软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(8):1749−1760 [doi: 10.3724/SP.J.1001.2011.04000]

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

带学习的同步隐私保护频繁模式挖掘*

郭宇红[1,2+], 童云海[2], 唐世渭[2], 吴冷冬[3]

[1](国际关系学院 信息科技系,北京 100091)

[2](北京大学 机器感知与智能教育部重点实验室,北京 100871)

[3](Department of Computing Science, University of Alberta, Edmonton T6G 2R3, Canada)

Learning and Synchronized Privacy Preserving Frequent Pattern Mining

GUO Yu-Hong[1,2+], TONG Yun-Hai[2], TANG Shi-Wei[2], WU Leng-Dong[3]

[1](Department of Information Technology, University of International Relations, Beijing 100091, China)

[2](Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China)

[3](Department of Computing Science, University of Alberta, Edmonton T6G 2R3, Canada)

+ Corresponding author: E-mail: yhguo@uir.cn

Guo YH, Tong YH, Tang SW, Wu LD. Learning and synchronized privacy preserving frequent pattern mining. *Journal of Software*, 2011,22(8):1749−1760. http://www.jos.org.cn/1000-9825/4000.htm

Abstract: To improve the accuracy of mining results, this paper proposes a method of privacy preserving frequent pattern mining, based on sample learning and synchronized randomization of itemset (LS-PPFM). The method utilizes the data of individuals who do not care privacy. First, the data that does not need protecting are learned, and some strongly associated items are obtained. Then, when the data is randomized, the associated items are bound together and randomized synchronously to try to keep their potential associations. Experimental results show that compared with independent randomization, LS-PPFM can achieve significant improvements on accuracy, while losing a little privacy.

Key words: supervised; learning-based; randomization; privacy preserving; frequent pattern mining

摘 要: 为了提高挖掘结果的准确性,提出基于样例学习和项集同步随机化的隐私保护频繁模式挖掘方法(learning and synchronized privacy preserving frequent pattern mining,简称 LS-PPFM).该方法充分利用不需要隐私保护的个体数据,首先对不需要保护的数据学习,得到样例数据中蕴涵的强关联项,然后在对数据随机化时,将强关联项绑定在一起作同步随机化变换,以保持项与项之间的潜在关联性.实验结果表明,相对于项独立随机化,LS-PPFM能够在略微牺牲一定的隐私保护性的情况下,显著提高频繁模式挖掘结果的准确性.

关键词: 有指导的;基于学习的;随机化;隐私保护;频繁模式挖掘

中图法分类号: TP311 **文献标识码**: A

频繁模式挖掘能够从数据中发现有趣的关联,在实际中有着十分广泛的应用.比如,医学研究人员希望通过

* 基金项目: 国家自然科学基金(60403041, 60473072)
收稿时间: 2010-04-16; 定稿时间: 2011-01-20
CNKI 网络优先出版: 2011-03-16 11:30, http://www.cnki.net/kcms/detail/11.2560.TP.20110316.1130.001.html

对医学普查数据的分析,发现不同疾病间的关联,获取并发症等病学知识,从而更好地指导医疗工作;药品公司希望对其客户进行疾病调查,分析不同疾病的关联性,从而更好地指导其药品销售——例如,通过发现多数人会同时患有糖尿病和冠心病,将治疗糖尿病的药物和治疗冠心病的药物捆绑销售.

然而人们对数据隐私的关注,往往限制和阻碍正常的数据采集和后期的频繁模式挖掘任务.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的各种任务和应用,成为数据挖掘和信息安全领域结合后的一个重要研究问题.隐私保护频繁模式挖掘正是针对这一问题,其目标是在不精确访问个体数据、不泄露个人隐私的情况下,仍能从数据中得到精确的频繁模式挖掘结果,既不丢失正常频繁模式,也不产生虚假频繁模式.

(1) 相关工作

针对隐私保护频繁模式挖掘问题,现有的方法可以分为两类:安全多方计算的方法[1,2]和随机化方法[3−11].

安全多方计算方法运用密码技巧和安全计算技术,对分布式环境下多个参与关联规则挖掘的实体,通过协同的分布式计算,获取全局的数据挖掘结果.文献[1]基于安全点积加密协议,针对垂直划分的分布式隐私保护关联规则挖掘.文献[2]基于安全求并运算,针对垂直划分的分布式隐私保护关联规则挖掘.安全多方计算的优点是,该方法得到的是精确的挖掘结果.其缺点是,该方法只能用于分布式环境,且每一步的安全计算都需大量的计算和通信开销;同时,分布式协同计算中的单点故障将导致错误结果的产生,甚至造成挖掘无法进行.

随机化是目前隐私保护数据挖掘中的主要方法.文献[3]提出了基于随机化回答的 mask(mining associations with secrecy konstraints)方法,通过数据干扰和支持度重构实现隐私保护的关联规则挖掘,mask 方法中随机化参数只有 1 个,所有的数据元素受控于唯一的参数.文献[4]对 mask 算法作了扩展,提出“特定于符号(1 和 0)”的随机化过程(symbol-specific 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[5]提出了“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,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[6]提出 FRAPP 框架,试图通过选择合适的干扰矩阵,最大化挖掘结果准确性.文献[7]提出了部分隐藏随机化回答的隐私保护关联规则挖掘方法.文献[8]提出了隐私保护频繁模式挖掘的增量算法.文献[9]对 mask 算法在支持度重构复杂度方面作了优化.文献[10]比较了不同的随机化策略,提出了用于布尔数据和分类数据的最优随机化策略.文献[11]利用多目标优化方法,力图寻找接近于最优随机化的变换概率矩阵.

文献[12]综合安全多方计算准确性高和随机化方法简单、高效的特点,针对 ID3 决策树和关联规则挖掘,提出了混合的基于属性分组的随机化方法.文献[13]基于随机化回答,实现了隐私保护决策树分类.文献[14]研究数据发布中的匿名保护,考虑属性顺序敏感的分析任务.近 10 年数据挖掘、发布中的隐私保护研究综述见文献[15].

(2) 本文动机

本文的想法源自我们多次的随机化实验发现和对相关工作的分析.

一方面,在多次的随机化实验中我们发现,如果不对数据随机化,而只利用原始数据的一部分数据作挖掘,挖掘结果的准确性有时反而比利用整个随机化的数据作挖掘的准确性好.这说明,部分数据会在一定程度上携带整体数据的统计特征.那么,在实际调查中是否能够得到一些个体的真实数据呢?是否能够恰当利用这部分数据对挖掘作指导和帮助呢?针对第 1 个问题,答案是肯定的.事实上,AT&T 实验室 1999 年的一项关于 Internet 用户对隐私问题态度的调查报告恰证实了这样的论断.报告显示:有 17%的被调查者对隐私极端重视,即使采取了有效的保护方法也不愿提供任何信息;有 56%的被调查者对隐私问题虽然关注,但在有效的保护措施下愿意提供相应信息;另外 27%的被调查者对隐私问题不关注,通常都愿意提供信息.这说明隐私保护需求因人而异,实际数据调查中是能够采集到部分真实数据的.上述第 2 个问题正是需要研究和探索的.

另一方面,仔细分析随机化的误差来源发现,除了文献[12,13]以外,已有方法对数据随机化时,对事务中每一个项都采取独立的随机化方式.独立随机化势必会在一定程度上破坏原数据中项与项之间的联系.由于频繁模式挖掘的任务正是要发现项与项间的关联,项与项的关联被破坏后必然造成挖掘结果出现大的偏差;且项集越长,其内在的关联关系被破坏的程度越大,挖掘得到的长模式的支持度就越偏离其原始值,误差也就越大.那么,

能否对存在关联关系的项采取同步随机变换方式,使得随机化后的数据能保留它们的内在联系呢?问题在于,如何知道哪些项存在关联,哪些项不存在关联.

结合第 1 个方面的分析,我们的想法是:既然部分数据一定程度上携带了整体数据的统计特征,那么能否从不要求保护的部分个体数据中学习和发现一些关联特征,然后利用这些关联特征指导其他个体作同步随机化呢?基于该想法,本文提出一种基于样例学习的同步随机化隐私保护频繁模式挖掘方法 LS-PPFM(learning and synchronized randomization in privacy preserving frequent pattern mining).其核心思想是"有指导的同步",LS-PPFM 首先对不要求隐私保护的个体记录学习其关联,然后在对数据随机化时,对从真实样本中学习到的、存在关联的项,采取同步随机化变换,以将学习到的关联注入到随机化后的数据,从而提高挖掘结果准确性.

本文提出的 LS-PPFM 弥补了文献[13]全同步和文献[12]盲目分组同步的不足.文献[13]提出将所有属性作为一个属性同步随机变换,但所有属性同步(1-group,全同步)会造成隐私保护度急剧下降,且文献[13]针对决策树分类.文献[12]在随机化时提出将属性强行分为若干组(如 2-group,3-group),每一组作为一个属性随机化.分组同步固然好,但其分组是硬性、盲目和缺乏指导的,这会出现将无关联的属性分在一组,而关联性强的属性被分开的情况.结果不但不会提升关联规则挖掘结果准确度,反而会降低挖掘结果准确度,并有可能因过度同步造成隐私保护度的大幅下降.且文献[12]是混合了安全多方计算的方法,而非纯随机化方法.本文 LS-PPFM 是有选择的、有指导的同步,对潜在的强关联属性同步,而对非潜在关联属性独立.在尽可能保持属性关联的同时,又减少了过多同步造成的隐私保护性的下降.我们的目标在于寻求挖掘结果准确度和隐私保护度的最佳平衡.

1 问题与架构

本文所提出的有指导的、同步随机化隐私保护频繁模式挖掘 LS-PPFM 架构如图 1 所示.它所解决的问题是:给定原始事务集 D、D 中一批不需要隐私保护的个体数据集 D_0 和最小支持度阈值 min_sup,如何利用 D_0 指导随机化模型 M,以及如何对利用 M 随机化后的事务集 D'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.LS-PPFM 用于面向挖掘的数据调查情景,分 4 个阶段:(1) 提交数据;(2) 特征学习;(3) 项集同步随机化;(4) 在支持度重构的基础上进行频繁模式挖掘.

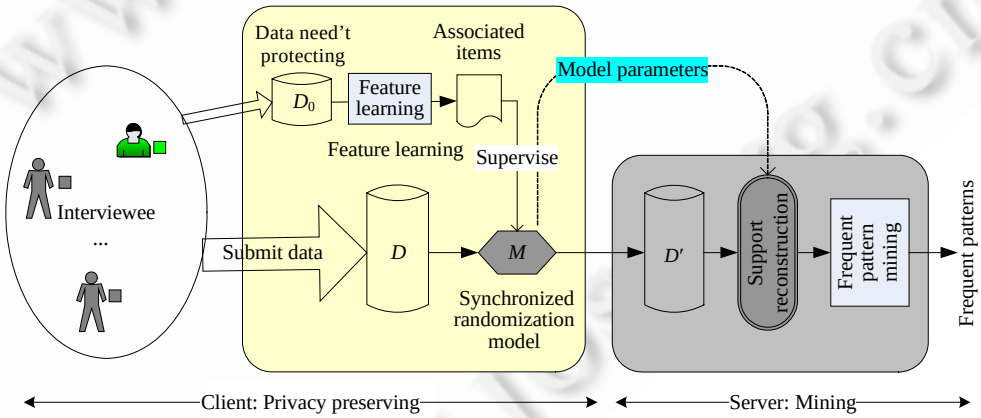


Fig.1 Framework of LS-PPFM

图 1 LS-PPFM 架构

第 1 阶段,被调查者提交数据.一方面收集所有被调查者的数据,提交至数据集 D;另一方面收集对隐私不敏感的个体数据,提交至不需要保护的数据集 D_0.

第 2 阶段,对不需要保护的数据集 D_0 进行特征学习,发现其中蕴含的强关联项,利用强关联项指导数据集 D 的随机化.本文第 2.1 节将给出具体的样例数据特征学习方法.

第 3 阶段,运用项集同步随机化模型 M 对数据集 D 进行随机化.M 对 D 中通过学习 D_0 所发现的强关联项

实施同步随机化,对其他项实施独立随机化.本文第 2.2 节对项集同步随机化进行说明.

第 4 阶段,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(或关联规则集).其中,一个很重要的部件是结合同步随机化模型参数进行支持度重构,第 2.3 节对此进行讨论.

2 有指导的项集同步随机化与挖掘方法

2.1 特征学习

在 LS-PPFM 中,很关键的一个阶段是对不需要隐私保护的样例数据 D_0 进行特征学习,发现其中的强关联项.本文基于频繁项集支持度,认为越频繁出现的项集,其关联性越强.这时,只需对 D_0 进行频繁模式挖掘,将支持度较高的项集作为强关联项.一般地,对于包含有 m 个项的 D_0,其对应的非空集项集格共包含 2^m–1 个项集,可以选择项集格某一层次中支持度最高的项集作为强关联项.

例如,若表 1 为特征学习样例数据,则可以选择如图 2 所示(图中脚标为项集的支持计数)的项集格的第 2 层中支持数最高的{*BD*}作为强关联项.在整个数据随机化时,将{*B*}{*D*}所在的两列绑定为一列作同步随机变换.具体选择第几层的频繁项集以及选择几个频繁项集,可以由用户指定.基本的原则是选支持度较高的频繁项集;同时,若选择多个频繁项集,则这些频繁项集之间不能有交叉项.比如,一旦选择了{*BD*}作为强关联项绑定,就不能再选择{*AB*},{*AD*},{*BC*},{*CD*}作绑定了,但仍可以选择{*AC*}作绑定(如果认为{*AC*}的支持度已足够高).若上移一层,则可以选{*ABD*}或{*BCD*}作 3 个项的绑定.但越往上移,项集的支持度越小,其对于项集关联性的支持越弱;而且参与同步的项越多,越易被恶意推理反推数据,使得隐私保护性下降.

Table 1 Sample data for feature learning

表 1 特征学习样例数据

TID	Items	*A*	*B*	*C*	*D*
1	*AC*	1	0	1	0
2	*AB*	1	1	0	0
3	*CD*	0	0	1	1
4	*BD*	0	1	0	1
5	*ABCD*	1	1	1	1
6	*D*	0	0	0	1
7	*AB*	1	1	0	0
8	*ABD*	1	1	0	1
9	*BD*	0	1	0	1
10	*BCD*	0	1	1	1

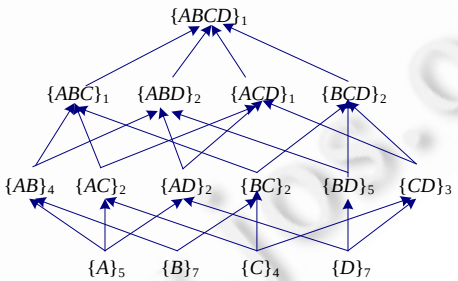


Fig.2 Support counts of the itemsets in the itemset lattice of the sample data

图 2 样例数据对应的项集格中各项集的支持计数

项集的支持度刻画了项集中的各项同时出现的频度,从一定程度上反映了各项之间的关联关系.实际中,可设定最小支持度阈值,筛选出频度高的项集集合,从而确定同步随机变换的项,以指导整个数据样本的随机化.

2.2 项集同步随机化

LS-PPFM 在对不要求隐私保护的个体记录进行特征学习得到强关联项后,即可在数据随机化时对强关联

项采取同步随机化.项集同步随机化的目的是保持强关联项间的关联关系,基本原理是对多个具有强关联关系的项作绑定,在随机化变换时,同时保持原值或同时取反.例如,假设 I_1 和 I_2 为强关联项,事务 t 在属性项 I_1,I_2 上的取值分别为 0,1,记作 $t=(I_1=0)\wedge(I_2=1)$,令 $\bar{t}=(I_1=1)\wedge(I_2=0)$,假定随机化概率为 p,则在对 t 随机化时,t 将以 p 的概率保持为 t,而以 $1-p$ 的概率变为 $\bar{t}$,即 $01\xrightarrow{p}01,01\xrightarrow{1-p}10$,而 01 变为 11,00 的概率均为 0.

表 2 给出了 LS-PPFM 项集同步随机化的例子,其中,项 A、项 B 同步随机化,其他列独立随机化,随机化参数 p=0.8,其对应的变换概率矩阵见表 3.作为对比,表 4、表 5 分别给出了现有 mask 方法随机化的例子及其对应的变换概率矩阵.可以看出,LS-PPFM 随机化时对项 A、项 B 同时取反或同时保持不变;mask 方法随机化时,项 A、项 B 是独立变换的.表 5 中,mask 方法对应的变换概率矩阵元素 $p_{ij}=p^r(1-p)^{k-r}(0\leqslant r\leqslant k)$,$r$ 为 i 和 j 对应的 k 位二进制 0-1 串中值相同的位数;而表 3 中,LS-PPFM 对应的变换概率矩阵元素 p_{ij} 由于同步项的引入变得复杂,假设 $r',\bar{r}'$ 分别为 i 和 j 对应的 k 位二进制 0-1 串中非同步项位串中值相同、不相同的位数,则有:

$$p_{ij}=\begin{cases}p^{r'+1}(1-p)^{\bar{r}'}, & i\text{和}j\text{对应的二进制0-1序列中,同步项完全相同}\\ p^{r'}(1-p)^{\bar{r}'+1}, & i\text{和}j\text{对应的二进制0-1序列中,同步项完全相反}\\ 0, & i\text{和}j\text{对应的二进制0-1序列中,同步项有的相同,有的相反}\end{cases}.$$

Table 2 Synchronized randomization of item *A*, *B* with LS-PPFM method

表 2 LS-PPFM 方法对项 *A*、项 *B* 绑定同步随机化

TID	Items	*p*=0.8					Items	*p*=0.8			
		A	*B*	*C*	*D*			*A*	*B*	*C*	*D*
1	*AC*	1	0	1	0		*BC*	0	1	1	0
2	*AB*	1	1	0	0		*AB*	1	1	0	0
3	*CD*	0	0	1	1	Synchronized randomization of binding *A* and *B* →	*C*	0	0	1	0
4	*BD*	0	1	0	1		*AD*	1	0	0	1
5	*ABCD*	1	1	1	1		*ABD*	1	1	0	1
6	*D*	0	0	0	1		*D*	0	0	0	1
7	*AB*	1	1	0	0		*AB*	1	1	0	0
8	*ABD*	1	1	0	1		*ABD*	1	1	0	1
9	*BD*	0	1	0	1		*BCD*	0	1	1	1
10	*BCD*	0	1	1	1		*BC*	0	1	1	0

Table 3 Transition matrix *P* of LS-PPFM method (synchronized randomization of item *A*, *B*)

表 3 LS-PPFM 方法中项 *A*、项 *B* 同步随机化变换概率矩阵 *P*

	00 00	00 01	…	0100	…	11 11
00 00	p^3	$p^2(1-p)$	…	0	…	$(1-p)^3$
00 01	$p^2(1-p)$	p^3	…	0	…	$p(1-p)^2$
…	…	…	…	…	…	…
11 11	$(1-p)^3$	$p(1-p)^2$	…	0	…	p^3

Table 4 Independent randomization of all items with mask method

表 4 mask 方法对所有项独立随机化

TID	Items	*p*=0.8					Items	*p*=0.8			
		A	*B*	*C*	*D*			*A*	*B*	*C*	*D*
1	*AC*	1	0	1	0		*C*	0	0	1	0
2	*AB*	1	1	0	0		*A*	1	0	0	0
3	*CD*	0	0	1	1	Independent randomization of all items →	*C*	0	0	1	0
4	*BD*	0	1	0	1		*ABD*	1	1	0	1
5	*ABCD*	1	1	1	1		*ABD*	1	1	0	1
6	*D*	0	0	0	1		*D*	0	0	0	1
7	*AB*	1	1	0	0		*A*	1	0	0	0
8	*ABD*	1	1	0	1		*ABD*	1	1	0	1
9	*BD*	0	1	0	1		*BCD*	0	1	1	1
10	*BCD*	0	1	1	1		*BC*	0	1	1	0

Table 5 Transition matrix P of mask method (independent randomization of all items)

表 5 mask 方法所有项独立随机化变换概率矩阵 P

	0000	0001	…	0100	…	1111
0000	p^4	$p^3(1-p)$	…	$p^3(1-p)$	…	$(1-p)^4$
0001	$p^3(1-p)$	p^4	…	$p^2(1-p)^2$	…	$(1-p)^3p$
…	…	…	…	…	…	…
1111	$(1-p)^4$	$(1-p)^3p$	…	$(1-p)^3p$	…	p^4

2.3 支持度重构

LS-PPFM 在完成项集同步随机化后,需要进行支持度重构,有以下 3 种方法:

(1) 公式直接求解法

借鉴独立随机化 mask 方法的支持度重构原理,假定数据集 D,其事务由 $I=\{I_1,I_2,...,I_m\}$中的项组成,$A=\{I_1,I_2,\ldots,I_k\}$为 k-项集,C_j, C'_j 为分别 A 的第 j 个子集在 $D(I_1,\ldots,I_k)$(D 中只含 A 中的项构成的数据集)和 $D'(I_1,\ldots,I_k)$中的净计数(即 $D(I_1,\ldots,I_k), D'(I_1,\ldots,I_k)$中恰等于 C_j 和 C'_j 的事务数).则向量 $\overrightarrow{C_A}=[C_0,C_1,...,C_{2^k-1}]^T$ 和 $\overrightarrow{C'_A}=[C'_0, C'_1,...,C'_{2^k-1}]^T$ 的期望值存在如下关系: $E(\overrightarrow{C'_A})=P_k\cdot\overrightarrow{C_A}, \overrightarrow{C_A}=P_k^{-1}\cdot E(\overrightarrow{C'_A})$. 其中,$P_k=[p_{ij}]$为随机化参数 p 构成的 $2^k\times 2^k$ 随机化变换概率矩阵,p_{ij} 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.

实际中,用从 D'中测得的 $\overrightarrow{C'_A}$ 近似代替 $E(\overrightarrow{C'_A})$,即得到对 $\overrightarrow{C_A}$ 的估计值 $\overrightarrow{\widehat{C_A}}=P_k^{-1}\cdot\overrightarrow{C'_A}$. 而向量 $\overrightarrow{\widehat{C_A}}$ 最后一个元素 $\widehat{C_A}=\widehat{C_{2^k-1}}$ 为 A 最后一个子集(A 本身)在 $D(I_1,\ldots,I_k)$中的净计数,恰等于 A 在 D 中的支持计数$\widehat{S_A}$.若 $P_k^{-1}=[a_{ij}]$,则

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

式中,$a_{2^k-1,j}(j=0,1,\ldots,2^k-1)$为矩阵 P_k^{-1} 中的最后一行元素.因此,只要构造出如表 3 所示的概率矩阵 P,就可以求得任意项集的重构支持计数和支持度.求取 k-项集支持度相当于求解 2^k 个线性方程,而求解整个项集空间的 2^m 个项集的支持度相当于求解的线性方程个数为 $\sum_{A\subseteq I}2^{|A|}=3^{|I|}=3^m$.

(2) 净计数、支持计数转换法

求解整个项集空间的 2^m 个项集支持度的较为快捷的方法是:先根据 $\overrightarrow{\widehat{C_I}}=P_m^{-1}\cdot\overrightarrow{C'_I}$ 重构出 2^m 个项集在 D 中的净计数,其中,$\overrightarrow{\widehat{C_I}}$ 为向量 $\overrightarrow{C_I}=[C_0,C_1,...,C_{2^m-1}]^T$ 的重构估计值,C_i 为 I 的第 i 个子集在 D 中的净计数;然后,由项集的支持计数与净计数的关系 $\overrightarrow{\widehat{S_I}}=T\cdot\overrightarrow{\widehat{C_I}}$ [5]一次性地求得此 2^m 个项集的重构支持数,进而得到其支持度.其中,$\overrightarrow{\widehat{S_I}}=[\widehat{S_0},\widehat{S_1},...,\widehat{S_{2^m-1}}]^T$ 为 I 的 2^m 个子集的重构支持计数构成的向量;$T=[t_{ij}]$为 $2^m\times 2^m$ 矩阵,当 $f_i\subseteq f_j$ 时,$t_{ij}=1$,其余元素均为 0.这里,行坐标 f_i 和列坐标 f_j 为 I 的子集.式 $\overrightarrow{\widehat{S_I}}=T\cdot\overrightarrow{\widehat{C_I}}$ 的依据是:对任意的项集,其支持计数等于其所有超集的净计数之和,即 $S_i=\sum_{i\subseteq j\subseteq I}C_{j,I\setminus j}$. 这是容易理解的,比如对于表 1 包含的事务数据集,项集$\{AB\}$的支持计数 $S_{\{AB\}}=C_{\{AB\}}+C_{\{ABC\}}+C_{\{ABD\}}+C_{\{ABCD\}}=2+0+1+1=4$.净计数 C_X 的含义为,事务数据集中恰好等于 X 的事务数.该方法相当于求解的方程个数为 $2^m+2^m=2\times 2^m$.

(3) 递归公式法

mask 方法使用唯一的随机化参数 p,可看作是文献[5]每个列随机化参数都相等时的特殊情况.本文先参照文献[5]中 RE 方法的支持计数重构递归公式(4),导出随机化 mask 方法对 k-项集 A 的支持计数重构递归公式为

$$\widehat{S_A}=\frac{S'_A-\sum_{f\subset A}(2p-1)^{|f|}(1-p)^{(|A|-|f|)}\widehat{S_f}}{(2p-1)^{|A|}},\ \widehat{S_\varnothing}=|D| \tag{1}$$

虽然同步随机化的引入导致随机化复杂,难以导出像公式(1)那样的项集支持计数重构递归公式,但有两类

项集的支持,计数重构公式易推导,一类是包含且仅包含同步项的项集,另一类是不包含任何同步项的非同步项集.

当 A 为仅包含同步项项集时,A 在 D 和 D'中的支持计数 S_A, S'_A 及 $\overline{A}$ 在 D 和 D'中的支持计数 $S_{\overline{A}}, S'_{\overline{A}}$,理论上满足下式:

$$\begin{cases} S'_A = pS_A + (1-p)S_{\overline{A}} \\ S'_{\overline{A}} = pS_{\overline{A}} + (1-p)S_A \end{cases} \tag{2}$$

用重构支持计数 $\widehat{S_A}, \widehat{S_{\overline{A}}}$ 近似替代公式(2)中的原始支持计数 $S_A, S_{\overline{A}}$,即得到:

$$\widehat{S_A} = \frac{pS'_A - (1-p)S'_{\overline{A}}}{2p-1} \tag{3}$$

其中, $S'_{\overline{A}}$ 代表 D'中 A 所包含的项取值全为 0 的事务个数.

当 A 为非同步项组成的项集时,其支持计数重构公式满足公式(1).

以上 3 种方法,方法(3)相对直接和简单,但只能求出所有项均为同步项或均为非同步项的项集的支持度;而方法(1)和方法(2)均可求出任意项集的重构支持度.其中,方法(1)适合于求单个项集的重构支持度,而方法(2)适合于批量求出所有项集的重构支持度.本文在实验中采用方法(2)进行支持度重构,以便一次性对所有频繁项集求解.

2.4 隐私性分析

(1) 独立随机化 mask 算法

文献[3]中,独立随机化 mask 算法定义的隐私保护度公式 $privacy=1-R(p)$,其中,$R(p)=aR_1(p)+(1-a)R_0(p)$,a 为分配给 1 和 0 的隐私保护权重.a 越大,表示数据库中对于 1 值的保护要求越高,而对于 0 值的保护要求越低.通常对于商场购物篮数据,需要保护的是顾客买了某个商品,即 1 值,而对于顾客没有购买某个商品并不实施保护,这时可以设置 $a=1$.此时,隐私保护度 $privacy=1-R_1(p)$.式 $R(p)$中,$R_1(p), R_0(p)$分别表示原始数据库中的 1,0 能够从随机化后的数据库中被还原出来的概率.$R_1(p)$计算公式如下:

$$R_1(p, s_i) = \frac{p^2 s_i}{(1-p)(1-s_i) + p\, s_i} + \frac{(1-p)^2 s_i}{p(1-s_i) + (1-p)\, s_i} \tag{4}$$

$$R_1(p) = \frac{\sum_i s_i R_1(p, s_i)}{\sum_i s_i} \tag{5}$$

其中,s_i 为项 i 的支持度,具体推导见文献[3].

(2) 项集同步随机化

下面推导项集同步随机化的隐私保护度计算公式.由于项集同步随机化模型对于每个单元来讲仍可看作是以 p 的概率保持不变、以 $1-p$ 的概率取反,假设 i 代表非同步项,j 代表同步项.

对于非同步项 i,其对应列中的 1 被还原的概率满足公式(4),根据公式(4)易知,$R_1(p,s_i)$与项 i 的支持度 s_i 成正比.说明对于非同步项,支持度越大,其 1 值被还原出来的概率就越大.

对于同步项 j,各个项对应列中的 1 能够被还原的概率取决于能够被还原的概率最大的项,即支持度最大的项.假定同步项中支持度最大的项的支持度分别为 $s_{\max}$,则同步项 j 对应列中的 1 能够被还原出来的概率为

$$R_1(p, s_j) = \frac{p^2 s_{\max}}{(1-p)(1-s_{\max}) + p\, s_{\max}} + \frac{(1-p)^2 s_{\max}}{p(1-s_{\max}) + (1-p)\, s_{\max}}.$$

考虑到所有的项,1 被还原的总概率为

$$R_1(p) = \frac{\sum_i s_i R_1(p, s_i) + \sum_j s_j R_1(p, s_j)}{\sum_i s_i + \sum_j s_j} \tag{6}$$

容易证明,对同样的数据,公式(6)≥公式(5).这表明同步随机化的原值被还原的概率大于等于独立随机化,即同步随机化的隐私保护度小于等于独立随机化.并且,随着参与同步的项增多以及同步项与同步项中支持度

最大的项的支持度差距增大,公式(6)和公式(5)的差距越大,同步随机化与独立随机化的隐私保护度差距就越大.这从理论上说明,为保持一定的隐私保护度,同步随机化时,要有选择地、针对性地选取某些项同步,不能全同步、盲目和过度同步.

3 实验评价

下面结合实验对 LS-PPFM 作分析评价,实验目的是验证项集同步(意同绑定,取名 bind)随机化模型是否比项独立随机化模型 mask 具有更高的挖掘结果准确性,分析不同随机化方式的差异.其中,不同随机化方式是指:

(1) 所有项独立随机化,记为 mask;

(2) 所有项同步随机化,记为 bindall;

(3) 支持度指导的项集同步随机化,记为 bindsupp.

3.1 实验方法

- 第 1 步,对样例数据进行学习,得到强关联项.实验采用 IBM Almaden 生成器生成的数据集 D 作为原始数据,生成器参数为 T=3,I=4,|D|=100K,N=10,即事务平均长度为 3,频繁项集平均长度为 4,总事务数为 100K,总项数为 10.本文利用 D 的前 30%作为不需要保护的样例数据 D_0.选取 30%主要源于本文开始提到的 AT&T 在 1999 年对隐私态度的调查中,有近 30%的人不关注隐私.具体操作如下:
  (1) 设置最小支持度阈值 min_sup=0.05%,对 D_0 挖掘,得到频繁 4-项集集合 F_4(D_0 中,频繁 4-项集居多);
  (2) 从 F_4 中挑选出支持度最高的频繁 4-项集 $supp$.
- 第 2 步,对 D 进行项集同步随机化.设置随机化参数 p 为 0.84,分别生成 D'_{mask},$D'_{bindall}$,$D'_{bindsupp}$ 这 3 个随机化数据集.其中,D'_{mask} 为对 D 中的所有项独立随机化生成的数据集,$D'_{bindall}$ 为对 D 中的所有 10 个项同时作绑定生成的数据集,$D'_{bindsupp}$ 为对第 1 步特征学习得到的 4-项集 $supp$ 中的 4 个项作绑定生成的随机化数据集.
- 第 3 步,挖掘随机化数据集.针对多个不同的支持度阈值 min_sup=0.05%,0.1%,0.2%,0.4%,0.6%,0.8%,1%,运用项集随机化支持度重构方法对第 2 步生成的 3 个随机化数据集 D'_{mask},$D'_{bindall}$,$D'_{bindsupp}$ 进行挖掘,记录每次挖掘得到的所有频繁项集及其支持数.
- 第 4 步,计算分析误差.计算第 3 步 3 次挖掘结果误差,包括项集支持度误差ρ、项集身份误差θ^-和θ^+.
  (1) 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.
  (2) 项集身份误差θ^-表示频繁项集丢失率,衡量原先频繁而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(错误识别率),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.假定 F 是从原数据挖掘得到的频繁项集集合,$\widehat{F}$ 是从随机化数据运用支持度重构挖掘得到的频繁项集集合,s_f 和 $\widehat{s_f}$ 分别表示频繁项集 f 实际的支持度和重构后的支持度,则所有频繁项集的平均支持度相对误差百分比、频繁项集丢失率θ^-、增加率θ^+分别为

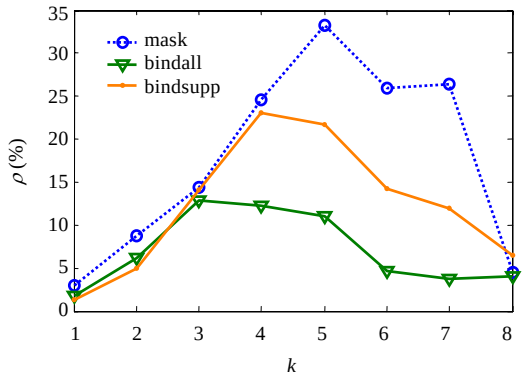
$$\rho=\frac{1}{|F|}\sum_{f\in F}\frac{|\widehat{s_f}-s_f|}{s_f},$$

$$\theta^-=\frac{|F-\widehat{F}|}{|F|},$$

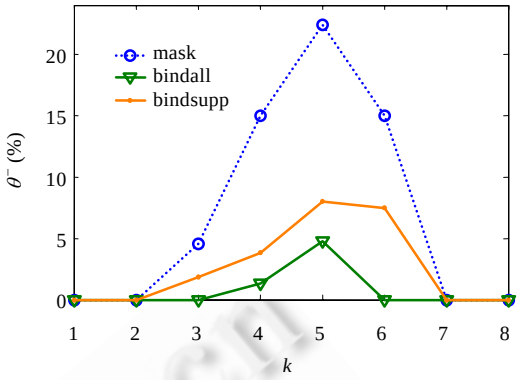
$$\theta^+=\frac{|\widehat{F}-F|}{|F|}.$$

3.2 实验结果

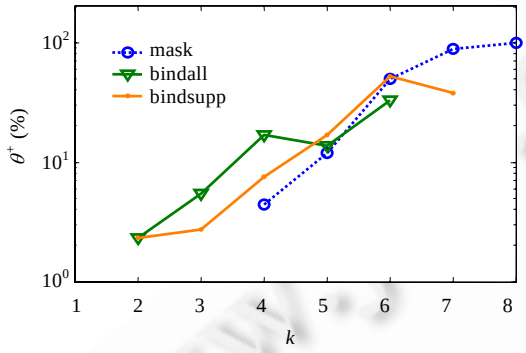
图 3(a)~图 3(f)给出了实验结果.



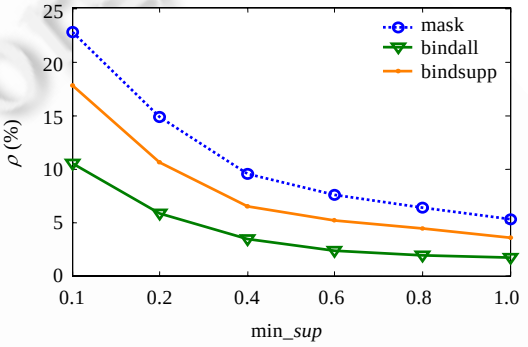
(a) Support error ρ vs. itemset length k
(a) 支持度误差ρ随项集长度 k 的变化



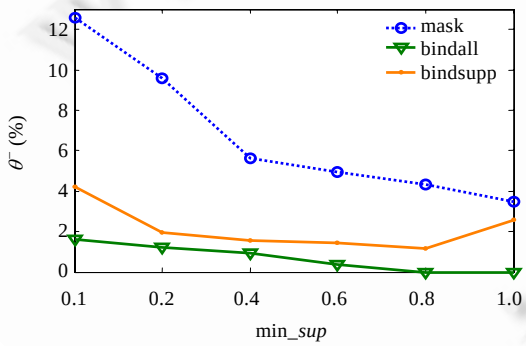
(b) Frequent itemsets lost ratio θ^- vs. itemset length k
(b) 频繁项集丢失率θ^-随项集长度 k 的变化



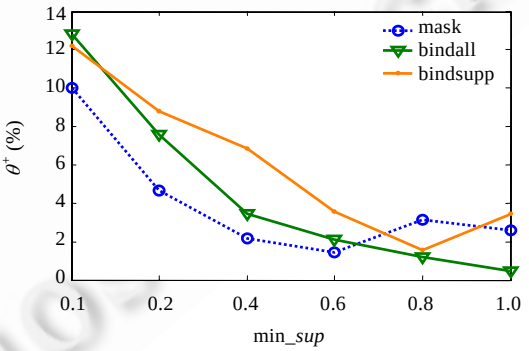
(c) Frequent itemsets added ratio θ^+ vs. itemset length k
(c) 频繁项集增加率θ^+随项集长度 k 的变化



(d) Support error ρ vs. min_*sup*
(d) 支持度误差ρ随 min_*sup* 的变化



(e) Frequent itemsets lost ratio θ^- vs. min_*sup*
(e) 频繁项集丢失率θ^-随 min_*sup* 的变化



(f) Frequent itemsets added ratio θ^+ vs. min_*sup*
(f) 频繁项集增加率θ^+随 min_*sup* 的变化

Fig.3 Experiment error of mask, bindall and bindsupp
图 3 mask,bindall 和 bindsupp 的实验误差

3.3 结果分析

(1) 误差随项集长度的变化曲线

图 3(a)~图 3(c)分别显示了当最小支持度阈值 min_*sup*=0.1%时,mask 方法和项集同步随机化的平均支持度

相对误差ρ、项集身份误差θ^-和θ^+随频繁项集长度 k 的变化曲线.

横向对比图 3(a)中各曲线可以发现,整体上,所有项集同步随机化 bindall 的误差最小,支持度指导的项集同步随机化 bindsupp 次之,而项独立随机化 mask 方法的误差最大,尤其是对占 F 数量较多的 3-项集、4-项集和 5-项集.即绝大多数情况下,平均支持度相对误差大小遵从 bindall<bindsupp<mask 的规律.其中,bindall 对所有项作绑定和同步处理,所以其能尽可能地保持所有长频繁项集的支持度;bindsupp 以样例数据中支持度最高的频繁项集作为同步项,所以其能尽可能地保持该频繁项集中各项所组成的项集的支持度.

另外,理论上,由项集支持度重构误差会从 1-项集的误差递归渗透和传导到 2-项集、3-项集直至最长的频繁项集,即支持度重构误差理应随频繁项集长度的变长而增大.而观察图 3(a)可发现,平均支持度误差ρ大体呈先上升、后下降趋势.主要原因在于,频繁-5 项集以后的频繁-6 项集、频繁-7 项集、频繁-8 项集个数非常少(实验中频繁-4 项集、5-项集个数分别为 160,125,而频繁 7-项集、8-项集个数分别只有 8 和 1),平均支持度误差受个别项集的误差影响较大.图中在 k=6 处出现拐点,也是由于频繁-7 项集个数太少无法反映整体规律造成的.

观察图 3(b)和图 3(c)频繁项集身份误差可知:(1) 对于频繁项集丢失率θ^-,几条曲线的大小顺序关系大致遵从 bindall<bindsupp<mask;(2) 对于频繁项集增加率θ^+,则呈现与θ^-恰好相反的规律,大致遵从 bindall>bindsupp>mask.这说明项集同步随机化与项独立随机化比,会增加非频繁项集变为频繁项集的机会,但会减少频繁项集丢失的机会.这是因为通常数据集中短事务多、长事务少,即 0 的个数较多,1 的个数较少.绑定多个项同步随机化,增加了事务中多个为 0 的项同时变为多个为 1 的项的机会,减少了多个为 1 的项变为 0 的机会.即绑定项越多,频繁项集增加的机会就越多,丢失的机会也就越少.实验中,bindall 绑定项最多,所以其θ^+最大、θ^-最小;而 mask 各项不绑定均独立随机化,所以其θ^+最小、θ^-最大.

(2) 误差随支持度阈值的变化曲线

图 3(d)~图 3(f)分别给出了所有频繁项集(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、项集身份误差θ^-和θ^+随最小支持度阈值 min_*sup* 的变化曲线.

横向对比图 3(d)中的各曲线可知,误差大小关系与图 3(a)基本一致,遵从 bindall<bindsupp<mask.说明所有项集同步随机化 bindall 得到的挖掘结果最准确,其次是支持度指导的项集同步随机化 bindsupp.

另外,理论上可推导相对误差ρ与支持度的平方根成反比.而由于支持度阈值 min_*sup* 越大,各频繁项集的支持度也越大,因此ρ随着支持度阈值 min_*sup* 的增大而减小.这与由图 3(d)所观测到的现象是一致的.

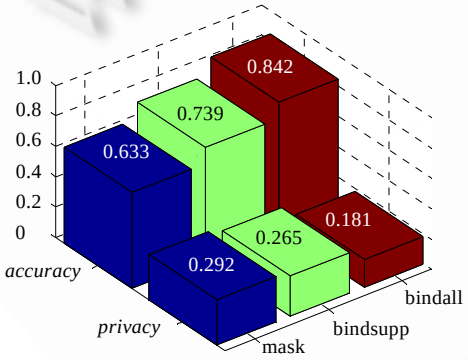


Fig.4 *privacy* and *accuracy* comparison of mask, bindall and bindsupp

图 4 mask,bindall 和 bindsupp 的 *privacy* 和 *accuracy* 对比

观察图 3(e)和图 3(f)可以看出:(1) 项集身份误差大体上随着 min_*sup* 的增大而减小,这与图 3(d)的规律是一致的.因为项集身份随 min_*sup* 的变化,本质上取决于项集支持度随 min_*sup* 的变化.(2) 相对于绑定随机化方式,项独立随机化 mask 的频繁项集增加率θ^+最小,而频繁项集丢失率θ^-最大.再次说明了项集绑定同步随机化会增大引入虚假频繁项集的机会,而减小频繁项集丢失的机会.

(3) 同步与非同步隐私保护程度、挖掘结果准确度的对比

图 4 为实验中当 min_*sup*=0.05%时,独立随机化 mask、半同步随机化 bindsupp 和所有项全部同步随机化 bindall 的隐私保护度 *privacy* 和挖掘结果准确度 *accuracy* 的对比情况.其中,*privacy*=1−$R_1(p)$,mask 的 $R_1(p)$ 由公式(5)求出,bindsupp,bindall 的 $R_1(p)$由公式(6)求出;*accuracy*=1−ρ,其中,ρ取所有频繁项集的平均支持度的相对误差.

由图 4 可以看出,从非同步随机化 mask 到半同步随机化 bindsupp,再到全部项同步随机化 bindall,频繁项集挖掘结果准确度逐步提高,而整体隐私保护度略有下降.这表明从非同步随机化到同步随机化,挖掘结果准确度

可以提高,但会牺牲一定程度的隐私保护性.但是,这种牺牲相对准确度的提高是很小和值得的,尤其是从 mask 到 bindsupp,隐私保护度下降了约 2 个百分点,但准确度却提高了约 10 个百分点.而从 bindsupp 到 bindall,隐私保护度则下降了约 8 个百分点(比 mask 到 bindsupp 下降的 2 个百分点大很多,隐私保护性牺牲程度较大),准确度上升了约 10 个百分点.以上事实说明,若想提高隐私保护频繁模式挖掘结果的准确度,可用项同步随机化代替项独立随机化,但同步项较多时,隐私保护度也会明显下降.因此,合理选择同步项——使挖掘结果准确度提高,同时不致牺牲太多的隐私保护性,是同步随机化的关键.本文同步随机化前的样例学习正是出于合理选择同步项的目的,否则,究竟如何同步、选择哪些项同步,会因缺乏依据而变得盲目鲁莽.这正是本文同步随机化需基于样例学习进行指导的思想根源.通过合理选取同步项,可在牺牲非常小的隐私保护性的情况下,大幅度提升挖掘准确度.

4 总结与展望

本文提出了一种基于样例学习和项集同步随机化的隐私保护频繁模式挖掘方法 LS-PPFM.该方法先对不需要隐私保护的个体数据进行学习,得到样例数据中蕴涵的强关联项;然后在对数据随机化时,将强关联项绑定在一起作同步随机化变换,以图保持项与项之间的关联性.实验结果表明:(1) 相对于项独立随机化,所有项绑定同步随机化会显著提高频繁项集支持度重构结果的准确性,但会牺牲较大程度的隐私保护性;(2) 相对于项独立随机化,支持度指导的项集同步随机化通过选取样例数据中支持度高的项集项作为强关联同步项,能够在牺牲非常小的隐私保护性情况下(比起所有项同步随机化,这种牺牲是非常小和值得的,尤其是当所有项数很多,而频繁模式项数较少时),显著提高频繁模式挖掘结果的准确性.

本文贡献在于:(1) 基于部分人群对隐私不关注的事实,提出了基于部分真实样例学习的、有指导的同步随机化数据保护方式;(2) 给出了同步随机化的3种支持度重构方法,理论分析得出了同步随机化的隐私保护度公式;(3) 实验分析比较了所有项独立随机化、所有项同步随机化、有指导的同步随机化这 3 种方法的效果.

LS-PPFM 可用于提高敏感性问题调查统计结果的准确性.未来的工作可从以下方面开展:(1) 探索其他的项集同步随机化指导方式,比如用强相关系数指导替代现有的支持度指导.(2) 探索既包含同步项又包含非同步项的项集的支持度重构公式.(3) 探索同步项的选择策略,究竟选多少组项同步,需结合所期待的挖掘结果准确性和隐私保护性给出进一步指导.本文实验对数据同步随机化时,只选了 1 组强关联项,实际上当项个数较多时,可选多组强关联项,这会进一步提高挖掘结果准确性,同时又不致牺牲太多的隐私保护性.(4) 进一步使用真实数据作实验.另外,可结合文献[16],根据要求的挖掘结果准确性,对究竟选取多少真实样本进行定量指导.

References:

[1] Vaidya J, Clifton C. Privacy preserving association rule mining in vertically partitioned data. In: Hand D, Keim D, Ng R, eds. Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (SIGKDD 2002). New York: ACM Press, 2002. 639−644. [doi:10.1145/775047.775142]

[2] Kantarcioglu M, Clifton C. Privacy-Preserving distributed mining of association rules on horizontally partitioned data. IEEE Trans. on Knowledge and Data Engineering (TKDE), 2004,16(9):1026−1037. [doi: 10.1109/TKDE.2004.45]

[3] Rizvi SJ, Haritsa JR. Maintaining data privacy in association rule mining. In: Bernstein PA, Ioannidis YE, Ramakrishnan R, Papadias D, eds. Proc. of the 28th Int'l Conf. on Very Large Data Bases (VLDB 2002). San Francisco: Morgan Kaufmann Publishers, 2002. 682−693.

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

[5] Xia Y, Yang Y, Chi Y. Mining association rules with non-uniform privacy concerns. In: Das G, Liu B, Yu PS, eds. Proc. of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2004). New York: ACM Press, 2004. 27−34. [doi: 10.1145/1008694.1008699]

[6] Agrawal S, Haritsa J. A framework for high-accuracy privacy-preserving mining. In: Kitagawa H, Ishikawa Y, Morishima A, Takayama T, eds. Proc. of the 21st IEEE Int'l Conf. on Data Engineering (ICDE 2005). Washington: IEEE Computer Society, 2005.

193−204. [doi: 10.1109/ICDE.2005.8]

[7] Zhang P, Tong YH, Tang SW, Yang DQ, Ma XL. An effective method for privacy preserving association rule mining. Journal of Software, 2006,17(8):1764−1774 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/1764.htm [doi: 10.1360/jos171764]

[8] Xu CF, Wang JL. An efficient incremental algorithm for frequent itemsets mining in distorted databases with granular computing. In: Nishida T, ed. Proc. of the 5th IEEE/WIC/ACM Int'l Conf. on Web Intelligence (WI 2006). Washington: IEEE Computer Society, 2006. 913− 918. [doi: 10.1109/WI.2006.37]

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

[10] Huang ZL, Du WL, Teng ZX. Searching for better randomized response schemes for privacy-preserving data mining. In: Kok JN, Koronacki J, Mantaras RL, Matwin S, Mladenic D, Skowron A, eds. Proc. of the 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007). LNCS 4702, Heidelberg: Springer-Verlag, 2007. 487−497. [doi: 10.1007/978-3-540-74976-9_50]

[11] Huang ZL, Du WL. OptRR: Optimizing randomized response schemes for privacy-preserving data mining. In: Alonso G, Blakeley J, Chen A, eds. Proc. of the 24th IEEE Int'l Conf. on Data Engineering (ICDE 2008). Washington: IEEE Computer Society, 2008. [doi: 10.1109/ICDE.2008.4497479]

[12] 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]

[13] Du WL, Zhan Z. Using randomized response techniques for privacy-preserving data mining. In: Getoor L, Senator TE, Domingos P, Faloutsos C, eds. Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (SIGKDD 2003). New York: ACM Press, 2003. 505−510. [doi: 10.1145/956750.956810]

[14] Gao AQ, Diao LH. Privacy preservation for attribute order sensitive workload in medical data publishing. Journal of Software, 2009,20:314−320 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/09036.htm

[15] Zhou SG, Li F, Tao YF, Xiao XK. Privacy preservation in database applications: A survey. Chinese Journal of Computers, 2009, 32(5):847−861 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2009.00847]

[16] Jacquemont S, Jacquenet F, Sebban M. A lower bound on the sample size needed to perform a significant frequent pattern mining task. Pattern Recondition Letters, 2009,30(11):960−967. [doi: 10.1016/j.patrec.2009.05.002]

附中文参考文献:

[7] 张鹏,童云海,唐世渭,杨冬青,马秀莉.一种有效的隐私保护关联规则挖掘方法.软件学报,2006,17(8):1764−1774. http://www.jos.org.cn/1000-9825/17/1764.htm [doi: 10.1360/jos171764]

[14] 高爱强,刁麓弘.医疗数据发布中属性顺序敏感的隐私保护方法.软件学报,2009,20:314−320. http://www.jos.org.cn/1000-9825/09036.htm

[15] 周水庚,李丰,陶宇飞,肖小奎.面向数据库应用的隐私保护研究综述.计算机学报,2009,32(5):847−861. [doi: 10.3724/SP.J.1016.2009.00847]

郭宇红(1979－),女,河南洛阳人,博士,讲师,主要研究领域为数据挖掘,数据仓库.

唐世渭(1939－),男,教授,博士生导师,CCF高级会员,主要研究领域为数据库,信息系统.

童云海(1971－),男,博士,副教授,主要研究领域为数据挖掘,联机分析处理.

吴冷冬(1982－),男,博士生,主要研究领域为人工智能,数据库系统.