DOI：10.11830/ISSN.1000-5013.201911025

隐私保护频繁项集挖掘中的分组随机化模型

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： 通过对隐私保护频繁项集挖掘问题的研究，发现现有的单参数随机化回答模型调控的数据范围宽、粒度粗，导致无法实现精细化、差异化的隐私保护的问题. 在沃纳模型、单参数等随机化模型的基础上，提出个体分组多参随机化 $P_{N/g}$ 模型，给出其在隐私保护频繁项集挖掘中的支持度重构方法. 研究结果表明：该模型面向多样化、差异化的隐私保护需求，将 N 个不同个体分为若干组，每组设置不同的随机化参数，可实现差异化的隐私保护效果. 实例分析表明：结合所提出的支持度重构方法，可实现基于分组随机化的隐私保护频繁项集挖掘，在保护不同群体隐私的同时，挖掘到有效的频繁项集与关联规则.

关键词： 随机化回答；隐私保护；频繁项集；支持度重构；数据挖掘；沃纳模型

中图分类号： TP 311 **文献标志码：** A **文章编号：** 1000-5013(2020)02-0230-07

Grouping Randomized Model in Privacy Preserving Frequent Item Set 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: Through the research of privacy preserving frequent item set mining, it is found that the existing single-parameter randomized response model regulates the data range wide and the granularity coarse, which leads to the problem that the privacy protection can not be refined and differentiated. Based on Warner model and single-parameter randomization model, an individual grouping multi-parameter randomized model of $P_{N/g}$ is proposed. The corresponding support degree reconstruction method in privacy preserving frequent item set mining is given. The research results show that the model is oriented to diversified and differentiated privacy protection needs, and N different individuals are divided into several groups, and each group is set with different randomization parameters, which can achieve differentiated privacy protection effects. Example analysis shows that combined with the proposed support reconstruction method, privacy preserving frequent item set mining based on grouping randomization can be realized, while protecting the privacy of different groups, effective frequent item sets and association rules can be mined.

Keywords: randomized response; privacy preserving; frequent item set; support reconstruction; data mining; Warner model

数据挖掘能从大量数据中发现新颖的、潜在有用的、可被用户理解的知识，成为一种有效的分析决

收稿日期： 2019-11-08

通信作者： 郭宇红(1979-)，女，副教授，博士，主要从事数据挖掘、推荐系统的研究. E-mail：yhguo@uir.cn.

基金项目： 国际关系学院中央高校基本科研业务费专项资金资助项目(3262017T48，3262018T02)

策手段，在企事业中得到广泛应用．频繁项集挖掘是数据挖掘中的一个重要分支，能从大量数据中发现有趣的关联关系．有效的数据分析需要有大量真实的数据做基础，而人们对数据隐私和安全问题的日益关注，使得在数据收集阶段中，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．因此，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]．

随机化[4-5]是目前隐私保护数据挖掘中运用的主要方法，基本思想是通过向原始数据中加入噪音的方式来对数据作干扰以达到隐私信息的保护，同时数据的统计性质在随机干扰后的数据中保持不变，以获取正确的挖掘结果，包括随机化干扰和随机化回答两种模型．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的调查中．在隐私保护频繁模式挖掘[7-11]、隐私保护关联规则挖掘[12-14]的应用方面，文献[14]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；文献[15]对 MASK(mining associations with secrecy Konstraints)算法进行了扩展，提出“特定于符号(1 和 0)”的随机化过程和相应的 eMASK 算法；文献[16]提出“非统一”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；文献[17]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．

上述随机化回答模型在隐私保护频繁项集挖掘中取得很大进展，但存在以下 2 点问题．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，提出个体分组多参随机化模型 $P_{N/g}$，并结合例子对水平分组随机化的支持度重构方法进行了探索．

1　沃纳模型

沃纳模型是最初由 Warner 在 1965 年针对“吸毒问题的调查”一类敏感问题提出的，可应用于单一属性敏感性问题的统计学调查和分析．在“吸毒问题的调查”这类问题中，调查者想要知道一定人群中吸毒者的比例，但当面对“你是否曾经吸过毒”这类敏感问题的回答时，被调查者(尤其是吸毒者)很可能不愿意回答，或者给出一个虚假的回答．针对这类问题，沃纳模型给出了解决办法．

该模型在调查中设计下面两个对立的问题供被调查者回答：1) 你是否吸过毒；2) 你是否没吸过毒．同时，分给每个被调查者一个随机数生成装置，被调查者可根据生成的随机数的不同，选择回答第 1 个问题，还是第 2 个问题．比如调查者可以跟被调查者事先约定：当生成的随机数小于 p 时，选第 1 个问题回答；大于等于 p 时，选第 2 个问题回答；无论选哪个问题，都要作出真实的回答．

一方面，对于被调查者而言，由于每个被调查者的随机数是随机生成，只他本人知道的，所以他究竟选择了哪个问题作回答也是随机的，只他本人知道的，而外界和调查者并不能从其答案中判断其到底吸过毒还是没吸过毒．因为他们并不知道被调查者究竟回答了哪个问题，这样，被调查者的隐私得到了很好的保护．另一方面，对于调查者，其目标是得到所调查人群中吸毒者的比例．假定分配给每个被调查者的随机数生成装置均相同，都以均匀的概率生成 0 到 1 之间的一个数，则生成的随机数小于 p 的概率为 p．这样，所有的被调查者都将以 p 的概率选择第 1 个问题作答，以 $1-p$ 的概率选择第 2 个问题作答．设 A 表示“吸过毒”，$\overline{\mathrm{A}}$ 表示“没吸过毒”，参与调查的被调查者的总人数为 N，其中“吸过毒”和“没吸过毒”的人数分别为 $N(\mathrm{A})$ 和 $N(\overline{\mathrm{A}})$，则收集到的调查数据中，回答“是”的人数 $N'(\mathrm{yes})$ 的期望值和回答“否”的人数 $N'(\mathrm{no})$ 的期望值分别为

$$\left.\begin{aligned} E[N'(\mathrm{yes})] &= N(\mathrm{A})p + N(\overline{\mathrm{A}})(1-p), \\ E[N'(\mathrm{no})] &= N(\mathrm{A})(1-p) + N(\overline{\mathrm{A}})p. \end{aligned}\right\} \tag{1}$$

式(1)中：$N(\mathrm{A})+N(\overline{\mathrm{A}})=N'(\mathrm{yes})+N'(\mathrm{no})=N$．用实际调查所得 $N'(\mathrm{yes})$，$N'(\mathrm{no})$ 作为近似值替代式(1)中的 $E[N'(\mathrm{yes})]$，$E[N'(\mathrm{no})]$，可估算出调查人群中吸毒者的比例 $\hat{s}=(N(\mathrm{A}))/N=[s'(\mathrm{yes})+p-1]/(2p-1)$．其中，$s'(\mathrm{yes})=(N'(\mathrm{yes}))/N$ 为所收集到的调查数据中回答“是”的被调查者的比例．而 $\hat{s}$ 则正是调查者想要得到的数值．

2 单参数随机化模型

现有隐私保护数据挖掘方法所使用的随机化回答技术,是在沃纳模型的基础上形成的.沃纳模型只能用于单一敏感性问题的调查和分析,其核心思想是在保护个体数据隐私的同时,能求得单一属性上的统计值.对于频繁模式挖掘而言,其对应的数据通常会有多个属性项,频繁模式挖掘的目标则是通过对项集支持度的计算,发现在总体样本中所占比例较高的项集(即频繁项集).因此,对隐私保护频繁模式挖掘,其目标是在保护个体隐私的同时,求取多属性上的统计值——项集支持度.沃纳模型中的公式只能解决隐私保护场景下1-项集支持度的计算.文献[14]提出的MASK方法解决了隐私保护场景下k-项集的支持度计算问题,从而很好地解决了隐私保护频繁模式挖掘问题,其原理如下.

1) 随机化过程.假定用二维0-1矩阵表示原始事务集D,"1"和"0"分别表示对应的项出现和不出现在事务中,则单参数随机化对于D中任意元素$v\in\{0,1\}$,以p的概率取原值v,以$1-p$的概率取$1-v$,生成随机化事务集D'.p称作随机化参数,p值越高,生成的D'中保留越多的原值v.

2) 支持度重构.假定$A=\{I_1,I_2,\cdots,I_k\}$为k-项集,A中的项可能全部或部分出现在D的事务T中.$A\cap T$共有2^k种可能的取值,每一种取值对应了A的一个子集$f_i(i\in\{0,1,\cdots,2^k-1\})$,并假设在二维0-1矩阵表示$D$时,$i$的$k$位二进制数字恰好对应$f_i$的从$I_1$到$I_k$的$k$项0-1序列.即

$$f_0=\varnothing=\underbrace{00\cdots0}_{k},f_1=I_k=\underbrace{00\cdots1}_{k},\cdots,f_{2^k-1}=A=\underbrace{11\cdots1}_{k}.$$

同时,假定$C_{f_i,A\backslash f_i}$表示$D(I_1\cdots I_k)$中仅包含f_i而不包含补集$A\backslash f_i$中的任何项的事务数(f_i在$D(I_1\cdots I_k)$中的净计数).即D中对应A的k列0-1序列等于f_i的事务数.当A在上下文中明确时,$C_{f_i,A\backslash f_i}$简记为C_{f_i}.$C_{f_0},C_{f_1},\cdots,C_{f_{2^k-1}}$(简记为$C_0,C_1,\cdots,C_{2^k-1}$)构成向量$\boldsymbol{C}_A$,即$\boldsymbol{C}_A=[C_0,C_1,\cdots,C_{2^k-1}]^{\mathrm{T}}$;相应地,$C'_f$表示$D'$中仅包含$f$的事务数,向量$\boldsymbol{C}'_A=[C'_0,C'_1,\cdots,C'_{2^k-1}]^{\mathrm{T}}$.则$\boldsymbol{C}_A$和$\boldsymbol{C}'_A$的期望值存在如下关系,即

$$E(\boldsymbol{C}'_A)=\boldsymbol{P}\cdot\boldsymbol{C}_A. \tag{2}$$

式(2)中:$\boldsymbol{P}=[p_{i,j}]$为随机化概率参数p构成的$2^k\times2^k$变换矩阵,$p_{i,j}$表示D中仅包含$f_i(f_i\subseteq A)$的事务(即对应的从I_1到I_k的k项0-1序列恰好为i的k位二进制值的事务)转换成D'中仅包含$f_j(f_j\subseteq A)$的事务的概率.若i和j对应的k位二进制0-1串中值相同的位数为r,则$p_{i,j}=p^r(1-p)^{k-r}(0\leqslant r\leqslant k)$.为便于理解,给出单参数随机化中3-项集的变换概率矩阵,如表1所示.

表1 单参数随机化中3-项集的变换概率矩阵

Tab.1 Transition probability matrix of 3-itemset in single-parameter randomization

			0	1	…	7
			00	001	…	111
			Φ	I_3	…	$I_1I_2I_3$
0	00	Φ	p^3	$p^2(1-p)$	…	$(1-p)^3$
1	001	I_1	$p^2(1-p)$	p^3	…	$(1-p)^2p$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
7	111	$I_1I_2I_3$	$(1-p)^3$	$p(1-p)^2$	…	p^3

据式(2)可得$\boldsymbol{C}_A=\boldsymbol{P}^{-1}E(\boldsymbol{C}'_A)$,实际中,用从$D'$中测得的$\boldsymbol{C}'_A$近似代替$E(\boldsymbol{C}'_A)$,即得到MASK方法对$\boldsymbol{C}_A$的估计值$\hat{\boldsymbol{C}}_A=\boldsymbol{P}^{-1}\cdot\boldsymbol{C}'_A$,而向量$\hat{\boldsymbol{C}}_A$的最后一个元素$\hat{C}_A=\hat{C}_{2^k-1}$正是$k$-项集的支持计数$\hat{S}_A$的估计值.假定$\boldsymbol{P}^{-1}=[a_{i,j}]$,则有

$$\hat{S}_A=\hat{C}_{2^k-1}=a_{2^k-1,0}C'_0+a_{2^k-1,1}C'_1+a_{2^k-1,2^k-1}C'_{2^k-1}=\sum_{j=0}^{2^k-1}a_{2^k-1,j}C'_j. \tag{3}$$

式(3)两边同除以事务总数$|D|$,可得MASK方法对项集A的重构支持度,即

$$\hat{s}_A=\frac{\hat{S}_A}{|D|}=\frac{1}{|D|}\cdot\sum_{j=0}^{2^k-1}a_{2^k-1,j}C'_j=\sum_{j=0}^{2^k-1}a_{2^k-1,j}c'_j. \tag{4}$$

式(4)中:$c'_j=\frac{C'_j}{|D|}$,c'_j表示D'中仅包含A的子集f_j的事务(即从I_1到I_k的k项0-1序列恰好为j的k位二进制值的事务)所占的比例.

以上即为单参数随机化MASK方法在隐私保护频繁模式挖掘中的工作原理.该方法能保证在不访问原始数据D的情况下,从随机化后的数据集D'中估算出各项集的原始支持计数和支持度,从而得到频繁项集和关联规则挖掘结果.

单参数随机化模型的缺点是，所有数据元素的隐私保护程度和最终挖掘结果的准确性全都受控于单一的随机化参数 p. 这不仅忽视不同数据元素隐私保护需求的差异性，使隐私数据不能得到充分有效的保护，而且挖掘结果的准确性也不理想．挖掘结果受 p 的制约很大，p 一旦确定，挖掘结果就确定，挖掘结果准确性上没有任何可调控的余地；而同时对隐私的保护也显得过于鲁棒、不够精准和粒度过粗．

3　分组多参随机化 $P_{N/g}$ 模型

3.1　$P_{N/g}$ 基本思想

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. 参与调查的多个个体的隐私保护要求差不多，则可按个体水平分组随机化，使同一组内共用一个随机化参数，而每个随机化参数控制组中的多行. 假设参与调查的个体总数为 N，若等分时，每组包含 g 行，则组数和随机化参数个数为 N/g，就形成分组多参随机化模型.

3.2　$P_{N/g}$ 模型举例

为简单起见，假定属性取值均为布尔值“1”和“0”. 由这 N 个个体的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．

个体分组随机化的例子，如表 2 所示. 表 2 中：TID 为事务标识号；左边为原始事务集 D，由 10 个被调查者的 3 个问题项（$I_1/I_2/I_3$）组成，10 个被调查者两两一组，同一组内共用同一个随机化参数. 对这五组数据分别随机化后，生成的随机化数据集如表 2 右边 3 列数据所示. 在表 2 中，由个体 1 和 2 构成的第 1 组数据选择的随机化概率参数 $p_1=1$，随机化过程对该组数据以 1 的概率保持为真，以 0 的概率取反，得到的第 1 组随机化数据完全保持不变. 表明该组中的个体完全不顾及隐私，愿意完全真实地贡献其数据. 相反的，由个体 9 和 10 构成的第 5 组数据选择的随机化参数 $p_5=0.6$，随机化过程对该组数据以 0.6 的概率保持为真，以 0.4 的概率取反，得到的第 5 组随机化数据中有 3 个值保持不变，3 个值被打乱取反. 表明该组中的个体相对比较在乎隐私，只肯贡献非常有限的数据．

表 2　数据集 D 分组随机化 $P_{N/g}$ 模型

Tab. 2　Grouping randomization model of $P_{N/g}$ on dataset D

TID	项目	I_1	I_2	I_3		项目	I_1	I_2	I_3
1($p_1=1$)	I_1I_3	1	0	1		I_1I_3	1	0	1
2($p_1=1$)	I_1I_2	1	1	0		I_1I_2	1	1	0
3($p_2=0.9$)	I_3	0	0	1		I_2I_3	0	1	1
4($p_2=0.9$)	I_2	0	1	0		I_2	0	1	0
5($p_3=0.8$)	$I_1I_2I_3$	1	1	1	分组随机化 →	$I_1I_2I_3$	1	1	1
6($p_3=0.8$)	I_4	0	0	0		I_3	0	0	1
7($p_4=0.7$)	I_1I_2	1	1	0		I_1I_2	1	1	0
8($p_4=0.7$)	I_1I_2	1	1	0		I_2	0	1	0
9($p_5=0.6$)	I_2	0	1	0		I_1	1	0	0
10 ($p_5=0.6$)	I_2I_3	0	1	1		I_2	0	1	0

3.3　$P_{N/g}$ 模型支持度重构

分组多参随机化时，需要求得变换概率矩阵 $\boldsymbol{P}$ 和进行支持度重构. 文中计算 $\boldsymbol{P}$ 中元素 $p_{i,j}$ 的基本思想是：D 分组随机化为 D' 作为整体来看时，项集 f_i 转变为 f_j 的概率为各个分组将项集 f_i 转变为 f_j 的概率之和. 相应地，k-项集 A 对应的 $2^k\times 2^k$ 变换概率矩阵 $\boldsymbol{P}_k$ 中的元素值为

$$P_{i,j}=\sum_{i=1}^{n}w_i p_i^r(1-p_i)^{k-r},\qquad 0\leqslant r\leqslant k. \tag{5}$$

对应于矩阵，假设 $\boldsymbol{P}_k^i$ 表示 k-项集 A 对应的第 i 个分组的变换概率矩阵，则有 $\boldsymbol{P}_k=w_1P_k^1+w_2P_k^2+\cdots+w_nP_k^n$. 得到矩阵 $\boldsymbol{P}_k$ 后，就可根据 $\hat{\boldsymbol{C}}_A=\boldsymbol{P}^{-1}\hat{\boldsymbol{C}}'_A$，求得 k-项集 A 的支持计数了，其支持计数恰等于向

量 $\hat{\boldsymbol{C}}_A$ 中的最后一个元素 $\hat{C}_A$.

例如表2中的分组多参随机化中,事务"000"转变"000"的概率为

$$\sum_{i=1}^{5} w_i p_i^3 = 0.2\times(1^3+0.9^3+0.8^3+0.7^3+0.6^3) = 0.56,$$

而"000"转变为"111"(即空集转变为事务$\{I_1I_2I_3\}$)的概率为

$$\sum_{i=1}^{5} w_i (1-p_i)^3 = 0.2\times(0^3+0.1^3+0.2^3+0.3^3+0.4^3) = 0.02.$$

这样便可得到3-项集$\{I_1I_2I_3\}$对应的8×8变换概率矩阵 $\boldsymbol{P}_3$ 中的所有元素.

分组随机化 $P_{N/g}$模型变换概率矩阵 $\boldsymbol{P}_3$,如表3所示.表3中:矩阵 $\boldsymbol{P}_3$ 中的某一元素表示某个项集随机化后转变为另一个项集的概率.

表3 分组随机化 $P_{N/g}$模型变换概率矩阵 $\boldsymbol{P}_3$

Tab.3 Transition probability matrix $\boldsymbol{P}_3$ in grouping randomization of $P_{N/g}$ model

			0	1	…	7
			000	001	…	111
			Φ	I_3	…	$I_1I_2I_3$
0	000	Φ	$\sum_{i=1}^{5} w_i p_i^3$	$\sum_{i=1}^{5} w_i p_i^2(1-p_i)$	…	$\sum_{i=1}^{5} w_i (1-p_i)^3$
1	001	I_3	$\sum_{i=1}^{5} w_i p_i^2(1-p_i)$	$\sum_{i=1}^{5} w_i p_i^3$	…	$\sum_{i=1}^{5} w_i (1-p_i)^2 p_i$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
7	111	$I_1I_2I_3$	$\sum_{i=1}^{5} w_i (1-p_i)^3$	$\sum_{i=1}^{5} w_i (1-p_i)^2 p_i$	…	$\sum_{i=1}^{5} w_i p_i^3$

以表2数据和表3的矩阵为例,根据 $\hat{\boldsymbol{C}}_{\{I_1I_2I_3\}} = \boldsymbol{P}_3^{-1}\boldsymbol{C}'_{\{I_1I_2I_3\}}$,可得

$$\hat{\boldsymbol{C}}_{\{I_1I_2I_3\}} = (\hat{C}_{000},\hat{C}_{001},\cdots,\hat{C}_{111})^{\mathrm{T}} = (\hat{C}_{\Phi},\hat{C}_{\{I_3\}},\cdots,\hat{C}_{\{I_1I_2I_3\}})^{\mathrm{T}} =$$

$$\begin{pmatrix} 0.56 & 0.10 & 0.10 & 0.04 & 0.10 & 0.04 & 0.04 & 0.02 \\ 0.10 & 0.56 & 0.04 & 0.10 & 0.04 & 0.10 & 0.02 & 0.04 \\ 0.10 & 0.04 & 0.56 & 0.10 & 0.04 & 0.02 & 0.10 & 0.04 \\ 0.04 & 0.10 & 0.10 & 0.56 & 0.02 & 0.04 & 0.04 & 0.10 \\ 0.10 & 0.04 & 0.04 & 0.02 & 0.56 & 0.10 & 0.10 & 0.04 \\ 0.04 & 0.10 & 0.02 & 0.04 & 0.10 & 0.56 & 0.04 & 0.10 \\ 0.04 & 0.02 & 0.10 & 0.04 & 0.10 & 0.04 & 0.56 & 0.10 \\ 0.02 & 0.04 & 0.04 & 0.10 & 0.04 & 0.10 & 0.10 & 0.56 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0 \\ 1 \\ 3 \\ 1 \\ 1 \\ 1 \\ 2 \\ 1 \end{pmatrix}. \tag{6}$$

式(6)中:$\hat{\boldsymbol{C}}_{\{I_1I_2I_3\}} = (\hat{C}_{000},\hat{C}_{001},\hat{C}_{010},\hat{C}_{100},\hat{C}_{101},\hat{C}_{110},\hat{C}_{111})^{\mathrm{T}} = (-1.57,1.26,4.90,0.40,0.96,1.01,2.37,0.66)^{\mathrm{T}}$.其中,$\hat{C}_{000},\hat{C}_{001},\cdots,\hat{C}_{111}$分别表示支持计数重构后数据集中 $I_1I_2I_3$ 三列恰好等于000,001,…,111的记录数,即 $\Phi,I_3,\cdots,I_1I_2I_3$ 在重构后数据集中出现的净次数 $\hat{C}_{\Phi},\hat{C}_{\{I_3\}},\cdots,\hat{C}_{\{I_1I_2I_3\}}$.式(6)中矩阵右侧列向量 $\boldsymbol{C}'_{\{I_1I_2I_3\}} = (C'_{000},C'_{001},C'_{010},C'_{011},C'_{100},C'_{101},C'_{110},C'_{111})^{\mathrm{T}} = (0,1,3,1,1,1,2,1)^{\mathrm{T}}$.其中,$C'_{000}=0$ 指表2右侧随机化后的数据集中 $I_1I_2I_3$ 三列恰等于000的记录数,$C'_{001}=1$ 指表2右侧随机化后的数据集中 $I_1I_2I_3$ 三列恰等于001的记录数,以此类推.式(6)中的矩阵就是将 $w_1=w_2=w_3=w_4=w_5=0.2$,$p_1=1$,$p_2=0.9$,$p_3=0.8$,$p_4=0.7$,$p_5=0.6$ 代入表3之后的结果.

3.4 $\boldsymbol{P}_{N/g}$模型支持度重构示例分析

得到 $\hat{\boldsymbol{C}}_{\{I_1I_2I_3\}}$ 后,可据表4中第3列"支持计数算式"求得所有项集的重构支持计数.即第1行算式表示 I_1 的支持计数(数据集中包含 I_1 的事务总数)$S(I_1) = C\{I_1\}+C\{I_1I_3\}+C\{I_1I_2\}+C\{I_1I_2I_3\} = C_{100}+C_{101}+C_{110}+C_{111}$.同样,$I_1$ 重构后的支持计数 $\hat{S}(I_1) = \hat{C}\{I_1\}+\hat{C}\{I_1I_3\}+\hat{C}\{I_1I_2\}+\hat{C}\{I_1I_2I_3\} = 0.96+1.01+2.37+0.66 = 5$.

表4给出了表2数据集 D 对应的项集空间中,所有项集的重构支持计数和重构误差.从表4可看

出：7 个项集支持计数重构总误差为－1.92，平均每个项集的支持计数重构误差为－0.27. 即相对原数据，每个项集支持计数估计值比真实值少了 0.27. 该误差较小，验证了文中所提 $P_{N/g}$ 模型支持度重构方法的可行性和有效性.

表 4　$P_{N/g}$ 和 MASK 支持计数重构对比

Tab. 4　Support count reconstruction comparison of $P_{N/g}$ and MASK

项集	原始支持计数	支持计数算式	$P_{N/g}$ 模型重构支持计数	$P_{N/g}$ 模型重构误差	MASK 重构支持计数	MASK 重构误差
I_1	5	$C_{100}+C_{101}+C_{110}+C_{111}$	0.96＋1.01＋2.37＋0.66＝5.0	0	5.01	＋0.01
I_2	7	$C_{010}+C_{011}+C_{110}+C_{111}$	4.90＋0.40＋2.37＋0.66＝8.33	＋1.33	8.33	＋1.33
I_3	4	$C_{001}+C_{011}+C_{101}+C_{111}$	1.26＋0.40＋1.01＋0.66＝3.33	－0.67	3.33	－0.67
I_1I_2	4	$C_{110}+C_{111}$	2.37＋0.66＝3.03	－0.97	2.78	－1.22
I_1I_3	2	$C_{101}+C_{111}$	1.01＋0.66＝1.67	－0.33	1.67	－0.33
I_2I_3	2	$C_{011}+C_{111}$	0.40＋0.66＝1.06	－0.94	0.56	－1.44
$\underline{I_1I_2I_3}$	1	C_{111}	0.66	－0.34	0.74	$\underline{-0.26}$
总误差				－1.92		－2.58
平均误差				－0.27		－0.37

3.5　与 MASK 方法对比

为进一步验证 $P_{N/g}$ 模型的有效性，将表 2 数据按照 MASK 单参数随机化方法进行支持度重构. 此时，MASK 随机化概率参数等同于平均概率 0.2×(1＋0.9＋0.8＋0.7＋0.6)＝0.8. 根据表 1 变换概率矩阵和 $\hat{\boldsymbol{C}}_{\{I_1I_2I_3\}}=\boldsymbol{P}_3^{-1}\boldsymbol{C}'_{\{I_1I_2I_3\}}$，可得 MASK 方法对表 2 数据重构后各项集的净计数计算公式，即

$$(\hat{C}_{000},\hat{C}_{001},\cdots,\hat{C}_{111})^{\mathrm{T}}=(\hat{C}_{\Phi},\hat{C}_{\{I_3\}},\cdots,\hat{C}_{\{I_1I_2I_3\}})^{\mathrm{T}}=$$

$$\begin{pmatrix}0.512&0.128&0.128&0.032&0.128&0.032&0.032&0.008\\0.128&0.512&0.032&0.128&0.032&0.128&0.008&0.032\\0.128&0.032&0.512&0.128&0.032&0.008&0.128&0.032\\0.032&0.128&0.128&0.512&0.008&0.032&0.032&0.128\\0.128&0.032&0.032&0.008&0.512&0.128&0.128&0.032\\0.032&0.128&0.008&0.032&0.128&0.512&0.032&0.128\\0.032&0.008&0.128&0.032&0.128&0.032&0.512&0.128\\0.008&0.032&0.032&0.128&0.032&0.128&0.128&0.512\end{pmatrix}^{-1}\cdot\begin{pmatrix}0\\1\\3\\1\\1\\1\\2\\1\end{pmatrix}. \tag{7}$$

式(7)中：$(\hat{C}_{000},\hat{C}_{001},\hat{C}_{010},\hat{C}_{011},\hat{C}_{100},\hat{C}_{101},\hat{C}_{110},\hat{C}_{111})^{\mathrm{T}}=(-2.41,1.85,5.74,-0.19,1.30,0.93,2.04,0.74)^{\mathrm{T}}$. 跟 $P_{N/g}$ 模型类似，由表 4 中的“支持计数算式”，可求得 MASK 方法对表 2 数据集 D 所有项集的重构支持计数和重构误差，见表 4 最右侧两列.

对比原始支持计数发现，相对于文中所提 $P_{N/g}$ 模型，MASK 方法仅在支持计数低的项集 $\underline{I_1I_2I_3}$ 上，重构支持计数误差绝对值($\underline{0.26}$)更小，而在支持计数相对高的其他项集上，$P_{N/g}$ 模型的重构误差绝对值小于或等于 MASK，这意味着对频繁项集挖掘，$P_{N/g}$ 模型在频繁项集的支持度重构上将更为准确. 由表 4 可知：整个项集空间支持计数重构的总误差和平均误差绝对值也小于 MASK. 这进一步验证了文中所提 $P_{N/g}$ 模型用于隐私保护频繁项集挖掘的有效性，即 $P_{N/g}$ 模型不仅能实现差异化的隐私保护，且能以小的误差重构频繁项集的支持度. 同时，相对单参数随机化 MASK，多参数随机化 $P_{N/g}$ 模型能在平均隐私保护度相同情况下，以更小的误差重构频繁项集的支持度，从而提高频繁项集挖掘的准确性.

4　结论

针对频繁项集挖掘中的隐私保护问题，提出个体分组多参随机化 $P_{N/g}$ 模型，给出其在隐私保护频繁项集挖掘中的支持度重构方法. 最后，通过示例验证了支持度重构方法的可行性和有效性.

作为个性化隐私保护挖掘的初步尝试，还有如下一些工作需要进一步探究. 1) 针对 $P_{N/g}$ 模型的支持度重构方法，理论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. 2) 设计相应算

法和基于大数据集进一步验证方法的有效性,特别是挖掘结果的准确性.3) 基于新的频繁项集挖掘算法[18],设计与之相适应的、更高效的隐私保护频繁项集挖掘算法.

参考文献:

[1] KENTHAPADI K, MIRONOV I, THAKURTA A G. Privacy-preserving data mining in industry[C]// Proc of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM'19). New York: ACM Press, 2019: 1308-1310. DOI: 10.1145/3308560.3320085.

[2] KOROLOVA A. Privacy-preserving WSDM[C]// Proc of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM'19). New York: ACM Press, 2019: 4. DOI: 10.1145/3289600.3291385.

[3] LI Yaliang, MIAO Chenglin, SU Lu, *et al*. An efficient two-layer mechanism for privacy-preserving truth discovery [C]// Proc of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'18). New York: ACM Press, 2018: 1705-1714. DOI: 10.1145/3219819.3219998.

[4] BULLEK B, GARBOSKI S, MIR D J, *et al*. Towards understanding differential privacy: When do people trust randomized response technique[C]// Proc of the 2017 CHI Conference on Human Factors in Computing Systems (CHI'17). New York: ACM Press, 2017: 3833-3837. DOI: 10.1145/3025453.3025698.

[5] ALDÀ F, SIMON H U. Randomized response schemes, privacy and usefulness[C]// Proc of the 2014 Workshop on Artificial Intelligent and Security Workshop (AISec'14). New York: ACM Press, 2014: 15-26. DOI: 10.1145/2666652.2666654.

[6] WARNER S L. Randomized response: A survey technique for eliminating evasive answer bias[J]. The American Statistical Association, 1965, 60(309): 63-69. DOI: 10.2307/2283137.

[7] 郭宇红,童云海,唐世渭,等.带学习的同步隐私保护频繁模式挖掘[J].软件学报,2011,22(8):1749-1760. DOI:10.3724/SP.J.1001.2011.04000.

[8] SUN Chongjing, FU Yan, ZHOU Junlin, *et al*. Personalized privacy-preserving frequent itemset mining using randomized response[J]. The Scientific World Journal, 2014, 2014: 1-10. DOI: 10.1155/2014/686151.

[9] 丁丽萍,卢国庆.面向频繁模式挖掘的差分隐私保护研究综述[J].通信学报,2014,35(10):200-209. DOI:10.3969/j.issn.1000-436x.2014.10.023.

[10] 许胜之.满足差分隐私保护的频繁模式挖掘关键技术研究[D].北京:北京邮电大学,2016.

[11] 蒋辰,杨庚,白云璐,等.面向隐私保护的频繁项集挖掘算法[J].信息网络安全,2019(4):73-81. DOI:10.3969/j.issn.1671-1122.2019.04.009.

[12] 张鹏,于波,童云海,等.基于随机响应的隐私保护关联规则挖掘[C]//第二十一届中国数据库学术会议论文集.厦门:中国计算机学会,2004:310-313.

[13] 邢欢.基于隐私保护的关联规则挖掘研究[D].南京:南京邮电大学,2016.

[14] RIZVI S J, HARITSA J R. Maintaining data privacy in association rule mining[C]// Proc of the 28th Int'l Conf on Very Large Data Bases (VLDB'02). San Francisco: Margan Kaufmann, 2002: 682-698. DOI: 10.1016/B978-155860869-6/50066-4.

[15] AGRAWAL S, KRISHNAN V, HARITSA J. On addressing efficiency concerns in privacy preserving mining[C]// Proc of the 9th Int'l Conf on Database Systems for Advanced Applications (DASFAA'04). Berlin: Springer-Verlag, 2004: 113-124. DOI: 10.1007/978-3-540-24571-1.

[16] XIA Yi, YANG Yirong, CHI Yun. Mining association rules with non-uniform privacy concerns[C]// Proc of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'04). New York: ACM Press, 2004: 27-34. DOI: 10.1145/1008694.1008699.

[17] ANDRUSZKIEWICZ P. Optimization for MASK scheme in privacy preserving data mining for association rules [C]// Proc of Int'l Conf on Rough Sets and Emerging Intelligent Systems Paradigms (RSEISP'07). Berlin: Springer-Verlag, 2007: 465-474. DOI: 10.1007/978-3-540-73451-2_49.

[18] 张健,刘韶涛.改进的频繁和高效用项集挖掘算法[J].华侨大学学报(自然科学版),2017,38(6):880-885. DOI:10.11830/ISSN.1000-5013.201603067.

(**责任编辑:** 黄仲一 **英文审校:** 吴逢铁)