文章编号：2096-1472(2019)-10-44-03
DOI:10.19644/j.cnki.issn2096-1472.2019.10.010

隐私保护频繁项集挖掘中的细粒度随机化模型

郭宇红[1]，童云海[2]

(1.国际关系学院信息科技学院，北京 100091；
2.北京大学智能科学系，北京 100871)

摘　要：已有的随机化回答模型调控的数据范围宽、粒度粗，对隐私数据的保护粒度缺乏灵活性，无法实现精细化、个性化、差异化的隐私保护。提出三类多参数随机化回答模型，包括行多参、复合多参、分组多参共11种随机化回答模型，给出了模型的分类框架和分类层次。细粒度多参数随机化模型可实现精细化、个性化、差异化的隐私保护效果。

关键词：随机化回答；隐私保护；频繁项集；敏感问题调查
中图分类号：TP311　**文献标识码**：A

Fine-grained Randomized Response Model in Privacy Preserving Frequent Itemset Mining

GUO Yuhong[1],TONG Yunhai[2]

(1.*School of Information Science and Technology*,*University of International Relations*,*Beijing* 100091,*China*;
2.*Department of Intelligence Science*,*Peking University*,*Beijing* 100871,*China*)

Abstract: The existing randomized response model regulates a wide range of data with coarse granularity and lacks flexibility in protecting privacy,unable to achieve fine,personalized and differentiated privacy protection.Three kinds,11 types of multi-parameter random response models are proposed,including row multi-parameter,compound multi-parameter and grouping multi-parameter.The classification framework and hierarchy of these models are given.The fine-grained multi-parameter randomized response models can realize fine,individualized and differentiated privacy protection effect.

Keywords:randomized response;privacy preserving;frequent item set;sensitivity survey

1 引言(Introduction)

频繁项集挖掘是数据挖掘中的一个重要分支，随着人们对数据隐私和安全的日益关注，频繁项集挖掘赖以生存的数据环境发生很大变化，表现在：(1)在数据收集阶段，出于隐私的考虑，人们可能不再愿意提供真实的数据供分析使用，比如在线调查，一家药品公司为了发现各种疾病之间的关联关系，需要开展疾病的调查以收集数据，而调查者出于隐私的考虑不愿意提供数据或提供虚假数据；(2)在企业试图将频繁模式挖掘任务外包给第三方，或多个企业合作进行频繁模式挖掘而需共享数据时，出于客户隐私和商业安全的考虑，在数据共享前需对数据进行预处理，以保护客户隐私和隐藏一些敏感规则；(3)频繁项集挖掘的科研人员对自己的算法进行测试时，难以得到真实数据作为benchmark。如何在基于隐私和安全考虑的环境中，很好地实施数据挖掘任务和各种应用，是隐私保护数据挖掘要解决的问题[1-3]。

2 相关工作(Related work)

随机化[4,5]是目前隐私保护数据挖掘中运用的主要方法，包括随机化干扰(randomized perturbation)和随机化回答(randomized response)两种模型。其中随机化干扰模型主要用于数值数据，通过在原始数值数据上增加随机干扰数实现；随机化回答模型主要用于分类数据，通过对分类属性值在不同取值间作随机变换实现，该模型最先由沃纳提出[6]，被广泛用于敏感性问题[7]的调查中。

在隐私保护频繁模式挖掘[8-11]、隐私保护关联规则挖掘[12,13]方面，文献[13]提出基于随机化回答的mask/MASK(Mining Associations with Secrecy Konstraints)方法，通过数据干扰和支持度重构实现了隐私数据保护的关联规则挖掘，mask方法中随机化参数只有一个，所有的数据元素受控于此唯一的参数。文献[14]对mask算法进行了扩展，提出“特定于符号(1和0)”的随机化过程(symbol-specific

基金项目：论文由国际关系学院中央高校基本科研业务费项目资助(项目编号3262017T48,3262018T02,3262019T06).

distortion)和相应的emask算法，emask对“1”“0”设置两个不同的随机化参数，使它们拥有不同的隐私保护级别。文献[15]提出“非统一”(non−uniform)参数的随机化过程和相应的项集支持度递归估计RE(Recursive Estimation)算法，RE在随机化过程中对不同属性设置不同的随机化参数，使不同属性可以拥有不同的隐私保护级别。文献[16]对mask算法在支持度重构复杂度方面进行了优化，提出了mmask算法。文献[17]使用“风险−效用”映射(Risk−Utility，R−U)比较了不同的随机化策略，并提出了用于布尔数据和分类数据的最优随机化策略，同emask一样，针对布尔数据的随机化采用两个随机化参数。

上述随机化回答模型在隐私保护频繁项集挖掘中，取得了很大进展，但存在以下问题。(1)随机化模型类型单一，随机化参数调控的数据范围宽、粒度粗，对隐私数据保护粒度的控制缺乏灵活性。(2)已有模型没有考虑不同个体隐私保护需求的差异性，而这种需求在现实应用中是客观存在和急需解决的。

针对以上问题，本文在沃纳模型、单参数等随机化模型的基础上，提出三类细粒度的随机化回答模型：个体多参随机化模型、复合多参随机化模型、分组多参随机化模型。个体多参随机化模型，面向不同个体需要不同保护的多样化隐私保护需求问题，可为不同个体设置不同的随机化参数；复合多参随机化模型，组合已有的单一型随机化模型，使随机化参数控制的数据范围更加细致；分组多参随机化模型，可对数据按不同方式分组，使随机化参数对于数据粒度的控制更加灵活。本文给出了随机化模型的分类框架和分类层次。本文旨在为随机化回答模型在敏感问题调查、隐私保护频繁项集挖掘的应用提供一个概览，推动隐私保护频繁项集挖掘方法的进一步研究。

3 多参数随机化模型 (Multi-parameter random response models)

不同于单参数随机化，多参数随机化用多个概率参数对数据进行随机化。其思想是对数据中的不同元素设置不同的隐私保护级别，不同的隐私保护级别对应不同的随机化参数，由参与调查的个体自行决定对其不同数据元素的隐私保护级别和相应的随机化参数。根据不同属性取值、不同属性和不同个体可以有以下几种多参数随机化模型。为简单起见，假设参与调查的个体总数为N，每个个体对应m个需要保护的属性，属性取值均为布尔值“1”“0”，由这N个个体的m个布尔属性组成需要保护的、二维布尔矩阵表示的数据表D。(事实上，数值类型属性可以通过离散化转变为多元分类属性，即枚举属性，而多元分类属性又可以转变为布尔属性，即一般的数据都可以转变为二维布尔矩阵形式)

3.1 二元随机化模型(简称P2模型)

这是最简单的多参数随机化模型，在该模型中，设置有两个不同的随机化参数p_1和p_2，对值“1”和“0”区别对待，p_1和p_2分别决定了对“1”和“0”的两个不同的隐私保护级别。随机化过程为：对于D中取值为“1”的元素，以p_1的概率保持不变，以$1-p_1$的概率取反；对于D中取值为“0”的元素以p_2的概率保持不变，$1-p_2$的概率取反。

文献[14]对[13]提出的单参数随机化模型和对应的mask算法进行了扩展，提出了“特定于符号(1和0)”的随机化过程和相应的emask算法，其中的“特定于符号”的随机化过程对应于上述两参数随机化模型。通过对“1”“0”设置两个不同的随机化参数，并对p_1和p_2进行认真选择和调节，emask可以在获得满意的隐私保护性、挖掘结果准确性的同时，提高频繁项集挖掘的时间性能。

3.2 列多参随机化模型(简称Pm模型)

列(属性)多参随机化模型，对不同属性(列)设置不同的隐私保护级别，假定有m个属性$A_1\cdots A_m$，则设置$p_1\cdots p_m$共m个随机化参数。随机化过程为：对于属性A_j所在列中的所有数据元素，以p_j的概率取原值，以$1-p_j$的概率取反。

列多参模型主要基于以下事实：通常人们对数据中的不同属性有不同的隐私关注度。反映在敏感问题问卷调查中，不同的问题其敏感度是有差别的，例如，对于“你是否喝过酒”“你是否抽过烟”和“你是否吸过毒”三个问题，其敏感度和需要的隐私保护级别是不一样的。再比如“性别”“年龄”和“收入”三个属性的敏感度和需要的保护级别也是不同的。文献[15]针对该类场景，即“不同属性需要不同保护”的关联规则挖掘进行了研究，提出了“非统一”参数的随机化过程和对应的项集支持度递归估计RE算法。其中“非统一”参数的随机化对应于上述列多参随机化模型。RE方法通过允许不同属性拥有不同的隐私级别，即对不同属性设置不同的随机化参数，可以提高支持度估计的精确性。

3.3 行多参随机化模型(简称P_N模型)

行(个体)多参随机化模型，对不同个体设置不同的隐私保护级别。假定有N个个体$I_1\cdots I_N$，则设置$p_1\cdots p_N$共N个随机化参数。随机化过程为：对于个体I_j所在行中的所有数据元素，以p_j的概率取原值，以$1-p_j$的概率取反。

列多参模型虽然解决了纵向“不同属性需要不同保护”的隐私保护需求问题，但没有解决“不同个体需要不同保护”的隐私保护需求问题。

事实上，通常不同的人对数据(即便是同样的数据)会有不同的隐私关注度，这来自个人对事物的主观态度和意愿。反映在调查问卷中，不同的人对问题(即便是对同一个问题)的敏感度是有差异的，例如，对于同样的调查问题“你是否吸过毒”，吸毒者张三、李四和王五的反应态度可能会截然不同。张三对此问题并不敏感，表示十分愿意真实地做出回答和贡献其个人对此问题的数据；李四对此问题的敏感度模棱两可，表示愿意以60%的概率贡献其真实数据；王五对此问题非常敏感，完全不愿意泄露对此问题的个人信息。同样，对于“年龄”属性，女性比男性可能更敏感，造成问卷调查中，女性组对于男性组更难于接受此项调查。这些例子均表明不同的个体会有不同的隐私保护要求和偏好。

然而，到目前为止，还没有文献对“不同个体需要不同保护”的隐私保护频繁模式挖掘(或关联规则挖掘)问题做研究。现有的隐私保护关联规则挖掘解决方案都是独立于个体的，即默认所有个体的隐私保护要求都相同，没有尊重不同个体的主观意愿。在提倡“以客户为中心”“以人为本”“尊重用户偏好”“强调个性化服务”的人本时代，如何能在获取相关信息的同时，尽可能地为不同的个体提供定制的隐私保护服务，满足不同个体的多样化隐私保护要求，是隐私保护的数据发布、数据分析和数据挖掘必须解决和亟待解决的问题。本章将结合上面所提出的行多参随机化模型，设计与之相对应的项集支持度估算方法，通过允许不同个体拥有不同的隐私级别，实现频繁模式挖掘中的个性化隐

私保护。

3.4 复合多参随机化模型

上述三种基本的多参随机化模型P_2、P_m、P_N可相互交叉组合，产生$P_{2\times m}$、$P_{2\times N}$、$P_{m\times N}$和$P_{2\times m\times N}$共四种复合的多参数随机化模型(脚标表示随机化参数个数)，不同模型中随机化参数所控制的范围或粒度不同。

三种基本多参随机化模型中，二元随机化P_2模型中的一个随机化参数控制一类取值，只对“1”和“0”区分，保护粒度最粗；P_m模型中的一个随机化参数控制一列，对每列作区分；P_N模型中的一个参数控制一行，对每行作区分。

四种复合多参随机化模型中，$P_{2\times m}$模型的每列有两个随机化参数(称其为二元随机化列)，一个控制该列中的“1”，一个控制该列中的“0”，对“1”“0”和列同时作区分；$P_{2\times N}$模型中的每行有两个随机化参数(称其为二元随机化行)，分别控制该行中的“1”和“0”，对“1”“0”和行同时作区分；而$P_{m\times N}$模型有$m\times N$个随机化参数，每个参数控制一个数据单元，对行和列同时区分；$P_{2\times m\times N}$模型有$2\times m\times N$个随机化参数，每一个数据单元上有两个随机化参数(称其为二元随机化数据单元)，该模型对取值、行、列同时作区分，保护粒度最细。

3.5 分组多参随机化模型

在细粒度的隐私保护随机化模型中，若保护粒度太细，随机化参数太多，并且存在多个属性的敏感度相当，或参与调查的多个个体的隐私保护要求差不多，则可按属性垂直分组或按个体水平分组，进行分组随机化，使同一组内共用一个随机化参数，形成分组多参随机化模型。将按个体、属性分组结合到以上P_m、PN和$P_{2\times m}$、$P_{2\times N}$、$P_{m\times N}$模型中(注：P_2模型不涉及个体、属性，不可分组)，可派生出下面六种分组多参随机化模型：

$P_{m/g}$模型：属性分组多参随机化模型，m个属性被分为若干组，每组一个随机化参数，每个随机化参数控制组中的多列。若等分时每组包含g列，则组数和随机化参数个数为m/g。

$P_{N/g}$模型：个体分组多参随机化模型，N个个体被分为若干组，每组一个随机化参数，每个随机化参数控制组中的多行。若等分时每组包含g行，则组数和随机化参数个数为N/g。

$P_{2\times m/g}$模型：二元属性分组多参随机化模型，m个二元随机化的列垂直分为m/g组，每组有两个随机化参数，一个控制该组中的“1”，一个控制该组中的“0”。

$P_{2\times N/g}$模型：二元个体分组多参随机化模型，N个二元随机化的行水平分为N/g组，每组有两个随机化参数，一个控制该组中的“1”，一个控制该组中的“0”。

$P_{m/g\times N/g}$模型：行、列交叉分组多参随机化模型，简称分块多参随机化模型。二维矩阵在行列方向同时分组后，被分为$m/g\times N/g$块，每块一个随机化参数。

$P_{2\times m/g\times N/g}$模型：二元行、列交叉分组多参随机化模型，简称二元分块多参随机化模型。二维矩阵在行列方向同时分组后，被分为$m/g\times N/g$块，每块两个随机化参数，一个控制该块中的“1”，一个控制该块中的“0”。

图1给出了随机化模型的分类框架、分类层次。该分类框架主要依据随机化过程是否根据取值、属性和个体的不同作了区分。依此分类框架构成的随机化模型分类层次中，最上端的P模型是最简单的单参数随机化模型，该模型不区分1—0取值、不区分行和列，以统一的随机化参数对一个二维的数据表进行随机化干扰。位于第二层的是三种基本的多参随机化模型P_2、P_m、P_N模型，分别对1—0取值、列和行作区分，并继承了P模型的基本随机化过程：以p的概率取原值，以$1-p$的概率取反。位于第三、第四层的是四种复合的多参数随机化模型$P_{2\times m}$、$P_{2\times N}$、$P_{m\times N}$和$P_{2\times m\times N}$，由相应的基本多参随机化模型交叉组合而产生。六种分组多参随机化模型未在图中列出，实际上每个分组随机化模型都可看作是相应模型的特例，比如$P_{m/g}$模型可看作是P_m模型在某些列随机化参数相等时的特例。

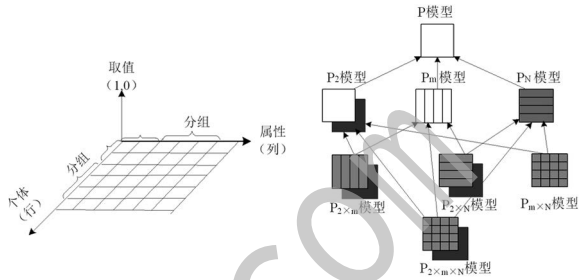

图1 随机化模型分类框架、分类层次

Fig.1 Randomized model classification framework, classification hierarchy

沿分类层次，越往上模型越简单、随机化参数越少，同一层次中，越往左模型越简单、随机化参数越少。且上层模型可看作是下层模型在某些随机化参数相等时的特例，比如最上端的、最简单的单参数随机化P模型，可看作是最下端的多参数随机化$P_{2\times m\times N}$模型在所有$2\times m\times N$个随机化参数都相等时的特例。

目前，研究者已先后提出了跟单参数随机化P模型、二元随机化P_2模型、属性多参随机化P_m模型(图1左上角)分别相匹配的mask、emask和RE三种项集支持度重构算法，用以解决隐私保护的频繁模式挖掘问题。对于除这三种模型外的其他几种较为复杂、保护粒度较细的随机化模型(图1右下角)，还没有相关的研究，而这几种细粒度随机化模型的研究以及与之相适应的项集支持度重构算法的设计，对于进一步加强随机化过程参数设置的灵活性、增强隐私保护挖掘方法的隐私保护性和提高隐私保护挖掘结果的准确性，都将有积极的推动作用。同时，图1右下角与“N”相关的四个模型在个性化方向的扩展，可作为个性化隐私保护方法研究的一个新起点。

4 结论(Conclusion)

本文针对频繁项集挖掘中的隐私保护问题，其主要贡献在于：(1)拓展和丰富了已有的三种简单、粗粒度随机化模型(P、P_2、P_m模型)，提出了三大类细粒度随机化模型，共11种，具体包括：一种行多参随机化模型(PN)，四种复合多参随机化($P_{2\times m}$、$P_{2\times N}$、$P_{m\times N}$和$P_{2\times m\times N}$)，六种分组多参随机化模型($P_{m/g}$、$P_{N/g}$、$P_{2\times m/g}$、$P_{2\times N/g}$、$P_{m/g\times N/g}$、$P_{2\times m/g\times N/g}$)；(2)给出了随机化模型的分类框架、分类层次。

作为个性化隐私保护频繁项集挖掘的一个研究分支，还有很多工作值得今后探索。(1)针对$P_{N/g}$模型的支持度重构方法，是否能推导出该方法所对应的支持计数重构公式和支持度重构偏差公式，是否能设计相应算法实验验证方法的有效性需要进一步研究；(2)这11种细粒度随机化模型的支持度重构方法尚属空白，对此内容研究将极大丰富隐私保护频繁模式挖掘方法；(3)能否将各种随机化模型应用到分类、聚类等挖掘任务，并构造与之适应的特征重构方法，也值得研究。

(下转第43页)