软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(12):3929−3944 [doi: 10.13328/j.cnki.jos.006101]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

分组随机化隐私保护频繁模式挖掘*

郭宇红[1], 童云海[2], 苏燕青[1]

[1](国际关系学院 网络空间安全学院,北京 100091)

[2](北京大学 智能科学系,北京 100871)

通讯作者: 郭宇红, E-mail: yhguo@uir.cn

摘 要: 已有的隐私保护频繁模式挖掘随机化方法不考虑隐私保护需求差异性,对所有个体运用统一的随机化参数,实施同等的保护,无法满足个体对隐私的偏好.提出基于分组随机化的隐私保护频繁模式挖掘方法 (grouping-based randomization for privacy preserving frequent pattern mining,简称 GR-PPFM).该方法根据不同个体的隐私保护要求进行分组,为每一组数据设置不同的隐私保护级别和与之相适应的随机化参数.在合成数据和真实数据中的实验结果表明:相对于统一单参数随机化 mask,分组多参数随机化 GR-PPFM 不仅能够满足不同群体多样化的隐私保护需求,还能在整体隐私保护度相同情况下提高挖掘结果的准确性.

关键词: 分组;随机化;个性化;隐私保护;频繁模式挖掘

中图法分类号: TP309

中文引用格式: 郭宇红,童云海,苏燕青.分组随机化隐私保护频繁模式挖掘.软件学报,2021,32(12):3929−3944. http://www.jos.org.cn/1000-9825/6101.htm

英文引用格式: Guo YH, Tong YH, Su YQ. Privacy preserving frequent pattern mining based on grouping randomization. Ruan Jian Xue Bao/Journal of Software, 2021,32(12):3929−3944 (in Chinese). http://www.jos.org.cn/1000-9825/6101.htm

Privacy Preserving Frequent Pattern Mining Based on Grouping Randomization

GUO Yu-Hong[1], TONG Yun-Hai[2], SU Yan-Qing[1]

[1](School of Cyber Science and Engineering, University of International Relations, Beijing 100091, China)

[2](Department of Machine Intelligence, Peking University, Beijing 100871, China)

Abstract: Existing randomization methods of privacy preserving frequent pattern mining use a uniform randomization parameter for all individuals, without considering the differences of privacy requirements. This equal protection cannot satisfy individual preferences for privacy. This study proposes a method of privacy preserving frequent pattern mining based on grouping randomization (referred to as GR-PPFM). In this method, individuals are grouped according to their different privacy protection requirements. Different group of data is assigned to different privacy protection level and corresponding random parameter. The experimental results of both synthetic and real-world data show that compared with the uniform single parameter randomization of mask, grouping randomization with multi parameters of GR-PPFM can not only meet the needs of different groups of diverse privacy protection, but also improve the accuracy of mining results with the same overall privacy protection.

Key words: grouping; randomization; personalization; privacy preserving; frequent pattern mining

频繁模式挖掘应用广泛,比如:医学研究人员希望通过分析医学普查数据,发现疾病间的关联,获取并发症等病学知识[1]——例如患糖尿病的人通常伴随着冠心病和高血压.然而在数据普查时,出于隐私的考虑,许多人

* 基金项目: 国家自然科学基金(60403041); 中央高校基本科研业务费专项资金(3262017T48, 3262018T02)

Foundation item: National Natural Science Foundation of China (60403041); Fundamental Research Funds for the Central Universities (3262017T48, 3262018T02)

收稿时间: 2019-08-28; 修改时间: 2019-12-24, 2020-04-04; 采用时间: 2020-06-16

在提供个人数据时会感到不安,有时拒绝提供数据或提供假数据.如何在保护好个人数据隐私的同时实施频繁模式、关联规则等挖掘任务,是隐私保护数据挖掘(privacy preserving in data mining,简称 PPDM)[2]要解决的重要问题,其目标是在不精确访问个体隐私数据的情况下,仍能挖掘到精确的结果.

(1) 相关工作

随机化回答 RR(randomized response)最先由 Warner 在 1965 年针对二元敏感性问题调查提出[3],称为沃纳模型.文献[4]提出了分层沃纳模型,但分层沃纳模型解决的仍是单属性敏感问题的调查,且敏感属性是二元变量.文献[5]使用“风险-效用”映射(risk-utility,简称 R-U)比较了不同的随机化策略,提出了用于单一布尔属性的最优随机化策略.文献[6]利用多目标优化方法,针对单一多元分类属性,力图寻找接近于最优随机化的变换概率矩阵.文献[7]提出了针对多个敏感属性的随机化回答技术.文献[8]通过不相关问题随机化回答技术估算多个混合类型敏感属性的依赖关系,其中,混合类型敏感属性包括你是否抽烟、是否有经济负担等二元分类属性,还包括睡眠质量、手机对学业的影响度等量化数值属性.Du 等人基于随机化回答技术实现了布尔类型数据的隐私保护决策树分类[9].不同文献的区别包括属性类型(二元、多元、量化数值)、属性数量(单属性、多属性)、目标(简单统计、相关性分析、决策树)、随机化问题(正/反问题、正/不相关问题)等.

随机化回答是隐私保护频繁模式和隐私保护关联规则挖掘中的主要方法[10−14].文献[10]提出了基于随机化回答的隐私保护关联规则挖掘方法 mask(mining associations with secrecy konstraints),mask 随机化过程只有一个参数 p,其基本思想是:对布尔数据中所有的“1”“0”值,以 p 的概率保持不变,以 $1–p$ 的概率取反.文献[11]针对数据中“1”“0”敏感度不同的问题提出了 emask 算法,该算法对“1”“0”设置两个不同的随机化参数 p_1 和 p_2,emask 随机化时,对所有的“1”值,以 p_1 的概率保持不变,以 $1–p_1$ 的概率取反;而对“0”值,则以 p_2 的概率保持不变,以 $1–p_2$ 的概率取反.从而使“1”“0”拥有不同的保护级别.文献[12]对 mask 支持度重构进行了性能优化,提出了 mmask 算法.文献[13,14]针对不同属性需要不同保护的场景,提出了“非统一”参数的隐私保护关联规则 RE(recursive estimation)算法.文献[15]提出了属性分组的随机化方法,实现隐私保护关联规则挖掘.近些年流行的差分隐私保护[16−18]通过在数据分析过程或结果中添加随机噪音,确保在数据库中插入或删除任意一条记录都不会显著影响数据分析结果,随机化回答是差分隐私的一种变体[17].

(2) 本文动机

本文动机来自于两方面:一是文献[13],二是客观存在的不同人群隐私保护需求的差异性.

文献[13]RE 算法认为:“性别”“年龄”和“收入”等不同属性的敏感度是不同的,应设置不同的随机化参数,使其拥有不同的隐私保护度.既然不同属性都需要不同保护,那不同个体是否需要不同保护呢?

AT&T 实验室 1999 年调查了 Internet 用户对隐私保护的态度,结果显示[1]:17%的用户对隐私保护极端重视,56%的用户对隐私保护中度重视,其余 27%的用户对隐私保护不重视.以上事实说明,不同人群对隐私态度和对隐私信息的保护需求是有差异的.然而,已有的隐私保护频繁模式挖掘方法没有考虑不同人群的隐私保护需求差异性,在对个体数据随机化时,运用统一的随机化参数对所有人实施同等的保护,无法满足个体对隐私的偏好和具体保护需求,造成的结果是对一部分人的隐私保护程度不足,而对另一部分人实施了过度保护.个性化隐私保护[18−21]应运而生.文献[18]提出一种个性化的差分隐私保护系统.文献[19]面向个性化的隐私保护数据发布.文献[20]提出一种精细化的个性化隐私保护框架.文献[21]提出一种个性化的隐私保护问题调查统计方法,与本文工作相似,它允许用户抽取自己的概率对答案进行干扰.然而,文献[21]的问题和应用场景与本文不同,文献[21]针对在线问题调查,而非频繁项集挖掘.

基于以上事实,本文在我们提出的 $P_{N/g}$ 模型[22]的基础上,提出一种基于个体分组多参随机化的隐私保护频繁模式挖掘方法 GR-PPFM(grouping-based randomization for privacy preserving frequent pattern mining).在 GR-PPFM 架构中,当人们参与数据调查提交自己的数据时,可以根据各自偏好进行分组,每一组数据设置不同的隐私保护级别,进行差异化的隐私保护.本文是我们在文献[22]工作的延续.文献[22]针对 $P_{N/g}$ 模型,总人数为 N,组数为 g,每一分组的人数相同,均为 $_{N/g}$.文献[22]通过简单的例子,手工计算探索了支持度重构的可行性,但没有公式推导、算法设计和实验评价.此外,本文的分组随机化是文献[22]$P_{N/g}$ 模型的更一般情况,分组内人数可以

不同.本文理论上推导了支持度重构递归公式,基于递归公式设计了完整的分组随机化隐私保护频繁项集挖掘算法,并基于大规模合成和真实数据集,针对支持度重构误差和隐私保护性能,与已有的 mask,emask,RE 方法进行了实验对比和评价,验证了方法的有效性.

事实上,隐私保护的内涵决定了其首要目标是为个体提供其所要求的保护,而差异化保护正体现了这一内在目标.GR-PPFM 可在兼顾这种差异化保护要求的同时,保证正常挖掘任务的执行.实验结果表明:相对于已有单参数随机化 mask 方法,GR-PPFM 不仅能满足不同群体多样化的隐私保护需求,加强随机化参数设置的灵活性,还能在整体隐私保护度相同情况下,提高挖掘结果准确性.

1 问题与架构

基于分组随机化的隐私保护频繁模式挖掘 GR-PPFM 的总体架构如图 1 所示,所解决的问题是:给定原始事务集 $D=\{D_1,D_2,\ldots,D_n\}$和最小支持度阈值 min_sup,如何利用 $M_1,M_2,\ldots,M_n$ 共 n 个随机化模型,分别对 $D_1,D_2,\ldots,D_n$ 随机化,以及如何对随机化生成的事务集 $D'=\{D'_1,D'_2,\ldots,D'_n\}$ 进行挖掘,得到跟集合 F 尽可能接近的频繁项集集合,其中,F 为从 D 挖掘得到的频繁项集集合.

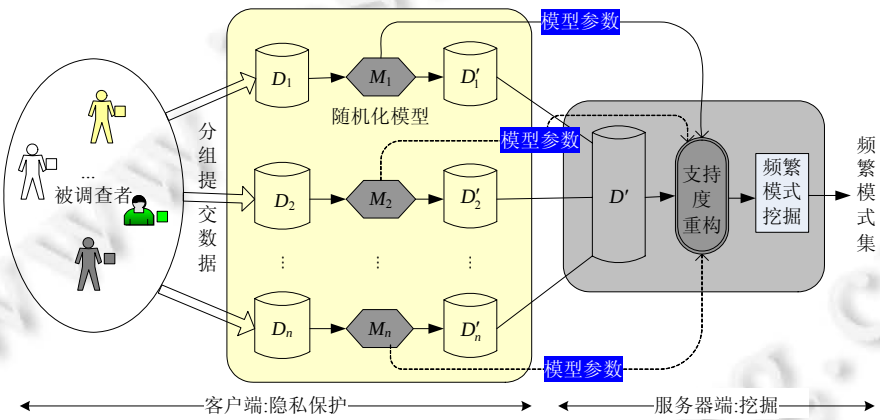


Fig.1 Framework of GR-PPFM

图 1 GR-PPFM 架构

GR-PPFM 分 3 个阶段:(1) 数据分组;(2) 分组随机化;(3) 在支持度重构的基础上,进行频繁模式挖掘.

(1) 第 1 阶段,隐私保护者对参与敏感问题的调查者(即数据提供者),按其对个人数据的保护程度要求进行分组,保护要求相同的进入同一组.可以预先设置若干个保护级别,被调查者可根据个人偏好选择一个合适的保护级别,一个级别对应一个分组.如图 1 所示,共有 n 个保护级别供用户选择,分组后形成了 $D_1,D_2,\ldots,D_n$ 共 n 组数据.假定隐私保护者已根据先验知识,设计好若干个不同的隐私保护级别和对应的分组供被调查者选择,并假设参与调查的个体对其隐私保护取向都比较明确,能够通过引导选定其想要的保护级别和“找到队伍”;

(2) 第 2 阶段,隐私保护者在客户端运用随机化模型,对分组后的数据分别进行随机化,生成随机化后的数据集.如图 1 所示,利用 $M_1,M_2,\ldots,M_n$ 共 n 个随机化模型,对 $D_1,D_2,\ldots,D_n$ 随机化,生成相应的 $D'_1,D'_2,\ldots,D'_n$ 共 n 个随机化数据集.具体的随机化过程和参数设置在第 2.1 节给出;

(3) 第 3 阶段,频繁模式挖掘者在服务器端,对随机化后的数据集 D'进行挖掘,生成想得到的频繁模式集.数据集 D'由 $D'_1,D'_2,\ldots,D'_n$ 共同组成.为了保证从随机化后的数据集能挖掘出正确的频繁模式,以得到正确的关联分析结果,一个很重要的部件是结合随机化模型参数进行项集支持度的重构,第 2.2 节讨论支持度重构.

上述 GR-PPFM 架构中,被调查者本人对持有的数据随机化,随后将随机化数据传给频繁模式挖掘服务器.

2 分组随机化与挖掘方法

2.1 分组随机化

GR-PPFM 分组随机化的基本思想是由参与调查的个体自行决定对其数据的隐私保护级别和相应的随机化参数,隐私保护级别差不多的分为一组,同一组内共用同一个随机化参数,可表示为如下形式:

$$\underbrace{\overset{w_1}{D_1}\overset{w_2}{D_2}\ldots\overset{w_n}{D_n}}_{D}\xrightarrow{p_1p_2\cdots p_n}\underbrace{\overset{w_1}{D'_1}\overset{w_2}{D'_2}\ldots\overset{w_n}{D'_n}}_{D'},$$

其中,原始事务集 D 由 N 个事务(个体)的 m 个项构成二维 $N\times m$ 布尔矩阵(数值属性可通过离散化变为分类属性,而分类属性又可变为布尔属性,即一般数据都可变为二维布尔矩阵形式).$D_1,D_2,\ldots,D_n$ 为 D 中的 n 个数据分组,分组中的个体数占总个体数$|D|=N$ 的比例分别为 $w_1,w_2,\ldots,w_n(0<w_g<1,g\in\{1,\ldots,n\})$.此 n 个数据分组分别以 $p_1,p_2,\ldots,p_n$ 的概率进行随机化,生成随机化后的数据分组 $D'_1,D'_2,\ldots,D'_n$,共同组成频繁模式挖掘所用的事务集 D'.每个 D_g 的随机化都遵循单参数随机化的基本过程,即:对该分组对应的 $N\times m$ 矩阵中的所有 0-1 元素,均以 p_g 的概率取原值,以 $1-p_g$ 的概率取反.单参数随机化使用随机化参数 p 与 $1-p$ 具有对等效果(隐私保护度和挖掘结果准确性关于 $p=0.5$ 对称,$p=0.5$ 隐私保护能力最强,但误差无穷大),故以下假定随机化参数均大于 0.5.

表 1 给出了个体分组随机化的例子.表的左边为原始事务集,由 10 个被调查者的 4 个问题项($I_1/I_2/I_3/I_4$)组成,10 个被调查者分为 5 组,分别包含 3 个、2 个、2 个、2 个、1 个被调查者.同一组内隐私保护需求相同,对应的随机化参数分别为 1,0.9,0.8,0.7,0.6.表的右边为分组随机化后的数据集,其中,前 3 行为第 1 组数据,这 3 名被调查者完全不考虑隐私保护,愿意全部贡献原始数据,随机化概率参数 $p_1=1$;相反,由被调查者 10 构成的第 5 组数据非常在乎隐私,选择的随机化参数 $p_5=0.6$,其数据随机化时以 0.6 的概率保持不变,以 0.4 的概率取反.得到的最后一条记录中,有 2 个值保持不变、2 个值取反.被调查者 4-5,6-7,8-9 的隐私保护需求介于第 1 组和第 5 组之间,随机化参数在 0.6 到 1 之间.

Table 1 Grouping randomization of with GR-PPFM method

表 1 GR-PPFM 方法分组随机化

TID	items	I_1	I_2	I_3	I_4		items	I_1	I_2	I_3	I_4
1 (p_1=1)	I_1I_3	1	0	1	0		I_1I_3	1	0	1	0
2 (p_1=1)	I_1I_2	1	1	0	0		I_1I_2	1	1	0	0
3 (p_2=1)	I_3I_4	0	0	1	1		$I_2I_3I_4$	0	0	1	1
4 (p_2=0.9)	I_2I_4	0	1	0	1	分组	I_2I_4	0	1	0	1
5 (p_3=0.9)	$I_1I_2I_3I_4$	1	1	1	1	随机化	$I_1I_2I_3$	1	1	1	0
6 (p_3=0.8)	I_4	0	0	0	1	——→	I_3I_4	0	0	1	1
7 (p_4=0.8	I_1I_2	1	1	0	0		$I_1I_2I_4$	1	1	0	1
8 (p_4=0.7)	$I_1I_2I_4$	1	1	0	1		I_2I_4	0	1	0	1
9 (p_5=0.7)	I_2I_4	0	1	0	1		I_1I_4	0	0	0	1
10 (p_5=0.6)	$I_2I_3I_4$	0	1	1	1		I_2I_4	0	1	0	0

GR-PPFM 方法采用分组多参随机化,允许对不同的人群使用不同的随机化参数,问题和挑战在于:

(1) GR-PPFM 对不同个体采取不同的随机化参数后,能像单参数随机化 mask 一样重构出原始事务集中各项集的支持度吗?如何重构呢?第 2.2 节针该问题给出解决方法;

(2) GR-PPFM 能从个体分组多参随机化模型中,得到真正的益处吗?第 3 节实验评价将针对该问题作答.

2.2 支持度重构

2.2.1 基本原理

设 $I=\{I_1,I_2,\ldots,I_m\}$是一组项的集合,$D=\{T_1,T_2,\ldots,T_N\}$为事务数据库,其中,事务 $T_u(u\in\{1,2,\ldots,N\})$为 I 的子集.项集 $A\subseteq I$ 的长度$|A|$是指 A 中项的个数,如果$|A|=k$,则称 A 为 k-项集.项集 A 在 D 中的支持计数(简称支持数)是指 D 中包含 A 的事务数,记作 $support_count(A)$或 S_A,$S_A=|\{T_u|A\subseteq T_u,T_u\in D\}|$.同时,将 D 中恰等于 A 的事务数,称作项集 A 在 D 中的净计数,记作 C_A,$C_A=|\{T_u|A=T_u,T_u\in D\}|$.

假定 $A=\{I_1,I_2,\ldots,I_k\}$为 k-项集,根据 mask 方法支持度重构原理,只要给出 k-项集 A 的变换概率矩阵 $P_k=[p_{ij}]$,

就可根据文献[22]中的公式(3):

$$\widehat{S_A} = a_{2^k-1,0}C'_0 + a_{2^k-1,1}C'_1 + \ldots + a_{2^k-1,2^k-1}C'_{2^k-1} = \sum_{j=0}^{2^k-1} a_{2^k-1,j}C'_j \tag{1}$$

估算出项集 A 的重构支持计数 $\widehat{S_A}$.而公式(1)中, $a_{2^k-1,j}$ (j=0,1…,2^k–1)为矩阵 P_k 的逆矩阵 P_k^{-1} 中的最后一行元素, C'_j 为 A 的第 j 个子集在 $D'(I_1\ldots I_k)$中的净计数(即 $D'(I_1\ldots I_k)$中恰等于第 j 个子集的事务数).因此,只要求出变换概率矩阵 P_k,就可求得任意项集的重构支持计数和支持度了.因为求得 P_k 就可得到 P_k^{-1},进而得到 a_{ij}.

2.2.2 变换概率矩阵

根据文献[22]表 1,易推出单参数随机化 4-项集的变换概率矩阵,见表 2.

Table 2 Transition probability matrix P_4 of mask

表 2 mask 变换概率矩阵 P_4

			0	1	…	15
			0000	0001	…	1111
			Φ	I_4	…	$I_1I_2I_3I_4$
0	0000	Φ	$\boldsymbol{p^4}$	$p^3(1-p)$	…	$(1-p)^4$
1	0001	I_4	$p^3(1-p)$	$\boldsymbol{p^4}$	…	$(1-p)^3p$
…	…	…	…	…	…	…
15	1111	$I_1I_2I_3I_4$	$(1-p)^4$	$p(1-p)^3$	…	$\boldsymbol{p^4}$

对于表 1 的分组多参随机化,如何求得变换概率矩阵 P 呢?文献[22]的公式(5)给出了 $P_{N/g}$ 模型 p_{ij} 的计算公式.$P_{N/g}$ 模型每个分组记录数相同,而本文表 1 每个分组记录数不同,但仔细分析发现,文献[22]的公式(5)同样适用于分组记录数不同的随机化.不过,文献[22]的公式(5)各组权重角标所用的组号 i,容易 与 p_{ij} 中的角标 i 混淆,本文使用 w_g 和 p_g,分别表示第 g 个分组所占的比例权重和对应的随机化参数,得到 GR-PPFM 方法中 k-项集 A 对应的 $2^k\times 2^k$ 变换概率矩阵 P_k 中的元素值:

$$p_{ij} = \sum_{g=1}^{n} w_g p_g^r (1-p_g)^{k-r} (0 \leqslant r \leqslant k).$$

例如表 1 中的分组随机化,事务“0000”转变“0000”的概率为

$$p_{00} = \sum_{g=1}^{5} w_g {p_g}^4 (1-p_g)^0 = 0.3\times 1^4 + 0.2\times 0.9^4 + 0.2\times 0.8^4 + 0.2\times 0.7^4 + 0.1\times 0.6^4 = 0.57412;$$

而“0001”(对应事务$\{I_4\}$)转变为“1110”(对应事务$\{I_1I_2I_3\}$)的概率为

$$p_{1,14} = \sum_{g=1}^{5} w_g {p_g}^0 (1-p_g)^4 = 0.3\times 0^4 + 0.2\times 0.1^4 + 0.2\times 0.2^4 + 0.2\times 0.3^4 + 0.1\times 0.4^4 = 0.00418.$$

这样便可得到 4-项集$\{I_1I_2I_3I_4\}$对应的 16×16 变换概率矩阵 P_4 中的所有元素,见表 3.

Table 3 Transition probability matrix P_4 of GR-PPFM

表 3 GR-PPFM 变换概率矩阵 P_4

			0	1	…	15
			0000	0001	…	1111
			Φ	I_4	…	$I_1I_2I_3I_4$
0	0000	Φ	$\sum_{g=1}^{5} w_g p_g^4$	$\sum_{g=1}^{5} w_g p_g^3(1-p_g)$	…	$\sum_{g=1}^{5} w_g (1-p_g)^4$
1	0001	I_4	$\sum_{g=1}^{5} w_g p_g^3(1-p_g)$	$\sum_{g=1}^{5} w_g p_g^4$	…	$\sum_{g=1}^{5} w_g (1-p_g)^3 p_g$
…	…	…	…	…	…	…
15	1111	$I_1I_2I_3I_4$	$\sum_{g=1}^{5} w_g (1-p_g)^4$	$\sum_{g=1}^{5} w_g (1-p_g)^3 p_g$	…	$\sum_{g=1}^{5} w_g p_g^4$

在得到矩阵 P_k 后,就可根据 $\overrightarrow{C_A} = P_k^{-1}\cdot\overrightarrow{C'_A}$,求得 k-项集 A 的支持计数了,其支持计数恰等于向量 $\overrightarrow{C_A}$ 中的最后

一个元素 $\widehat{C_A}$.

文献[22]第 3.3 节、第 3.4 节给出了手工进行支持度重构的完整例子.

2.2.3 支持计数重构递归公式

有两种方法可加快求解整个项集空间的 2^m 个项集支持度的计算过程:第 1 种方法是根据 $\overrightarrow{\widehat{C_I}}=P_m^{-1}\cdot\overrightarrow{C_I'}$ 求取 2^m 个项集在 D 中的净计数,然后由项集的支持计数与净计数的关系 $\overrightarrow{\widehat{S_I}}=T\cdot\overrightarrow{\widehat{C_I}}$ 求得此 2^m 个项集的支持计数;第 2 种方法是导出项集支持计数重构递归公式,见后文公式(4),根据该递归公式只需 2^m 次求解,便可求取整个项集空间的 2^m 个项集的支持度.下面给出公式(4)的推导过程:

假设 $\overrightarrow{S_A'}=[S_0',S_1',...,S_{2^k-1}']$ 为 k-项集 A 的 2^k 个子集在 D' 中的支持数构成的向量,由文献[13]命题 1,知 $\overrightarrow{S_A'}=TP_kT^{-1}\overrightarrow{\widehat{S_A}}$.其中,$T=[T_{ij}]$为 $2^k\times 2^k$ 矩阵,满足:

$$T_{ij}=T(i,j)=T(f_i,f_j)=\begin{cases}1, & f_i\subseteq f_j;\\ 0, & \text{others.}\end{cases}$$

其中,f_i 和 f_j 均为 A 的子集.令 $U=P_kT^{-1}$,则根据 $P_k=w_1P_k^1+w_2P_k^2+...+w_nP_k^n$,可得:

$$U=(w_1P_k^1+w_2P_k^2+...+w_nP_k^n)T^{-1}=\sum_{g=1}^{n}w_gP_k^gT^{-1}.$$

令 $V=TU$,则有 $V=T\sum_{i=1}^{n}w_iP_k^iT^{-1}=\sum_{i=1}^{n}w_iTP_k^iT^{-1}$,由文献[13]的公式(11),易推得:

$$V(i,j)=V(f_i,f_j)=\begin{cases}\sum_{g=1}^{n}w_g(2p_g-1)^{|f_j|}(1-p_g)^{|f_i|-|f_j|}, & f_j\subseteq f_i;\\ 0, & \text{others.}\end{cases}\tag{2}$$

结合 $\overrightarrow{S_A'}=TP_kT^{-1}\overrightarrow{\widehat{S_A}}=V\overrightarrow{\widehat{S_A}}$ 和公式(2),得到向量 $\overrightarrow{S_A'}$ 的最后一个元素 S_A' 满足:

$$S_A'=\sum_{f\subseteq A}\left(\sum_{g=1}^{n}w_g(2p_g-1)^{|f|}(1-p_g)^{|A|-|f|}\right)\widehat{S_f}=\left(\sum_{g=1}^{n}w_g(2p_g-1)^{|A|}\right)\widehat{S_A}+\sum_{f\subset A}\left(\sum_{g=1}^{n}w_g(2p_g-1)^{|f|}(1-p_g)^{|A|-|f|}\right)\widehat{S_f}\tag{3}$$

根据公式(3),得到项集支持计数重构递归公式如下:

$$\widehat{S_A}=\frac{S_A'-\sum_{f\subset A}\left(\sum_{g=1}^{n}w_g(2p_g-1)^{|f|}(1-p_g)^{(|A|-|f|)}\right)\widehat{S_f}}{\sum_{g=1}^{n}w_g(2p_g-1)^{|A|}},\widehat{S_\varnothing}=|D|\tag{4}$$

2.3 GR-PPFM挖掘方法

GR-PPFM 利用支持数重构递归公式(4),在频繁项集生成算法 Apriori 基础上形成,支持数重构递归公式是算法的核心,Apriori 构成了方法的主框架.

算法. 分组随机化隐私保护频繁项集生成方法 GR-PPFM.

输入:分组多参随机化数据 D';分组比例和随机化参数$(p_g,w_g)(g=1,\ldots,n)$;最小支持度阈值 min_sup;

输出:从 D' 重构出的频繁项集集合 $\hat{F}$.

1. 扫描事务集 D',记录每个项 j 的支持计数 $j.S'$,所有项的集合构成 I;
2. **for** each item $j\in I$:
   (a) $j.s'\leftarrow\frac{j.S'}{|D|}$; //得到项 j 在 D' 中的支持度;

(b) $j.\hat{s} \leftarrow \dfrac{j.s' - \left(1 - \sum_{g=1}^{n} w_g p_g\right)}{2\sum_{g=1}^{n} w_g p_g - 1}$; //重构项 j 在 D 中的支持度;

3. $\widehat{F_1} \leftarrow \{j \in I \mid j.\hat{s} \geqslant \min\_\sup\}$; //得到重构频繁 1-项集集合;

4. **for** (k=2; $\widehat{F_{k-1}} \neq \varnothing$; k++):

(a) $\widehat{C_k} \leftarrow$ ***apriori_gen***$(\widehat{F_{k-1}}, \min\_\sup)$; //由 $\widehat{F_{k-1}}$ 生成候选频繁 k-项集集合 $\widehat{C_k}$;

(b) **for** each transaction $t \in D'$, //扫描 D'记录每个候选 k-项集的支持计数

for each candidate $c \in \widehat{C_k}$,

if $c \subseteq t$ **then** $c.S'$++;

(c) **for** each candidate $c \in \widehat{C_k}$:

(i) $c.s' \leftarrow \dfrac{c.S'}{|D|}$; //得到候选频繁 k-项集 c 在 D'中的支持度;

(ii) $c.\hat{s} \leftarrow \dfrac{c.s' - \sum_{f \subset c}\left(\sum_{g=1}^{n} w_g (2p_g - 1)^{|f|}(1 - p_g)^{(k-|f|)}\right) f.\hat{s}}{\sum_{g=1}^{n} w_g (2p_g - 1)^k}$; //重构 c 在 D 中的支持度;

(d) $\widehat{F_k} \leftarrow \{c \in C_k | c.\hat{s} \geqslant \min\_\sup\}$; //得到重构频繁 k-项集集合;

5. **Return** $\widehat{F} \leftarrow \bigcup_k \widehat{F_k}$;

3 实验评价

3.1 实验数据

分别用人工合成购物篮数据集、真实购物篮数据集进行实验评价.

人工合成购物篮数据集.人工合成购物篮数据集 D 由 IBM Almaden 生成器生成,生成器参数为 T=3, I=4,$|D|$=100K,N=10,即事务平均长度为 3,频繁项集的平均长度为 4,总事务数为 100K,总项数为 10.直接生成的数据集为项集形式,可将其转化为 0,1 布尔表示的数据集;

真实购物篮数据.真实购物篮数据集 D 为某食品超市的购物数据 basket.txt,事务平均长度为 3,总事务数 940, 总项数为 11, 包括 fruitveg,freshmeat,dairy,cannedveg,cannedmeat,frozenmeal,beer,wine,softdrink,fish, confectionery.该数据可从以下网址获取:https://download.csdn.net/download/lol000/8693253(2020 年 2 月).

3.2 实验方法

- 第 1 步,挖掘原始数据集 D.

针对多个不同的最小支持度阈值,分别运用 Apriori 算法对数据集 D 进行挖掘,记录每次挖掘得到的所有频繁项集和其支持数.

- 第 2 步,生成分组多参数随机化数据集.

对数据集 D 进行分组多参数随机化干扰,生成干扰后的数据集 D'.具体地讲,对数据集 D 按行分为 Group1~Group5 共 5 组数据,这 5 组数据所占的比例分别为 w_1=30%,w_2=20%,w_3=20%,w_4=20%,w_5=10%,对应的随机化参数分别为 p_1=1,p_2=0.9,p_3=0.8,p_4=0.7,p_5=0.6.即:第 1 组数据保持不变;第 2 组数据以 0.9 的概率保持原来的值,以 0.1 的概率取反;第 3 组~第 5 组数据分别以 0.8,0.7,0.6 的概率保持原值,以 0.2,0.3,0.4 的概率取反.直观地,数据集 D 对应的分组多参随机化模型参数设置见表 4.

以上 5 组数据所占比例,大致依据本文开始提到的 AT&T 实验室 1999 年隐私态度调查报告中不同用户的

比例和进一步细分.隐私保护级别设置大致遵从国家保密局 2007 年 6 月 22 日颁布的《信息安全等级保护管理办法》中的五级分类.表 4 中:1 级表示信息密级为公开,不需保护;2 级表示信息密级为限制,需弱保护;3 级表示信息密级为秘密,需保护;4 级表示信息密级为机密,需强保护;5 级表示信息密级为绝密,需特别强的保护.5 级分类不仅考虑了信息保护需求的差异性,同时级数设置较少易于管理,实际中也可根据需要进一步分级细化.

- 第 3 步,生成单参数随机化数据集.

为便于同单参数 mask 方法对比,对数据集 D 进行单参数随机化干扰,生成 mask 方法所用的单参数随机数据集 D'_{mask},随机化概率 p 设置为多参数随机化的平均概率,即:

$$p=\sum_{g=1}^{5}w_g p_g=30\%\times 1+20\%\times 0.9+20\%\times 0.8+20\%\times 0.7+10\%\times 0.6=0.84.$$

- 第 4 步,挖掘随机化数据.

针对多个不同的最小支持度阈值,分别运用 Apriori 算法以及带有支持数重构的 GR-PPFM 方法,对第 3 步生成的多参数随机化数据集 D'进行挖掘,记录每次挖掘得到的所有频繁项集和其支持数.同时,运用 Apriori 算法和 mask 方法,对第 4 步生成的单参数随机化数据集 D'_{mask} 挖掘,记录挖掘得到的频繁项集和对应的支持数.

- 第 5 步,计算分析误差.

对比第 4 步的挖掘结果,计算分析 mask 方法、GR-PPFM 方法的挖掘结果误差,包括项集支持度相对误差ρ、项集身份误差θ^-和θ^+.其中,ρ反映频繁项集在随机化数据中重构后的支持度与其在原数据中的实际支持度间的相对误差;θ^-表示频繁项集丢失率,衡量原先频繁而被错误识别为不频繁的项集占原频繁项集总数的比例;θ^+表示频繁项集增加率,衡量原先不频繁而被错误识别为频繁的项集占原频繁项集总数的比例.具体公式见文献[1].

Table 4 Randomization parameter settings of dataset D

表 4 数据集 D 的随机化模型参数设置

分组	所占比例(%)	随机化参数	隐私保护级别(密级)
Group1	30	1	1 级(公开)
Group2	20	0.9	2 级(限制)
Group3	20	0.8	3 级(秘密)
Group4	20	0.7	4 级(机密)
Group5	10	0.6	5 级(绝密)

3.3 结果分析

本节对实验结果分析,考察不同方法挖掘结果的误差随项集长度 k、最小支持度阈值 min_sup 的变化情况,并对不同方法的误差进行对比.其中,图 2 为合成数据上的实验结果,图 3 为真实数据上的实验结果.

3.3.1 误差与项集长度的关系

3.3.1.1 支持度误差

图 2(a)给出了针对合成数据,当最小支持度阈值 min_sup=0.1%时,mask 和本文 GR-PPFM 方法的平均支持度相对误差ρ随频繁项集长度 k 的变化曲线.图 3(a)给出了针对真实数据,当 min_sup=1%时,ρ随 k 的变化曲线.

(1) 横向比较:很明显,本文提出的多参数随机化 GR-PPFM 的误差小于单参数随机化 mask 方法.说明了在整体隐私保护度相同的情况下(实验中,mask 的随机化参数 p 等于多参数随机化的平均概率 $\bar{p}$),GR-PPFM 挖掘结果好于 mask;

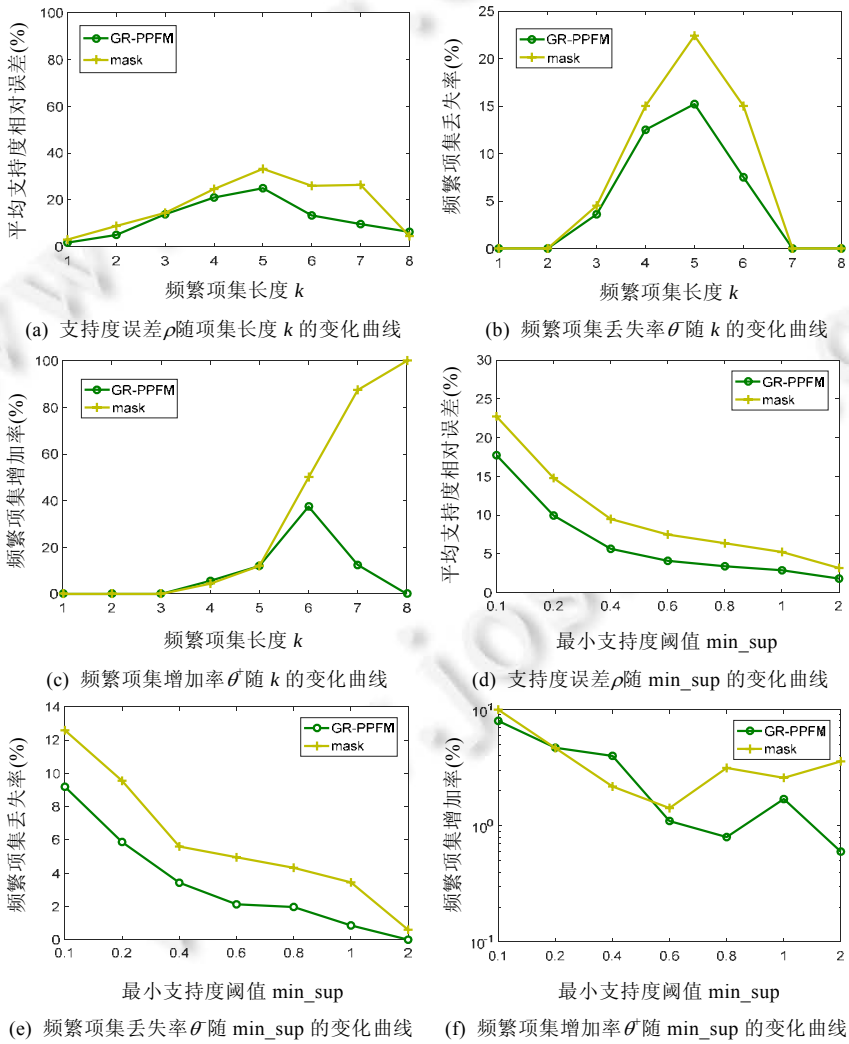
(2) 误差随频繁项集长度 k 的变化:

a. 理论上,从 k-项集的支持度重构递归公式(4)分析,k-项集的重构支持度 $\hat{s}_k$ 依赖于该 k-项集的所有子集的重构支持度 $\hat{s}_{k-1},\hat{s}_{k-2},...,\hat{s}_1$,作为一个递归的过程,$\hat{s}_{k-1}$ 的支持度依赖于 $\hat{s}_{k-2},...,\hat{s}_1$,依此类推.这种递归依赖关系将导致误差的级联传导,使误差从 1-项集逐渐向上层传递和累积,因此直观上,k 越大,理论偏差也越大;

b. 观察图 2(a)发现,平均支持度相对误差随 k 大体呈现先上升、后下降的趋势.在频繁 5-项集处,平均支持度误差最大,5-项集之前误差随 k 的增加而增大,5-项集之后误差大致平缓下降.为什么图中的误差在频繁 5-项集后会出现回落呢?回到ρ的计算公式,不难发现,ρ计算的是频繁项集的平均支持度相对误差,这意味着ρ与最小

支持阈值 min_sup 紧密相关.因为 min_sup 不同,其划定的某个长度的频繁项集集合 F 也就不同,造成该集合的平均支持度误差ρ也会不同.这表明,F 对应的ρ值会因特定数据的偏斜和最小支持度阈值 min_sup 的不同而出现偶然变大变小的情况.因为 F 中即使出现 1 个误差特别大或特别小的项集,就能显著改变 F 的平均支持度误差ρ值,尤其当 F 中的项集个数较少时.

图 2(a)中,在 min_sup=0.1%时,原始数据对应的频繁 1-项集集合、2-项集集合、…、8-项集集合的项集个数依次为 10,44,112,160,125,40,8,1.其中,频繁 4 项集的个数最多(这是 IBM Almaden 生成器生成的数据集特征决定的,因为选取的参数指定了生成数据集的频繁项集的平均长度为 4),而频繁 6-项集、7-项集和 8-项集集合的个数较少,尤其是频繁 8-项集集合,只有 1 个项集,因此其平均支持度误差的偶然性就较大,这也是图中 5-项集以后误差回落的原因.当 F 中的项集个数较多时,偶然性因素会被淹没,误差大小会顺从一般的规律随项集长度的增加而增大.事实上,实验中发现:当设置 min_sup=0.001%时,即只要出现 1 次的项集就为频繁项集(因为实验所用的数据集 D 的大小为 100 000)时,所测得的ρ正是按项集长度的递增而递增的.因为此时每一级长度对应的频繁项集个数都较多,平均误差随项集长度的变化能呈现理论分析的规律.



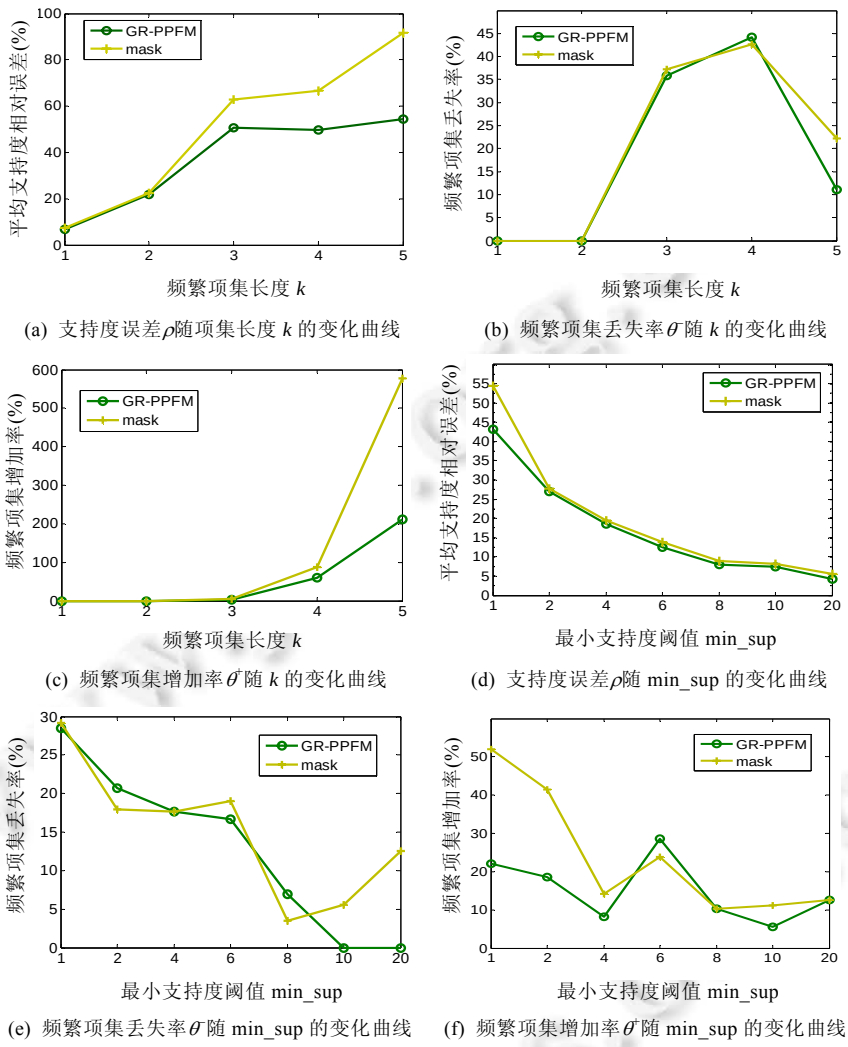
(a) 支持度误差ρ随项集长度 k 的变化曲线 (b) 频繁项集丢失率θ^-随 k 的变化曲线

(c) 频繁项集增加率θ^+随 k 的变化曲线 (d) 支持度误差ρ随 min_sup 的变化曲线

(e) 频繁项集丢失率θ^-随 min_sup 的变化曲线 (f) 频繁项集增加率θ^+随 min_sup 的变化曲线

Fig.2 Experiment error of mask, and GR-PPFM on synthetic data

图 2 mask、GR-PPFM 在人工合成数据中的实验误差



(a) 支持度误差ρ随项集长度 k 的变化曲线 (b) 频繁项集丢失率θ^-随 k 的变化曲线

(c) 频繁项集增加率θ^+随 k 的变化曲线 (d) 支持度误差ρ随 min_sup 的变化曲线

(e) 频繁项集丢失率θ^-随 min_sup 的变化曲线 (f) 频繁项集增加率θ^+随 min_sup 的变化曲线

Fig.3 Experiment error of mask and GR-PPFM on real-world data

图 3 mask 与 GR-PPFM 在真实数据中的实验误差

图 3(a)测得的ρ正是按项集长度的递增而递增的,同理论分析一致.

3.3.1.2 项集身份误差

图 2(b)、图 3(b)和图 2(c)、图 3(c)给出了项集身份误差随频繁项集长度 k 的变化情况,可以看出:(1) 分组随机化 GR-PPFM 方法误差小于单参数随机化 mask 方法;(2) 项集身份误差随 k 的变化跟图 2(a)、图 3(a)中支持度误差ρ随 k 的变化情况相近,误差大致随 k 增大而增大.

θ^-,θ^+随 k 变化规律与ρ随 k 变化规律的相似性是容易理解的,因为追根溯源,项集支持度大小决定了项集作为频繁项集还是非频繁项集的身份,项集支持度误差从最深层次反映了随机化过程对于数据的影响,项集身份误差是项集支持度误差的外在表现.

3.3.2 误差与支持度阈值的关系

3.3.2.1 支持度误差

图 2(d)给出了合成数据所有频繁项集(从频繁 1-项集到频繁 8-项集,k=ALL)的平均支持度相对误差ρ随最小支持度阈值 min_sup 的变化曲线.图 3(d)给出了真实数据上ρ随 min_sup 的变化曲线.

(1) 横向比较:横向比较图 2(d)、图 3(d)可发现:误差大小关系跟图 2(a)、图 3(a)一致,仍是 GR-PPFM<mask.说明在整体隐私保护度相同时,多参数随机化方法 GR-PPFM 的挖掘结果优于单参数 mask 方法;

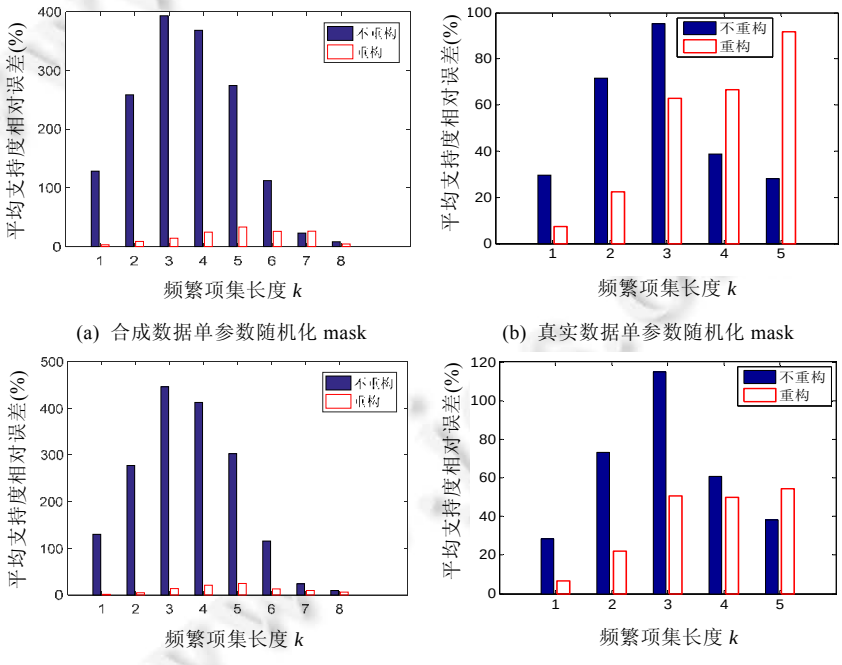
(2) 误差随支持度阈值 min_sup 的变化:观察曲线随 min_sup 的变化可发现,平均支持度误差随支持度阈值的增大而减小.说明随着支持度阈值的增大,挖掘结果越好.原因是什么呢?由于支持度相对误差等于绝对误差与原始支持度值的比值,假定项集 I_1 和 I_2 的绝对误差相等,则项集 I_1 和 I_2 的相对误差完全取决于其原支持度值:原支持度值越大,其相对误差越小;原支持度值越小,其相对误差越大.当支持度阈值增大时,其对应的频繁项集集合 F 中各项集的支持度相对越大,造成 F 中各项集的平均支持度相对误差越小,误差随 min_sup 的变化呈现图 2(d)、图 3(d)中的趋势.

3.3.2.2 项集身份误差

图 2(e)、图 3(e)和图 2(f)、图 3(f)给出了项集身份误差随支持度阈值 min_sup 的变化情况.可看出,项集身份误差随 min_sup 的变化跟图 2(d)、图 3(d)支持度误差ρ随 min_sup 的变化情况大体相近,其基本规律:(1) 大多数情况下,分组随机化 GR-PPFM 方法的误差小于单参数随机化 mask 方法;(2) 误差随 min_sup 增大而减小.θ^-,θ^+随 min_sup 变化规律与ρ随 min_sup 变化规律相似.

3.3.3 支持度重构与不重构误差对比

通常,数据随机化后,由于数据被扰乱,项集的支持度将发生变化,若直接从随机化后的数据挖掘,不进行支持度重构,项集的支持度跟原始支持度比究竟会发生多大的变化呢?图 4(a)~图 4(d)分别给出了实验中针对合成数据和真实数据、单参数随机化 mask(随机化概率 p=0.84)支持度重构与不重构的误差对比及分组随机化 GR-PPFM 支持度重构与不重构误差对比.图 4 中,合成数据、真实数据设置的最小支持度阈值分别为 0.1%,1%.



(a) 合成数据单参数随机化 mask (b) 真实数据单参数随机化 mask

(c) 合成数据个体分组随机化 GR-PPFM (d) 真实数据个体分组随机化 GR-PPFM

Fig.4 Error of support reconstruction vs. non-reconstruction

图 4 支持度重构与不重构误差对比

图 4(a)、图 4(c)针对 IBM Almaden 生成器生成的数据,可发现支持度不重构的误差远远大于重构误差.说明数据经随机化后,项集的支持度已发生显著变化,直接从随机化后的数据得到的挖掘结果已远远偏离从原始数

据挖掘得到的结果,必须通过支持度重构恢复原数据库中各项集的支持度,以得到跟原数据尽可能相近的挖掘结果.

图 4(b)、图 4(d)针对真实购物篮数据,可发现当频繁项集长度不大时,不重构的误差远远大于重构误差;但当频繁项集长度较大时(单参数随机化后 k=4,5,分组随机化 k=5),不重构的误差反而小于重构的误差.说明对于高阶项集重构的误差大,这跟之前分析的误差随 k 增大而增大的规律一致.综合图 4,针对分组随机化,对于低阶项集(k<5)使用重构支持度,对于高阶项集(k≥5)使用不重构支持度.

3.3.4 隐私保护对比

下面分析两种方法的隐私保护性能,从隐私保护度(定量)、个体隐私保护差异性和暴露的信息(定性)等方面进行对比.本文隐私保护性能考虑的场景是敏感问题调查或购物时,对于敏感问题回答为“是”和“购买”敏感物品的保护,不考虑对于敏感问题回答为“否”和“不购买”敏感物品的保护.并假定被调查者运用随机化装置给出了真实的回答,购物数据提供者对数据进行了相应的随机化变换.

3.3.4.1 隐私保护度

文献[10]将单参数随机化 mask 的隐私保护度 *privacy* 定义为 $1-R(p)$,其中,$R(p)=aR_1(p)+(1-a)R_0(p)$.其中,$R_1(p)$表示原始数据库中的“1”能从随机化后的数据库中被还原的概率,$R_0(p)$表示原始数据库中的“0”能从随机化后的数据库中被还原的概率,a 为隐私保护权重.本文隐私保护场景只考虑敏感问题回答为“是”和“购买”敏感物品记录的保护,即对如表 1 所示的 0-1 购物篮数据,只考虑“1”值的保护.设保护权重系数 a=1,则隐私保护度公式为

$$privacy=1-R_1(p) \tag{5}$$

假定随机化概率为 p,项的平均支持度为 s_0,$R_1(p)$的计算公式为

$$R_1(p)=\frac{p^2 s_0}{(1-p)(1-s_0)+p\,s_0}+\frac{(1-p)^2 s_0}{p(1-s_0)+(1-p)\,s_0} \tag{6}$$

根据公式(5)、公式(6)的分析,隐私保护度 $1-R_1(p)$跟 $p(p>0.5)$和 s_0 均成反比.说明随机化概率越大,项的平均支持度越高,隐私保护度越低;反之亦然.极端地,当 p=1 时,数据完全保持不变,$R_1(p)$=1,隐私保护度最低,为 0.

- 当 s_0=1 时,数据是全 1 数据,$R_1(p)$=1,隐私保护度也是最低,为 0.此时,无论 p 取多少,“1”均能从随机化后的数据中被还原,随机化均无法保护数据;
- 当 s_0=0 时,数据是全 0 数据,$R_1(p)$=0,隐私保护度最高,为 1.

对于分组随机化,由于不同分组随机化概率 p 不同,所以不同分组的隐私保护度也不同.假定 w_g 为第 g 个分组个体数占总个体数的比例,p_g 为第 g 个分组对应的随机化概率,$R_1(p_g)$为第 g 个分组中的“1”能从随机化后的数据库中被还原的概率,$privacy(g)=1-R_1(p_g)$为第 g 个分组对应的隐私保护度.对于分组随机化,在已知每个分组对应的随机化概率的条件下,定义如下 4 个隐私保护度:最低隐私保护度、最高隐私保护度、平均隐私保护度、整体隐私保护度.

定义 1(最低隐私保护度 minPrivacy). 分组随机化中,隐私保护度最小的分组对应的隐私保护度.公式为

$$minPrivacy=\min\{privacy(g)|g=1,2,\ldots,n\} \tag{7}$$

定义 2(最高隐私保护度 maxPrivacy). 分组随机化中,隐私保护度最大的分组对应的隐私保护度.公式为

$$maxPrivacy=\max\{privacy(g)|g=1,2,\ldots,n\} \tag{8}$$

定义 3(平均隐私保护度 avgPrivacy). 分组随机化中,多个分组隐私保护度的平均值称为平均隐私保护度.计算公式为

$$avgPrivacy=\sum_{g=1}^{n} w_g\,privacy(g) \tag{9}$$

定义 4(整体隐私保护度 overallPrivacy). 将分组随机化的平均概率 $\bar{p}$ 代入公式(5),求得的隐私保护度称为整体隐私保护度.计算公式为

$$overallPrivacy = 1 - R_1(\overline{p}), \overline{p} = \sum_{g=1}^{n} w_g p_g \tag{10}$$

实验中,IBM Almaden 生成器生成的合成数据中,项的平均支持度 s_0=40.69%;真实数据中,项的平均支持度 s_0=27.08%.项的平均支持度 s_0 取决于原数据,事实上,由于原数据是无法知道的,真实的 s_0 无从得知,实际中,可通过先验知识或抽样估计 s_0 值,也可通过重构得到的项的平均支持度代替 s_0 值.将 s_0 和 p_1=1,p_2=0.9,p_3=0.8,p_4=0.7,p_5=0.6, $\overline{p}$ =0.84 代入公式(5)~公式(10),可分别求得合成数据和真实数据分组随机化时,在已知每个分组对应的随机化概率的条件下,GR-PPFM 对应的最低隐私保护度、最高隐私保护度、平均隐私保护度和整体隐私保护度,结果见表 5.

Table 5 Privacy of GR-PPFM vs. mask

表 5 GR-PPFM 与 mask 方法隐私保护性能对比

	分组多参数随机化 GR-PPFM		整体单参数随机化 mask	
	合成数据	真实数据	合成数据	真实数据
最低隐私保护度	0	0	32.4%	43.4%
最高隐私保护度	57.0%	70.6%		
平均隐私保护度	27.8%	35.9%		
整体隐私保护度	32.4%	43.4%		
隐私保护差异性	个体有差异		个体无差异	
个体随机化概率	不暴露		暴露	
暴露的信息	D',(p_1,w_1),(p_2,w_2),(p_3,w_3),(p_4,w_4),(p_5,w_5)		D',p	

单参数随机化 mask 仅对应一个隐私保护度,实验中,单参数 mask 方法的随机化概率 p=0.84,分组多参随机化的平均概率 $\overline{p}$ =0.84,两者相等.因此,单参数随机化的隐私保护度与分组随机化的整体隐私保护度相同,合成数据是 32.4%,真实数据是 43.4%.

3.3.4.2 隐私保护差异性

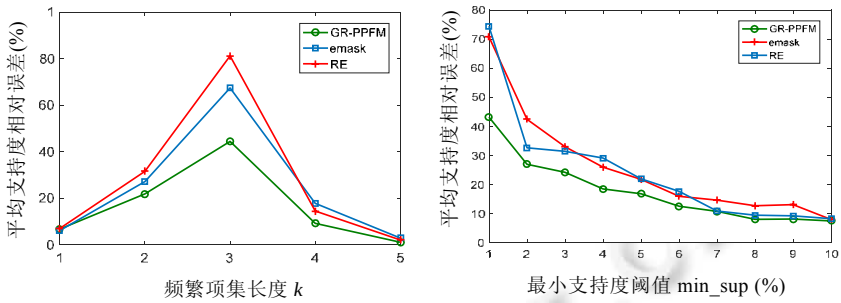
除定量考察隐私保护度外,还可定性分析两种方法对个体隐私保护的差异性及为支持度重构需暴露的信息的差异.GR-PPFM 对个体实施有差异的保护,而 mask 实施无差异的保护.

从个体对应的随机化概率是否暴露给挖掘者来分析.在分组随机化 GR-PPFM 方法中,为了进行支持度重构,挖掘者只需要知道个体大致使用了哪些随机化参数以及相应的个体比例,而不需要知道具体的个体与使用的随机化参数的确切对应关系.但 mask 支持度重构需要知道 p,由于所有的个体都使用了同样的 p,所以挖掘者确切知道每个个体与随机化参数的对应关系,即任意指定一个个体,挖掘者都能确切知道该个体使用了什么样的参数进行了随机化变换.

两种方法的隐私保护度、隐私保护差异性和为支持度重构需要暴露的信息见表 5,其中,D'表示 D 随机化后的全部数据, D'_g 表示 D 中以 p_g 的概率随机化后的第 g 组数据,其占全部数据的比例为 w_g.实验中,p_1=1,p_2=0.9,p_3=0.8,p_4=0.7,p_5=0.6;w_1=0.3,w_2=0.2,w_3=0.2,w_4=0.2,w_5=0.1.

3.4 拓展实验

为进一步探究本文工作 GR-PPFM 与其他相关工作的效能差异,本节设计实验对 GR-PPFM,emask 和 RE 这 3 种算法进行对比.GR-PPFM 按行分组随机化,emask 对 0,1 分别随机化,RE 按列分组随机化,实验数据采用第 3.1 节提到的真实购物篮数据,GR-PPFM 采用第 3.2 节实验方法第 2 步使用的分组比例和相应的随机化参数设置,平均随机化概率 $\overline{p}$ =0.84.Emask 设置两个随机化参数:p_1=0.6,p_0=0.9,即:对所有的“1”值,以 0.6 的概率保持不变,以 0.4 的概率取反;而对“0”值,则以 0.93 的概率保持不变,以 0.07 的概率取反.由于“1”和“0”在数据中的占比分别为 72.9%和 27.1%,平均随机化概率 $\overline{p}$ =0.6×72.9%+0.93×27.1%=0.84.RE 将购物篮数据中的 11 种商品分为 3 组:(1) fruitveg,freshmeat,dairy;(2) cannedveg,cannedmeat,frozenmeal,beer,wine;(3) softdrink,fish,confectionery.第(1)组~第(3)组随机化参数分别设为 0.68,0.84,1,平均随机化概率 $\overline{p}$ =(3×0.68+5×0.84+3×1)/11=0.84.3 种方法的平均随机化概率均为 0.84,保证三者的整体隐私保护度相同.图 5 给出了 GR-PPFM,emask 和 RE 这 3 种方法在整体隐私保护度相同的情况下挖掘结果的差异.



(a) 支持度误差ρ随项集长度 k 的变化曲线 (b) 支持度误差ρ随 min_sup 的变化曲线

Fig.5 Support error of GR-PPFM/emask/RE on real-world basket data

图 5 GR-PPFM/emask/RE 在真实购物数据中的支持度误差

图 5(a)为支持度阈值 1%时,支持度相对误差随挖掘出的频繁项集长度变化的比较;图 5(b)为支持度误差随最小支持度阈值变化的比较.可以看出:在整体隐私保护度相同情况下,本文提出的个体分组随机化 GR-PPFM 的误差均小于相关工作提到的 0,1 差异随机化 emask 方法及属性分组随机化 RE 方法.

4 总结与展望

已有以 mask 为代表的隐私保护频繁模式挖掘方法,在对数据随机化时,采用统一的随机化参数,对所有个体实施同等、无差异的保护.考虑到不同人群对隐私保护需求的差异性,本文提出一种基于分组随机化的隐私保护频繁模式挖掘 GR-PPFM 方法.针对 GR-PPFM,探索了支持度重构方法以及与之相适应的频繁模式挖掘算法,实现了粗粒度的个性化隐私保护频繁模式挖掘.实验结果表明:

(1) 在整体隐私保护度相同情况下,分组随机化 GR-PPFM 挖掘结果准确性高于整体单参数随机化 mask;

(2) 分组随机化对不同人群实行有差异的保护,个体隐私保护度更强,为支持度重构暴露的信息更少;

(3) 支持度重构误差随项集长度的增大而增大、随支持度阈值的增大而减小,支持度重构误差远远小于直接挖掘不重构的误差.

本文的贡献在于,基于不同人群隐私保护需求不同的事实,提出了分组多参数随机化的数据保护方式;给出了分组随机化的支持度重构方法以及与之相适应的频繁模式挖掘算法,实现了粗粒度的个性化隐私保护频繁模式挖掘;实验分析比较了分组多参数随机化和单参数随机化两种方法的效果.

GR-PPFM 方法可广泛应用于社会、经济生活中涉及隐私的敏感性问题的调查和关联分析任务.未来工作包括:(1) 探索误差随分组参数 w、随机化概率参数 p 和数据集大小参数$|D|$的变化情况;(2) 本文只通过实验证实了分组随机化 GR-PPFM 方法的可行性及相对于单参数随机化 mask 方法的优越性,能否从理论上导出两种方法 k-项集的支持度重构偏差公式,并证明 GR-PPFM 方法的优越性值得探究;(3) 能否将分组随机化模型应用到其他类型的个性化隐私保护挖掘算法,如分类、聚类[23],并构造与之适应的特征重构方法,也是十分值得期待的研究工作.

References:

[1] Guo YH, Tong YH, Tang SW, *et al*. Learning and synchronized privacy preserving frequent pattern mining. Ruan Jian Xue Bao/ Journal of Software, 2011,22(8):1749−1760 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4000.htm [doi: 10.3724/SP.J.1001.2011.04000]

[2] Kenthapadi K, Mironov I, Thakurta AG. Privacy-preserving data mining in industry. In: Proc. of the 12th ACM Int'l Conf. on Web Search and Data Mining (WSDM 2019). New York: ACM, 2019. 840−841. [doi: 10.1145/3308560.3320085]

[3] Warner SL. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 1965,60(309):63−69. [doi: 10.1080/01621459.1965.10480775]

[4] Kim JM, Warde WD. A stratified Warner's randomized response model. Journal of Statistical Planning and Inference, 2004, 120(1-2):155−165. [doi: 10.1016/S0378-3758(02)00500-1]

[5] Huang ZL, Du WL, Teng ZX. Searching for better randomized response schemes for privacy-preserving data mining. In: Proc. of the 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007). LNCS 4702, Springer-Verlag, 2007. 487−497. [doi: 10.1007/978-3-540-74976-9_50]

[6] Huang ZL, Du WL. OptRR: Optimizing randomized response schemes for privacy-preserving data mining. In: Proc. of the 24th IEEE Int'l Conf. on Data Engineering (ICDE 2008). IEEE Computer Society, 2008. 705−714. [doi: 10.1109/ICDE.2008.4497479]

[7] Tamhane AC. Randomized response techniques for multiple sensitive attributes. Journal of the American Statistical Association, 1981,76(376):916−923. [doi: 10.1080/01621459.1981.10477741]

[8] Chu AM, So MK, Chan TW, *et al*. Estimating the dependence of mixed sensitive response types in randomized response technique. Statistical Methods in Medical Research, 2020,29(3):894−910. [doi: 10.1177/0962280219847492]

[9] Du W, Zhan Z. Using randomized response techniques for privacy-preserving data mining. In: Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2003). ACM, 2003. 505−510. [doi: 10.1145/956750.956810]

[10] Rizvi SJ, Haritsa JR. Maintaining data privacy in association rule mining. In: Proc. of the 28th Int'l Conf. on Very Large Data Bases (VLDB 2002). Morgan Kaufmann Publishers, 2002. 682−698. [doi: 10.1016/B978-155860869-6/50066-4]

[11] Agrawal S, Krishnan V, Haritsa J. On addressing efficiency concerns in privacy preserving mining. In: Proc. of the 9th Int'l Conf. on Database Systems for Advanced Applications (DASFAA 2004). LNCS 2973, Springer-Verlag, 2004. 113−124. [doi: 10.1007/978-3-540-24571-1_9]

[12] Andruszkiewicz P. Optimization for mask scheme in privacy preserving data mining for association rules. In: Proc. of Int'l Conf. on Rough Sets and Emerging Intelligent Systems Paradigms (RSEISP 2007). LNAI 4585, Springer-Verlag, 2007. 465−474. [doi: 10.1007/978-3-540-73451-2_49]

[13] Xia Y, Yang Y, Chi Y. Mining association rules with non-uniform privacy concerns. In: Proc. of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2004). ACM, 2004. 27−34. [doi: 10.1145/1008694.1008699]

[14] Sun CJ, Fu Y, Zhou JL, *et al*. Personalized privacy-preserving frequent itemset mining using randomized response. Scientific World Journal, 2014. 1−10. [doi: 10.1155/2014/686151]

[15] Teng ZX, Du WL. A hybrid multi-group approach for privacy-preserving data mining. Knowledge and Information Systems, 2009, 19(2):133−157. [doi: 10.1007/s10115-008-0158-y]

[16] Xu SZ. Research on differentially private frequent pattern mining techniques [Ph.D. Thesis]. Beijing: Beijing University of Posts and Telecommunications, 2016 (in Chinese with English abstract).

[17] Bullek B, Garboski S, Darakhshan JM, *et al*. Towards understanding differential privacy: When do people trust randomized response technique? In: Proc. of the 2017 CHI Conf. on Human Factors in Computing Systems (CHI 2017). New York: ACM, 2017. 3833−3837. [doi: 10.1145/3025453.3025698]

[18] Vu XS, Jiang LL, Brändström A, *et al*. Personality-based knowledge extraction for privacy-preserving data analysis. In: Proc. of the Knowledge Capture Conf. (K-CAP 2017), Vol.45. New York: ACM, 2017. 1−4. [doi: 10.1145/3148011.3154479]

[19] Xiao XK, Tao YF. Personalized privacy preservation. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD 2006). New York: ACM, 2006. 229−240. [doi: 10.1145/1142473.1142500]

[20] Song XM, Wang X, Nie LQ, *et al*. A personal privacy preserving framework: I let you know who can see what. In: Proc. of the 41st Int'l ACM SIGIR Conf. on Research & Development in Information Retrieval (SIGIR 2018). New York: ACM, 2018. 295−304. [doi: 10.1145/3209978.3209995]

[21] Li YL, Miao,CL, Su L, *et al*. An efficient two-layer mechanism for privacy-preserving truth discovery. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining (KDD 2018). New York: ACM, 2018. 1705−1714. [doi: 10.1145/3219819.3219998]

[22] Guo YH, Tong YH. Grouping randomized model in privacy preserving frequent item set mining. Journal of Huaqiao University (Natural Science), 2020,41(2):230−236 (in Chinese with English abstract). [doi: 10.11830/ISSN.1000-5013.201911025]

[23] Li B. Research on personalized privacy protection method for clustering mining [Ph.D. Thesis]. Harbin: Harbin Engineering University, 2017 (in Chinese with English abstract).

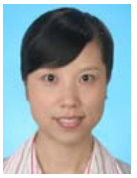
附中文参考文献:

[1] 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.软件学报,2011,22(8):1749−1760. http://www.jos.org.cn/1000-9825/4000.htm [doi: 10.3724/SP.J.1001.2011.04000]

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

[22] 郭宇红,童云海.隐私保护频繁项集挖掘中的分组随机化模型.华侨大学学报(自然科学版),2020,41(02):230−236.

[23] 李保.面向聚类挖掘的个性化隐私保护方法研究[博士学位论文].哈尔滨:哈尔滨工程大学,2017.



郭宇红(1979－),女,博士,副教授,主要研究领域为数据挖掘,隐私保护,信息系统.

苏燕青(1997－),男,学士,主要研究领域为网络空间安全.

童云海(1971－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为数据挖掘与知识发现,智能媒体发现,隐私保护.