尽可能多的背景知识.因此,为了满足差分隐私保护的要求,必须在发布结果中引入足够大的噪声,但噪声过大可能导致数据完全失去意义.针对这个问题,一些研究者试图通过降低差分隐私的要求,在适当降低隐私性的情况下,提高结果的可用性.例如文献[56]提出的"安全 k-匿名"模型以及文献[95]提出的 Crowd-Blending 隐私定义等,但这些定义在实现上都比较困难.尽管如此,我们认为在差分隐私的基础上提出新的隐私定义是对差分隐私定义的完善与延伸,对于扩展差分隐私的应用领域具有重要的意义.

总之,差分隐私保护是目前信息安全领域的研究热点之一,也取得了丰富的研究成果.但从实际应用的角度来看,还有许多内容需要继续深入研究.本文从理论和应用的角度对差分隐私保护目前的研究状况进行综述,希望能够为该领域的研究者提供有价值的参考信息.

参 考 文 献

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

[2] Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets//Proceedings of the 2008 IEEE Symposium on Security and Privacy. Oakland, USA, 2008: 111-125

[3] Srivatsa M, Hicks M. Deanonymizing mobility traces: Using social network as a side-channel//Proceedings of the 2012 ACM Conference on Computer and Communications Security. Raleigh, USA, 2012: 628-637

[4] Blond S L, Zhang C, Legout A, et al. I know where you are and what you are sharing: Exploiting P2P communications to invade users' privacy//Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference. Berlin, Germany, 2011: 45-60

[5] Dalenius T. Towards a methodology for statistical disclosure control. Statistik Tidskrift, 1977, 15(2): 429-444

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

[7] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007, 1(1): 3

[8] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity//Proceedings of the IEEE International Conference on Data Engineering (ICDE). Istanbul, Turkey, 2007: 106-115

[9] Wong R C-W, Li J, Fu A W-C, Wang K. (α,k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 754-759

[10] Xiao X, Tao Y. M-invariance: Towards privacy preserving re-publication of dynamic datasets//Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. Beijing, China, 2007: 689-700

[11] Wong R C-W, Fu A W-C, Wang K, Pei J. Minimality attack in privacy preserving data publishing//Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, Austria, 2007: 543-554

[12] Ganta S R, Kasiviswanathan S P, Smith A. Composition attacks and auxiliary information in data privacy//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 265-273

[13] Wong R C-W, Fu A W-C, Wang K, et al. Can the utility of anonymized data be used for privacy breaches? ACM Transactions on Knowledge Discovery from Data (TKDD), 2011, 5(3): 1-24

[14] Kifer D. Attacks on privacy and deFinetti's theorem// Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. Providence, Rhode Island, USA, 2009: 127-138

[15] Dwork C. Differential privacy//Proceedings of the 33rd International Colloquium on Automata, Languages and Programming. Venice, Italy, 2006: 1-12

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

[17] Dwork C. A firm foundation for private data analysis. Communications of the ACM, 2011, 54(1): 86-95

[18] Haeberlen A, Pierce B C, Narayan A. Differential privacy under fire//Proceedings of the 20th USENIX Conference on Security. San Francisco, USA, 2011: 33-33

[19] Nissim K, Raskhodnikova S, Smith A. Smooth sensitivity and sampling in private data analysis//Proceedings of the 39th Annual ACM Symposium on Theory of Computing. San Diego, USA, 2007: 75-84

[20] McSherry F. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. Communications of the ACM, 2010, 53(9): 89-97

[21] Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis//Proceedings of the 3rd Conference on Theory of Cryptography. New York, USA, 2006: 265-284

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

[23] Dwork C, Kenthapadi K, McSherry F, et al. Our data, ourselves: Privacy via distributed noise generation//Proceedings of the 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques. St. Petersburg, Russia, 2006: 486-503

[24] Dwork C, McSherry F, Talwar K. The price of privacy and the limits of LP decoding//Proceedings of the 39th Annual ACM Symposium on Theory of Computing. San Diego, USA, 2007: 85-94

[25] Dwork C, Naor M, Pitassi T, Rothblum G N. Differential privacy under continual observation//Proceedings of the 42nd ACM Symposium on Theory of Computing. Cambridge, USA, 2010: 715-724

[26] Dwork C, Naor M. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. Journal of Privacy and Confidentiality, 2008, 2(1): 8

[27] Dwork C. The differential privacy frontier (extended abstract)//Proceedings of the 6th Theory of Cryptography Conference on Theory of Cryptography. San Francisco, 2009: 496-502

[28] Dwork C, Naor M, Reingold O, et al. On the complexity of differentially private data release: Efficient algorithms and hardness results//Proceedings of the 41st Annual ACM Symposium on Theory of Computing. Bethesda, USA, 2009: 381-390

[29] Dwork C, Lei J. Differential privacy and robust statistics// Proceedings of the 41st Annual ACM Symposium on Theory of Computing. Bethesda, USA, 2009: 371-380

[30] Dwork C. Differential privacy in new settings//Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms. Austin, Texas, 2010: 174-183

[31] Friedman A, Schuster A. Data mining with differential privacy//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 493-502

[32] Roth A. New algorithms for preserving differential privacy [Ph.D. dissertation]. Carnegie Mellon University, Pittsburgh, USA, 2010

[33] Roth A, Roughgarden T. Interactive privacy via the median mechanism//Proceedings of the 42nd ACM Symposium on Theory of Computing. Cambridge, USA, 2010: 765-774

[34] Hardt M, Rothblum G N. A multiplicative weights mechanism for privacy-preserving data analysis//Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science. 2010: 61-70

[35] Hardt M, Talwar K. On the geometry of differential privacy// Proceedings of the 42nd ACM Symposium on Theory of Computing. Cambridge, USA, 2010: 705-714

[36] Gupta A, Roth A, Ullman J. Iterative constructions and private data release//Proceedings of the 9th International Conference on Theory of Cryptography. Sicily, Italy, 2012: 339-356

[37] Gupta A, Hardt M, Roth A, Ullman J. Privately releasing conjunctions and the statistical query barrier//Proceedings of the 43rd Annual ACM Symposium on Theory of Computing. San Jose, California, USA, 2011: 803-812

[38] Muthukrishnan S, Nikolov A. Optimal private halfspace counting via discrepancy//Proceedings of the 44th Symposium on Theory of Computing. New York, USA, 2012: 1285-1292

[39] Xiao Y, Xiong L, Yuan C. Differentially private data release through multidimensional partitioning//Proceedings of the 7th VLDB Conference on Secure Data Management. Singapore, 2010: 150-168

[40] Xiao Y, Gardner J, Xiong L. DPCube: Releasing differentially private data cubes for health information//Proceedings of the 2012 IEEE 28th International Conference on Data Engineering. Washington, USA, 2012: 1305-1308

[41] Xu J, Zhang Z, Xiao X, et al. Differentially private histogram publication//Proceedings of the 2012 IEEE 28th International Conference on Data Engineering. Washington, USA, 2012: 32-43

[42] Hay M, Rastogi V, Miklau G, Suciu D. Boosting the accuracy of differentially private histograms through consistency. Proceedings of the VLDB Endowment, 2010, 3(1-2): 1021-1032

[43] Dinur I, Nissim K. Revealing information while preserving privacy//Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. San Diego, USA, 2003: 202-210

[44] Xiao X, Wang G, Gehrke J. Differential privacy via wavelet transforms. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(8): 1200-1214

[45] Li C, Hay M, Rastogi V, et al. Optimizing linear counting queries under differential privacy//Proceedings of the 29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Indianapolis, USA, 2010: 123-134

[46] Cormode G, Srivastava D, Shen E, Yu T. Aggregate query answering on possibilistic data with cardinality constraints//Proceedings of the 2012 IEEE 28th International Conference on Data Engineering. Washington, USA, 2012: 258-269

[47] Barak B, Chaudhuri K, Dwork C, et al. Privacy, accuracy, and consistency too: A holistic solution to contingency table release//Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Beijing, China, 2007: 273-282

[48] Dobra A, Fienberg S. Bounding entries in multi-way contingency tables given a set of marginal totals//Haitovsky Y, Ritov Y, Lerche H eds. Foundations of Statistical Inference. Heidelberg, German: Physica-Verlag HD, 2003: 3-16

[49] De Loera J, Onn S. All rational polytopes are transportation polytopes and all polytopal integer sets are contingency tables//Proceedings of 10th International IPCO Conference. New York, USA, 2004: 338-351

[50] De Loera J, Onn S. The complexity of three-way statistical tables. SIAM Journal on Computing, 2004, 33(4): 819-836

[51] De Loera J, Onn S. Markov bases of three-way tables are arbitrarily complicated. Journal of Symbolic Computation, 2006, 41(2): 173-181

[52] Kasiviswanathan S P, Rudelson M, Smith A, Ullman J. The price of privately releasing contingency tables and the spectra of random matrices with correlated rows//Proceedings of the 42nd ACM Symposium on Theory of Computing. Cambridge, USA, 2010: 775-784

[53] Fienberg S E, Rinaldo A, Yang X. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables//Proceedings of the 2010 International Conference on Privacy in Statistical Databases. Corfu, Greece, 2010: 187-199

[54] Yang X, Fienberg S E, Rinaldo R. Differential privacy for protecting multi-dimensional contingency table data: Extensions and applications. Journal of Privacy and Confidentiality, 2012, 4(1): 101-125

[55] Machanavajjhala A, Gehrke J, Götz M. Data publishing against realistic adversaries. Proceedings of the VLDB Endowment, 2009, 2(1): 790-801

[56] Li N, Qardaji W, Su D. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy//Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security. Seoul, Korea, 2012: 32-33

[57] Mohammed N, Chen R, Fung B C M, Yu P S. Differentially private data release for data mining//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 493-501

[58] Zhu T, Xiong P, Xiang Y, Zhou W. An effective differentially private data releasing algorithm for decision tree//Proceedings of the 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications. Melbourne, Australia, 2013: 388-395

[59] Anthony M, Bartlett P L. Neural Network Learning: Theoretical Foundations. Cambridge, UK: Cambridge University Press, 2009

[60] Kasiviswanathan S P, Lee H K, Nissim K, et al. What can we learn privately? SIAM Journal on Computing, 2011, 40(3): 793-826

[61] Blum A, Ligett K, Roth A. A learning theory approach to non-interactive database privacy//Proceedings of the 40th Annual ACM Symposium on Theory of Computing. Victoria, Canada, 2008: 609-618

[62] Valiant L G. A theory of the learnable//Proceedings of the 16th Annual ACM Symposium on Theory of Computing. Washington, USA, 1984: 436-445

[63] Kearns M J, Schapire R E, Sellie L M. Toward efficient agnostic learning. Machine Learning, 1994, 17(2-3): 115-141

[64] Hardt M, Rothblum G N, Servedio R A. Private data release via learning thresholds//Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms. Kyoto, Japan, 2012: 168-187

[65] Dwork C, Rothblum G N, Vadhan S. Boosting and differential privacy//Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science. Las Vegas, USA, 2010: 51-60

[66] Acs G, Castelluccia C, Chen R. Differentially private histogram publishing through lossy compression//Proceedings of the 2012 IEEE 12th International Conference on Data Mining. Brussels, Belgium, 2012: 1-10

[67] Yuan G, Zhang Z, Winslett M, et al. Low-rank mechanism: Optimizing batch queries under differential privacy. Proceedings of the VLDB Endowment, 2012, 5(11): 1352-1363

[68] Li C, Miklau G. An adaptive mechanism for accurate query answering under differential privacy. Proceedings of the VLDB Endowment, 2012, 5(6): 514-525

[69] Yaroslavtsev G, Procopiuc C M, Cormode G, Srivastava D. Accurate and efficient private release of datacubes and contingency tables//Proceedings of the 2013 IEEE International Conference on Data Engineering (ICDE 2013). Brisbane, Australia, 2013: 745-756

[70] Blum A, Dwork C, McSherry F, Nissim K. Practical privacy: The SuLQ framework//Proceedings of the 24th ACM SIGMOD- SIGACT-SIGART Symposium on Principles of Database Systems. Baltimore, USA, 2005: 128-138

[71] Dwork C, Nissim K. Privacy-preserving datamining on vertically partitioned databases//Proceedings of the 24th Annual International Cryptology Conference. Santa Barbara, USA, 2004: 528-544

[72] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011

[73] Quinlan J R. Induction of decision trees//Bruce G B, David C W. Readings in knowledge acquisition and learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993: 349-361

[74] Feldman D, Fiat A, Kaplan H, Nissim K. Private coresets//Proceedings of the 41st Annual ACM Symposium on Theory of Computing. Bethesda, USA, 2009: 361-370

[75] Har-Peled S, Mazumdar S. On coresets for k-means and k-median clustering//Proceedings of the 36th Annual ACM Symposium on Theory of Computing. Chicago, USA, 2004: 291-300

[76] Fan W, Wang H, Yu P S, Ma S. Is random model better? On its accuracy and efficiency//Proceedings of the 3rd IEEE International Conference on Data Mining. Melbourne, USA, 2003: 51

[77] Jagannathan G, Pillaipakkamnatt K, Wright R N. A practical differentially private random decision tree classifier. Transactions on Data Privacy, 2012, 5(1): 273-295

[78] Chaudhuri K, Monteleoni C. Privacy-preserving logistic regression//Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS'2008). Vancouver, Canada, 2008: 289-296

[79] Bhaskar R, Laxman S, Smith A, Thakurta A. Discovering frequent patterns in sensitive data//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 503-512

[80] Li N, Qardaji W, Su D, Cao J. PrivBasis: Frequent itemset mining with differential privacy. Proceedings of the VLDB Endowment, 2012, 5(11): 1340-1351

[81] Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization. The Journal of Machine Learning Research, 2011, 12: 1069-1109

[82] Zeng C, Naughton J F, Cai J-Y. On differentially private frequent itemset mining. Proceedings of the VLDB Endowment, 2012, 6(1): 25-36

[83] Shen E, Yu T. Mining frequent graph patterns with differential privacy//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 545-553

[84] McSherry F, Mironov I. Differentially private recommender systems: building privacy into the net//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 627-636

[85] Machanavajjhala A, Korolova A, Sarma A D. Personalized social recommendations: accurate or private. Proceedings of the VLDB Endowment, 2011, 4(7): 440-450

[86] Zhu T, Li G, Ren Y, et al. Differential privacy for neighborhood-based collaborative filtering//Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Niagara Falls, Canada, 2013: 752-759

[87] Zhu T, Li G, Ren Y, et al. Privacy preserving for tagging recommender systems//Proceedings of the 2013 IEEE/WIC/ACM International Conference on Web Intelligence. Atlanta, USA, 2013(to be appeared)

[88] McSherry F, Mahajan R. Differentially-private network trace analysis//Proceedings of the ACM SIGCOMM 2010 Conference. New Delhi, India, 2010: 123-134

[89] Chen R, Fung B C M, Desai B C, Sossou N M. Differentially private transit data publication: A case study on the montreal transportation system//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 213-221

[90] Gotz M, Machanavajjhala A, Wang G, et al. Publishing search Logs——A comparative study of privacy guarantees. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(3): 520-532

[91] Kifer D, Machanavajjhala A. No free lunch in data privacy//Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. Athens, Greece, 2011: 193-204

[92] Chan T-H H, Shi E, Song D. Private and continual release of statistics//Proceedings of the 37th International Colloquium Conference on Automata, Languages and Programming. Bordeaux, France, 2010: 405-417

[93] McGregor A, Mironov I, Pitassi T, et al. The limits of two-party differential privacy//Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science. Las Vegas, USA, 2010: 81-90

[94] Beimel A, Nissim K, Omri E. Distributed private data analysis: Simultaneously solving how and what//Proceedings of the 28th Annual Conference on Cryptology: Advances in Cryptology. Santa Barbara, USA, 2008: 451-468

[95] Gehrke J, Hay M, Lui E, Pass R. Crowd-blending privacy//Safavi-Naini R, Canetti R eds. Advances in Cryptology-CRYPTO 2012. Heidelberg, German: Springer, 2012: 479-496

XIONG Ping, born in 1974, Ph. D., associate professor. His research interests include information security, machine learning and data mining.

ZHU Tian-Qing, born in 1979, Ph. D. candidate, lecturer. Her research interests include privacy preserving and network security.

WANG Xiao-Feng, born in 1978, Ph. D., assistant researcher. His research interests include artificial intelligence, data mining and wireless sensor networks.

Background

Differential privacy has become an important research area since the appearance of the first paper about this topic in 2006. There has been much work done in both computing theory and statistical fields on developing this new privacy notion to existing privacy preserving problems in the recent years. The interest in this area becomes very high because it constitutes a rigorous and provable privacy notion that can be implemented in various of research areas, not only in computing theory, but also in machine learning and data mining. In 2012, Microsoft releases a whitepaper titled Differential Privacy for Everyone, striving to translate this research into new privacy-enhancing technologies.

This paper provides a review of the main theoretical developments and applications of differential privacy. The authors summarize the principles on differential privacy and the mechanisms used to achieve it. They categorize the research on differential privacy into two main subjects: privacy preserving data release and privacy preserving data mining. The authors survey the related methods and algorithms in each of the subjects followed by discussion and comparison on these techniques, and point out the possible research trends in the future.

The work is done when the first author Xiong Ping is visiting Deakin University, Australia. It is supported by TULIP Lab at Deakin University. The work is also partially supported by the National Natural Science Foundation of China under Grant Nos. 61202211 and 61304067, Young Foundation of Ministry of Education, Humanities and Social Science Research Project (Project No. 12YJC630078), the Fundamental Research Funds for the Central Universities under Grant Nos. 31541311302 and 31541111305.