软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(6):1457−1472 [doi: 10.13328/j.cnki.jos.004576]

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

一种分布式事务数据的差分隐私发布策略*

欧阳佳, 印 鉴, 刘少鹏

(中山大学 计算科学系,广东 广州 510275)

通讯作者: 印鉴, E-mail: issjyin@mail.sysu.edu.cn

摘 要: 目前隐私保护的事务数据发布研究多是基于集中式结构.针对分布式结构下事务数据发布问题,为保护数据隐私,同时最大化数据效用,提出一种满足差分隐私约束的发布策略.首先,将结果效用性优化与差分隐私约束相结合,构建分布式非线性规划模型.然后,基于全局与局部数据设计两种解决方案安全求解该分布式模型.理论分析与实验结果均表明,所提出的发布策略是安全的且满足差分隐私要求,具有很好的实用性.

关键词: 隐私保护;差分隐私;分布式结构;事务数据发布;优化

中图法分类号: TP311

中文引用格式: 欧阳佳,印鉴,刘少鹏.一种分布式事务数据的差分隐私发布策略.软件学报,2015,26(6):1457−1472. http://www.jos.org.cn/1000-9825/4576.htm

英文引用格式: Ouyang J, Yin J, Liu SP. Differential privacy publishing strategy for distributed transaction data. Ruan Jian Xue Bao/Journal of Software, 2015,26(6):1457−1472 (in Chinese). http://www.jos.org.cn/1000-9825/4576.htm

Differential Privacy Publishing Strategy for Distributed Transaction Data

OUYANG Jia, YIN Jian, LIU Shao-Peng

(Department of Computer Science, Sun Yat-Set University, Guangzhou 510275, China)

Abstract: In the research of privacy preserving transaction data publishing, the existing methods are always designed for the centralized structure. The paper proposes a differential privacy publishing strategy to protect data privacy and maximize utility of the output data in the distributed environment. The new method combines the utility optimization of the output with differential privacy constraints and builds a distributed nonlinear programming model. Furthermore, two solutions based on global and local data respectively are designed to solve the distributed model securely. As shown in the theoretical analysis and the experimental results, the publishing strategy can achieve significant improvements in terms of privacy, security, and applicability.

Key words: privacy preservation; differential privacy; distributed environment; transaction data publishing; optimization

信息时代,互联网成为人们不可缺少的部分.人们在互联网上留下大量历史记录,如购物记录等事务信息,这些信息记录被公司或机构广泛收集.为挖掘数据中的知识,需要发布这些数据,但同时也会泄露数据中的隐私信息给研究者或恶意用户(称为攻击者)[1].隐私保护数据发布研究的一个关键问题是:不泄露隐私的同时,最大化数据的效用性.

近年来,由于通信技术的日益成熟以及网络通信带宽的不断增加,个体完整的事务信息通常被划分多个子集,每个子集存储在独立的站点上.为提供更好的服务和最大化各自收益,各站点之间共享这些分布式数据以进行数据分析及数据挖掘任务,如关联规则挖掘[2,3]、用户行为预测[4]等.

隐私模型与通信安全是分布式事务数据发布的两个关键问题.

- 一方面,由于如下两个原因,已有的关系数据库匿名模型如 *k*-匿名模型[5]、*l*-diversity[6]等,不能直接应

* 基金项目: 国家自然科学基金(61033010, 61272065, 61472453); 广东省自然科学基金(S2011020001182, S2012010009311); 广东省科技计划(2011B040200007, 2012A010701013)

收稿时间: 2012-12-25; 修改时间: 2013-09-02; 定稿时间: 2014-01-24

用于事务数据[7]:

- 首先,事务数据的准标识符(quasi-identifies)与敏感信息没有明显区别,攻击者易获得个体部分信息,通过已知项组合查询数据集,攻击者能唯一识别个体对应的事务,导致受害者购买记录泄露.显然,预先知道攻击者所有可能已知项的组合是不可能的;
- 其次,由于事务数据的高维与稀疏性,匿名处理后的数据效用性不足[8].尽管已提出众多基于划分的事务数据匿名模型及对应的匿名算法[9−17],如 k-anonymity[9],k^m-anonymity[13,14]等.但由于攻击者背景知识的复杂性及匿名算法过程的确定性,导致新的攻击方式出现[18,19],泄露个人隐私.因此,合理有效的隐私模型是事务数据发布的核心;

- 另一方面,分布式结构下,各站点拥有自己的商业机密与隐私信息,泄露这些信息会造成巨大的经济损失.数据共享过程中,各站点互相不泄露信息给对方,是分布式数据发布的难点.

本文针对分布式结构下事务数据发布的隐私保护问题,基于差分隐私模型,提出一种分布式事务数据发布策略.该策略能很好地保护个人隐私和最大化数据的效用,有效地保证分布式结构下的通信与计算安全.本文的主要贡献如下:

(1) 基于差分隐私模型,设计一种垂直划分数据格式的随机抽样机制$\mathcal{R}$.以不同的概率从每个项中抽取事务,抽样过程满足差分隐私约束,能有效保护个人隐私;

(2) 通过结合差分隐私约束与数据效用性优化,构造分布式非线性规划模型.该模型能得到最优的数据效用.设计两种解决方案安全地求解该模型:第 1 种为全局解决方案(GS),基于全局数据,设计数量积和协议(sum of scalar product protocol)进行全局求解;第 2 种为局部解决方案(LS),各站点基于局部数据单独求解该模型.理论证明,集成各站点抽样数据同样满足差分隐私;

(3) 理论分析与实验证明,该发布策略具有隐私性、效用性和安全性.

1 相关工作

分布式结构下,事务数据发布的隐私保护问题涉及到隐私模型、效用性和分布式数据发布这 3 个方面,本节简要介绍其相关工作.

1.1 隐私模型

近年来,因为数据挖掘的广泛应用,事务数据发布的隐私保护问题已成为研究热点.事务数据发布的隐私模型可分为两类:一种是传统的基于分组的匿名模型,另一种是忽略攻击者背景知识的差分隐私模型.

- 基于分组的匿名模型

传统的基于分组的匿名模型[9−17],根据项的敏感性又可划分为区分敏感项和不区分敏感项两类.

文献[10,15,20]与文献[17]将项分为敏感的(sensitive)或非敏感的(non-sensitive),假设攻击者的背景知识只包含非敏感项,并试图推断敏感项的信息.Ghinita 等人[20]提出一种基于桶(bucketization)的方法,限定推断敏感项的概率不能超过某个阈值,同时为频繁模式挖掘保留项之间的关系.Xu 等人[10]对攻击者的背景知识进行限定,假设攻击者最多拥有 p 个非敏感项,采用全局消除的方式保留了更多的信息,有效地提升了数据的效用性.文献[15]通过保留频繁项集以及对边界的表示对文献[10]的方法进行改进.Cao 等人[17]假设攻击者的背景知识同时包含敏感项和非敏感项,提出ρ-uncertainty 隐私概念,要求包含敏感项的项集的置信度不能超过ρ.但由于事务数据的高维性与稀疏性,导致该类方法的效用性不足.

文献[9,13,14]没有区分项的敏感性,适应性更强.Terrovitis 等人[14]假设攻击者的背景知识最多包含 m 个项,提出一种新的隐私模型 k^m-anonymity,通过自底向上的泛化方式对数据进行匿名.为增强数据效用性,Terrovitis 等人[13]采用局部编码的方式满足 k^m-anonymity 要求.He 等人[9]指出,k^m-anonymity 的隐私保护力度低于 k-anonymity,并基于 k-anonymity 提出一种自顶向下的局部泛化方法.该类方法的效用性有所提高,但匿名过程均是确定性的,一旦攻击者具有丰富的背景知识,将会存在新的隐私泄露问题[18,19].

总的来说,针对事务数据的隐私保护发布问题,由于事务数据的高维性与稀疏性,导致传统的基于分组的匿

名模型不能取得较好的数据效用性.又因为基于分组的匿名模型对攻击者背景知识作过多假设,限制了模型的应用范围,且隐私保护力度不强.而本文采用的差分隐私模型能较好的弥补了上述不足,成为国内外研究热点.

- 差分隐私

差分隐私(differential privacy)[21−23]是一种完全独立于攻击者背景知识和计算能力的强隐私概念,近年来已成为研究热点.它假定攻击者拥有任意的背景知识,无论特定个体记录是否在数据集中,对该数据集的任意计算分析或查询的结果在形式上不可区分.差分隐私随机算法对任意两个邻近数据集进行操作,得到的结果几乎是一致的.目前,差分隐私已被应用于各种不同数据结构的隐私发布[24−27].特别地,Xiao 等人[24]介绍了一种ε-差分隐私发布策略,通过将 Haar 小波变换应用于差分隐私保护中,对小波系数添加噪音,提高了计数查询的准确度,为区间查询提供准确的结果.Hay 等人[25]针对图的度分布估计问题提出一种有效的差分隐私算法,通过引入度的有序约束,有效地降低了度分布估计的错误率.Mcsherry 等人[26]首次将差分隐私概念应用于推荐系统中,为用户行为提供推荐的同时满足差分隐私约束.Chen 等人[27]首次提出集中式结构下事务数据的差分隐私发布机制.总之,差分隐私是一种有效的隐私保护机制,它在能保护隐私的同时,为数据分析保留足够的有用信息.

1.2 效用最大化

数据发布要保护隐私,也要为数据分析提供足够多的信息.因此,数据效用性是数据发布的核心问题.Bayardo 等人[28]针对 *k*-anonymity 模型效用性较差的问题,提出了一种 *k*-anonymity 的优化方法.Lefevre 等人[29]针对高维匿名问题,提出一种简单但非常有效的贪心算法对匿名过程进行优化,得到了高质量的匿名数据.Ghosh 等人[30]提出一种发布统计数据集的效用最大化机制.文献[8]提出搜索日志的效用最大化发布机制,在效用性优化方面取得非常好的效果,但该机制不能直接应用于分布式结构.本文沿用了文献[8]严格保护隐私同时最大化结果效用性的优点,为适应分布式结构的要求,本文进一步提出分布式结构下的事务数据差分隐私发布策略.

1.3 分布式隐私保护数据发布

目前,隐私保护的数据发布研究多是基于集中式结构,随着分布式应用越来越多,分布式结构下的数据发布的隐私保护问题已得到越来越多的关注.分布式结构可分为垂直划分与水平划分.

针对垂直划分结构,Jiang 和 Clifton[31]提出一种分布式的 *k*-匿名框架(*DkA*),实现两个站点数据的安全集成,结果满足 *k*-匿名模型要求.Mohammed 等人[32]提出一种有效的匿名算法安全集成多方数据.Jurczyk 与 Xiong[33]针对垂直划分的数据,安全地从多方集成数据.对于水平划分结构,Mohammed 等人[34]提出一种分布式算法集成水平划分的高维医疗数据.但是他们所采用的隐私模型都是 *k*-匿名或者是 *k*-匿名的扩展,容易受到新的隐私攻击.因此,Dima 等人[35]首次提出安全两方数据发布算法,该算法同时满足差分隐私及安全多方计算的安全定义.目前,已有的分布式数据发布方法多针对关系数据,关于分布式事务数据发布这方面的工作较少.

2 预备知识

2.1 差分隐私

ε-差分隐私(ε-differential privacy)[23]:随机算法Π满足差分隐私约束,如果任意的两个事务数据集 D 和 D',$|D\Delta D'|=1$,对于所有输出数据集 O,下列不等式成立:

$$\Pr[\Pi(D)=O]\leqslant e^{\varepsilon}\Pr[\Pi(D')=O] \tag{1}$$

其中,$|D\Delta D'|=1$ 表示事务数据集 D 和 D'只有一条记录不同,本文称为邻近事务数据集.ε-差分隐私保证 D 中任意事务的改变(添加项或删除项)对算法Π的输出影响不显著.

然而,ε-差分隐私的强隐私性也导致有些应用不能满足差分隐私要求.本文采用 Machanavajjhala 等人[36]提出的概率差分隐私概念,其隐私强度低于ε-差分隐私.

概率差分隐私(probabilistic differential privacy)[36]:随机算法Π满足概率差分隐私要求,如果任意事务数据集 D 的输出空间Ω能划分为不相交的两部分Ω_1与Ω_2,同时满足如下两个条件:

(1) $\Pr[\Pi(D)\in\Omega_1]\leqslant\delta$;

(2) $e^{-\varepsilon}\leqslant\dfrac{\Pr[\Pi(D)=O]}{\Pr[\Pi(D')=O]}\leqslant e^{\varepsilon}$, $|D\Delta D'|=1, \forall O\in\Omega_2$.

概率差分隐私保证随机机制Π以较高的概率($\geqslant 1-\delta$)满足ε-差分隐私.即,输出空间Ω_1 包含所有隐私泄露的结果输出.条件(1)表示概率差分隐私允许一定的隐私泄露,但其概率不超过δ.

2.2 安全模型

安全多方计算与安全数量积协议是求解分布式非线性规划模型的基础,本节简要介绍半诚信模型(semi-honest model)下的安全多方计算及安全数量积协议.

- 安全多方计算(secure multiparty computation,简称 SMC)[37].

半诚信模型中,攻击者遵循协议的同时也会尝试从得到的信息中推断其他信息.协议在半诚信模型中是安全的,如果在计算结束时,参与者除了知道自身和结果,不知道其他任何信息.安全多方计算定义为[37]:在一个互不信任的、独立但有联系的多用户网络中,各用户在相互不泄漏私有信息的情况下合作执行某项可靠的计算任务.该问题由 Yao 在文献[38]中首次提出,并得到广泛的理论研究.

- 安全数量积协议(secure scalar product protocol)[39−41].

2002 年,Vaidya 与 Clifton[41]为解决垂直划分结构下事务数据关联规则挖掘的隐私保护问题,首次提出安全数量积协议.安全数量积技术在面向分布数据的各类隐私保护数据挖掘中应用十分广泛,并起着重要作用.本文以数量积协议为子协议安全地计算两个数量积的和.

数量积安全求解问题可定义如下:设有两方 A 与 B,A 方拥有向量 $\vec{X}=(x_1,...,x_n)$,B 方拥有向量 $\vec{Y}=(y_1,...,y_n)$,需要安全地计算两个向量的数量积: $A\cdot B=\sum_{i=1}^{n}(x_i\cdot y_i)$.

近几年来该问题的研究越来越多,通信耗费和保护隐私程度都有提高.

3 分布式事务数据的差分隐私发布策略

分布式事务数据的差分隐私发布策略的核心是:随机抽样机制$\mathcal{R}$,为使得$\mathcal{R}$满足差分隐私要求,同时最大化结果数据的效用性,通过结合差分约束与效用性目标函数构造分布式非线性规划模型.为安全求解该模型,本文提出两种解决方案:全局解决方案与局部解决方案.在全局解决方案中,为保证各站点间通信与计算安全,设计了两方安全数量积和协议;而局部解决方案中各站点独立完成计算,计算过程中不存在机密信息泄露.

本节首先介绍随机抽样机制$\mathcal{R}$;其次,针对事务数据发布问题推导抽样机制的差分隐私约束;然后介绍分布式非线性规划模型的构造;最后,提出两种解决方案安全地求解该模型.

假设事务数据被垂直划分并存储在站点 A 与站点 B 上,每个站点包含一组共同的个体事务记录,但每条记录只包含部分属性,每个事务分配一个唯一标识 ID,完整的事务数据集可通过事务 ID 连接重构.设站点 A 拥有项 a,b,c,站点 B 拥有项 d,e,见表 1,Count 表示项的支持计数.在垂直划分的分布式结构下,随机抽样算法$\mathcal{R}$根据各站点所拥有的每个项,从事务数据集中抽取事务 ID,并保证抽样过程满足差分隐私要求.

Table 1 Vertically distributed transaction data

表 1 垂直划分的事务数据

ID	站点 A				ID	站点 B	
	a	b	c			d	e
001	1	0	1		001	1	1
002	1	0	1		002	0	1
…	...	...	...	...	...	...	...
049	1	1	0		049	1	0
050	1	1	1		050	0	1
Count	30	20	40		Count	48	45

3.1 随机抽样机制

随机抽样机制$\mathcal{R}$采用 x 个样本无放回简单随机抽样的方法.首先,根据隐私模型约束及效用性目标函数,计算各站点每个项的最优支持计数 $item.x$;然后,各站点根据每个项 $item$ 依次抽取 $item.x$ 个事务,抽样概率为 $item.x/item.count$,其中,$item.count$ 为项 $item$ 的已知支持计数;然后,各站点通过事务 ID 连接得到结果数据集 O.对每个项的最优支持计数添加约束,使抽样过程满足差分隐私约束,同时最优化效用性目标函数.

例如,考虑表 1 的事务数据集,第 1 步得到每个项的最优计数,分别为 13,9,12,22,17.第 2 步各站点依次从项中抽取包含该项的事务,如,项 a 的最优计数为 13,则从原来的 30 个包含 a 事务中以 13/30 的概率无放回抽取 13 个事务.最后,各站点通过事务 ID 连接(⊕)得到集成事务数据集.

算法 1. 随机抽样机制.

输入:站点 A 与站点 B 事务数据集 D_A 与 D_B,$D=D_A\oplus D_B$;差分隐私参数(ε,δ);

输出:结果数据集 O.

1. 计算站点 A 与站点 B 所拥有项的最优支持计数.
   1.1. 结合效用目标函数与差分隐私约束,构造分布式非线性规划模型,使抽样过程满足差分隐私约束,求解得到项的最优支持计数.
   1.2. 提出两种解决方案求解该分布式非线性规划模型:一种基于全局数据的全局解决方案(GS),该方案采用本文的提出两方安全数量积和协议进行安全通信与计算;另一种基于局部数据的局部解决方案(LS).
2. 生成结果数据集 $O=O_A\oplus O_B$.
   2.1. 各站点分别对每个项抽取包含该项的 x 个事务,站点 A 生成数据集 O_A,站点 B 生成数据集 O_B;
   2.2. 根据事务 ID 连接后得到 O.结果数据集 O 与原数据集 D 中的事务 ID 在顺序上保持不变.

3.2 概率差分隐私约束

文献[8]指出,算法 1 的随机抽样机制不能满足ε-差分隐私约束.因此,本文采用概率差分隐私模型.针对事务数据发布的隐私保护问题,通过定义概率差分隐私约束条件得到优化问题的约束条件.

令 $D=D'+T_k$,如果初始数据集为 D,得到的结果 O 中可能包含事务 T_k,$\Pr[\mathcal{R}(D)=O]>0$;由于 D'不包含 T_k,结果不可能含 T_k,$\Pr[\mathcal{R}(D')=O]=0$.可将输出空间Ω划分为两部分:Ω_1:所有含 T_k 的结果 O;Ω_2:所有不含 T_k 的结果 O.下面根据该划分讨论概率差分隐私中的概率定义.

(1) 第 1 种情况,$\forall O\in\Omega_1$.

对于空间Ω_1 中的输出结果 O,$\Pr[\mathcal{R}(D')=O]=0$,则 $\Pr[\mathcal{R}(D')\in\Omega_1]=0$,因此,$\mathcal{R}(D')$不泄露 T_k 的隐私,只需考虑 $\Pr[\mathcal{R}(D)\in\Omega_1]$即可.

特别地,对于事务 T_k,从 D 中抽样得到结果 O 包含事务 T_k 的概率 $\Pr[\mathcal{R}(D)=O]$等于从各项中抽到 T_k 的概率.对于所有的项 $item$,从包含该项的所有事务中抽取 $item.x$ 个事务,包含事务 T_k 的概率为

$$\Pr[\mathcal{R}(D)\in\Omega_1]=1-\prod_{\forall item_j\in T_k}\left(1-\frac{item_j.x}{item_j.count}\right) \tag{2}$$

其中,$\prod_{\forall item_j\in T_k}(1-item_j.x/item_j.count)$ 为 T_k 不在结果 O 中的概率.

(2) 第 2 种情况,$\forall O\in\Omega_2$.

因为空间Ω_2 中的所有输出结果 O 不包含事务 T_k,则 $\Pr[\mathcal{R}(D)=O]>0$ 且 $\Pr[\mathcal{R}(D')=O]>0$.由于 D 中含有事务 T_k,$\mathcal{R}(D)$的输出空间中必然有结果 $O'\in\Omega_1$,如图 1 所示.因此,$\Pr[\mathcal{R}(D)=O]/\Pr[\mathcal{R}(D')=O]\leqslant 1\leqslant e^{\varepsilon}$,已满足概率差分隐私条件.

为了推导 $\Pr[\mathcal{R}(D')=O]/\Pr[\mathcal{R}(D)=O]$的值,考虑到从 D'中抽样得到的结果不包含 T_k 的概率为 1,且从 D 中抽样得到的结果包含 T_k 的概率为 $\prod_{\forall item_j\in T_k}(1-item_j.x/item_j.count)$,得到公式(3).

$$\frac{\Pr[\mathcal{R}(D')=O]}{\Pr[\mathcal{R}(D)=O]}=1\Big/\prod_{\forall item_j\in T_k}\left(1-\frac{item_j.x}{item_j.count}\right) \tag{3}$$

对于最多相差一个事务的任意 D 与 D',由于不能断定攻击者对哪个事务未知,因此,D 中所有事务都必须满足概率差分隐私约束条件:

(1) $\forall T_i\in D,1-\prod_{\forall item_j\in T_i}\left(1-\frac{item_j.x}{item_j.count}\right)\leqslant\delta$;

(2) $\forall T_i\in D,1\Big/\prod_{\forall item_j\in T_k}\left(1-\frac{item_j.x}{item_j.count}\right)\leqslant e^{\varepsilon}$.

为了统一形式,对条件(1)与条件(2)两边同时取对数,整理后得到新的概率差分隐私约束条件:

(3) $\forall T_i\in D,\sum_{\forall item_j\in T_i}\ln\left(1-\frac{item_j.x}{item_j.count}\right)\geqslant\theta,\theta=\max\left[\ln\left(\frac{1}{e^{\varepsilon}}\right),\ln(1-\delta)\right]$.

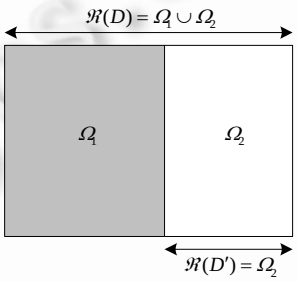


Fig.1 Sample space

图 1 抽样结果空间

3.3 分布式非线性规划模型

分布式结构下,为了分析数据及进行数据挖掘任务,各站点协作共享数据.从数据应用出发,不同应用对数据质量的要求不同,有些应用只需要数据的部分信息.例如,关联规则挖掘仅需要频繁项集.发布数据时,一方面要保护个体隐私,另一方面要为数据分析保留足够的数据信息.可见,优化输出数据的效用性非常必要.本文定义数据效用性优化目标为:最大化支持计数.尽可能最大化结果中项的支持计数,对于 Top-K 查询等应用具有重要的意义.

定义优化目标函数为 $\sum_{\forall item_j}item_j.x$,令概率差分隐私约束条件为优化问题的约束条件,定义如公式(4)的优化问题,该优化问题的解为最优输出效用性.最后,对结果 x 的取值为$\lfloor x\rfloor$.

$$\begin{cases}\max:\sum_{\forall item_j}item_j.x\\ \text{s.t. }\forall T_i\in D,\sum_{\forall item_j\in T_i}\ln\left(1-\frac{item_j.x}{item_j.count}\right)\geqslant\max\left[\ln\left(\frac{1}{e^{\varepsilon}}\right),\ln(1-\delta)\right]\end{cases} \tag{4}$$

优化问题(4)的标准形式为

$$\begin{cases}\min:f(X)=-sum(X)\\ \text{s.t. }g_i(X)=-\sum_{j=1}^{m}\left(\ln\left(1-\frac{X[j]}{C[j]}\right)\times D[i,j]\right)+\theta\leqslant 0,i=1,...,n\\ X[i]\geqslant 0,i=1,...,n\end{cases} \tag{5}$$

其中,n 为事务数;m 为项集数,站点 A 拥有前 m_1 个项,站点 B 拥有后 $m-m_1$ 个项;θ为全局共享参数;X 是一个$(1\times m)$的向量,为各项未知的最优支持计数,全局共享;C 是一个$(1\times m)$的向量,为各项的已知支持计数,站点 A 拥有前 m_1

个值,站点 B 拥有后 $m-m_1$ 个值;D 为垂直划分的分布式事务数据集,$D[i,j]=\{0,1\}$.非线性规划问题(5)的条件约束与项集数$|U|$以及事务数$|T|$相关.

3.4 分布式非线性规划安全求解

随机抽样机制$\mathcal{M}$的第 1 步是求解非线性规划问题(5),得到项的最优支持计数.本文提出两种解决方案:第 1 种为全局解决方案(global solution,简称 GS),基于全局数据,在分布式结构下,首先通过 *K-T* 条件等价转换得到非线性方程组,然后应用 Marquardt 法[42],基于安全数量积和协议求解方程组,得到全局各项最优支持计数;第 2 种为局部解决方案(local solution,简称 LS),基于局部数据,各站点采用 SQP 方法[42]独立求解非线性规划问题(5),得到各自所拥有项的最优支持计数.概率差分隐私组合定理表明,LS 同样满足概率差分隐私约束.

3.4.1 全局解决方案

求解非线性规划问题最有效的方法是采用 SQP[42]方法,由于 SQP 算法的复杂性,难以在分布式结构下实现,根据定理 1,通过 *K-T* 等价条件将非线性规划问题转化为非线性方程组.

定理 1. 非线性规划问题(5)为凸规划问题.

证明:非线性规划问题满足如下两个条件:

(1) 所有的线性函数均为凸函数,因此,$f(X)$为凸函数;

(2) $g_i(X)(i=1,\ldots,n)$的 Hessian 矩阵为

$$H=\nabla^2 g_i(X)=\begin{bmatrix}\dfrac{D[i,1]}{(X[1]-C[1])^2} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \dfrac{D[i,m]}{(X[m]-C[m])^2}\end{bmatrix} \tag{6}$$

其中,$D[i,j]=\{0,1\},j=1,\ldots,m$.对于所有不为 0 的 $X\in R^n$,都有 $XHX^T\geqslant 0$,所以 H 为半正定矩阵,因此,$g_i(X)(i=1,\ldots,n)$为凸函数.

根据凸规划问题的定义[42],非线性规划问题(5)满足同时条件(1)与条件(2),证明非线性规划问题(5)为凸规划问题. □

对于凸规划问题,根据最优解的一阶充分条件[42],如果满足 *K-T* 必要条件的点 $\bar{X}$ 存在,则 $\bar{X}$ 为非线性规划问题(5)的全局最优解.公式(5)的 *K-T* 条件可以等价写为

$$\begin{cases}\nabla f(\bar{X})-\sum_{i=1}^{n} w_i\nabla g_i(\bar{X})=0\\ w_i g_i(\bar{X})=0, i=1,...,n\\ w_i\geqslant 0, i=1,...,n\end{cases} \tag{7}$$

方程组(7)含有 $m+n$ 个未知数(m 个 x,n 个 w)、$m+n$ 个等式方程,为非线性方程组,它的解为优化问题(5)的全局最优解.计算 $f(X)$与 $g_i(X)(i=1,\ldots,n)$的梯度,方程组(7)化解为如下 $m+n$ 个等式方程:

- 前 m 个等式方程为

$$h_j=1+\sum_{i=1}^{n}\left(\frac{w_i\cdot D[i,j]}{X[j]-C[j]}\right)=0, j=1,...,m \tag{8}$$

- 后 n 个等式方程为

$$l_i=w_i\cdot\left(\sum_{j=1}^{m}\left(\ln\left(1-\frac{X[j]}{C[j]}\right)\times D[i,j]\right)-\theta\right)=0, i=1,...,n \tag{9}$$

令 $f=[h_1\ldots h_m\ l_1\ldots l_n]$,该方程组可以应用非线性最小二乘法的改进方法 Marquardt 法[42]求解.在 Marquardt 法的第 k 次迭代中,令:

$$d^{(k)}=-(A_k^TA_k+\alpha_k I)^{-1}A_k^T f^{(k)} \tag{10}$$

其中,I 为单位矩阵,α_k 为一个正实数.显然,当α_k=0 时,$d^{(k)}$就是高斯-牛顿方向.后继点为

$$x^{(k+1)}=x^{(k)}+d^{(k)} \tag{11}$$

公式(10)包含 $f^{(k)}$与 A_k,其计算公式分别为

$$f^{(k)}=(f_1(x^{(k)}),\ldots,f_m(x^{(k)}),f_{m+1}(x^{(k)}),\ldots,f_{m+n}(x^{(k)}))^T \tag{12}$$

$$A_k=\begin{bmatrix} \dfrac{\partial f_1(x^{(k)})}{\partial x_1} & \cdots & \dfrac{\partial f_1(x^{(k)})}{\partial x_m} & \dfrac{\partial f_1(x^{(k)})}{\partial x_{m+1}} & \cdots & \dfrac{\partial f_1(x^{(k)})}{\partial x_{m+n}} \\ \vdots & & \vdots & \vdots & & \vdots \\ \dfrac{\partial f_{m+n}(x^{(k)})}{\partial x_1} & \cdots & \dfrac{\partial f_{m+n}(x^{(k)})}{\partial x_m} & \dfrac{\partial f_{m+n}(x^{(k)})}{\partial x_{m+1}} & \cdots & \dfrac{\partial f_{m+n}(x^{(k)})}{\partial x_{m+n}} \end{bmatrix} \tag{13}$$

$f^{(k)}$是[m+n,1]的列向量,A_k 是[m+n,m+n]的矩阵.分布式计算环境下,站点 A 与站点 B 共享目标函数 $f(X)$、参数(ε,δ)以及每次迭代中的 $X^k=(x_1^k,...,x_m^k),W^k=(w_1^k,...,w_n^k)$.由于数据集 D 的分布式垂直划分,导致向量$f^{(k)}$与矩阵 A_k 也被划分为两部分,分布在站点 A 与站点 B 中.为安全求解公式(10),需要分析 $f^{(k)}$与 A_k的分布.首先,分析前 m 个方程组公式(8)的计算需要站点 A 与站点 B 协作进行,常数项 1 不涉及隐私信息,可由任意站点计算.将公式(8)分解为如下两部分:

$$h_{[1,m_1]}(P_A)=1+\sum_{i=1}^{n}\left(\frac{w_i\cdot D[i,j]}{X[j]-C[j]}\right), j=1,...,m_1 \tag{14-A}$$

$$h_{[m_1+1,m]}(P_B)=\sum_{i=1}^{n}\left(\frac{w_i\cdot D[i,j]}{X[j]-C[j]}\right), j=m_1+1,...,m \tag{14-B}$$

令 $H(P_A)=[h_1(P_A)...h_{m_1}(P_A)]^T$,$H(P_B)=[h_{m_1+1}(P_B)...h_m(P_B)]^T$,公式(8)可统一表示为

$$H=[H(P_A)\ H(P_B)]^T=0 \tag{15}$$

同理,后 n 个方程组公式(9)可分解为两部分,常数θ全局共享,可由任意一方计算:

$$l_{[1,n]}(P_A)=w_i\cdot\sum_{j=1}^{m_1}\left(\ln\left(1-\frac{X[j]}{C[j]}\right)\times D[i,j]\right), i=1,...,n \tag{16-A}$$

$$l_{[1,n]}(P_B)=w_i\cdot\sum_{j=m_1+1}^{m}\left(\ln\left(1-\frac{X[j]}{C[j]}\right)\times D[i,j]\right), i=1,...,n \tag{16-B}$$

令 $L(P_A)=[l_1(P_A)\ldots l_n(P_A)]^T$,$L(P_B)=[l_1(P_B)\ldots l_n(P_B)]^T$,$\Theta=[w_1\theta\ldots w_n\theta]^T$,公式(9)可统一表示为

$$\Gamma=L(P_A)+L(P_B)-\Theta=0 \tag{17}$$

结合公式(15)与公式(17),得到 $f^{(k)}$为

$$f^{(k)}=[[H(P_A)^{(x_k)}]_{[m_1,1]}\ [H(P_B)^{(x_k)}]_{[m-m_1,1]}\ [L(P_A)^{(x_k)}+L(P_B)^{(x_k)}-\Theta]_{[n,1]}]^T_{[1,m+n]}} \tag{18}$$

公式(18)表明,$f^{(k)}$由 3 大部分构成,分布在站点 A 与站点 B 中.令:

$f_A^{(k)}=[[H(P_A)^{(x_k)}]_{[m_1,1]}\ [O]_{[m-m_1,1]}\ [L(P_A)^{(x_k)}]_{[n,1]}]^T_{[1,m+n]}$, $f_B^{(k)}=[[O]_{[m_1,1]}\ [H(P_B)^{(x_k)}]_{[m-m_1,1]}\ [L(P_B)^{(x_k)}-\Theta]_{[n,1]}]^T_{[1,m+n]}$.

$f^{(k)}$最终可表示为两方的和:

$$f^{(k)}=f_A^{(k)}+f_B^{(k)} \tag{19}$$

对公式(15)与公式(17)求导,得到 A_k 为

$$A_k=\begin{bmatrix} \begin{bmatrix} \dfrac{\partial H(P_A)}{\partial x_{[m_1,m]}} & \dfrac{\partial H(P_A)}{\partial w_{[m_1,n]}} \end{bmatrix} \\ \begin{bmatrix} \dfrac{\partial H(P_B)}{\partial x_{[m-m_1,m]}} & \dfrac{\partial H(P_B)}{\partial w_{[m-m_1,n]}} \end{bmatrix} \\ \begin{bmatrix} \dfrac{\partial L(P_A)}{\partial x_{[n,m_1]}} \end{bmatrix} \begin{bmatrix} \dfrac{\partial L(P_B)}{\partial x_{[n,m-m_1]}} \end{bmatrix} \begin{bmatrix} \dfrac{\partial L(P_A)}{\partial w_{[n,n]}}+\dfrac{\partial L(P_B)}{\partial w_{[n,n]}}-\theta \end{bmatrix} \end{bmatrix}_{[m+n,m+n]} \tag{20}$$

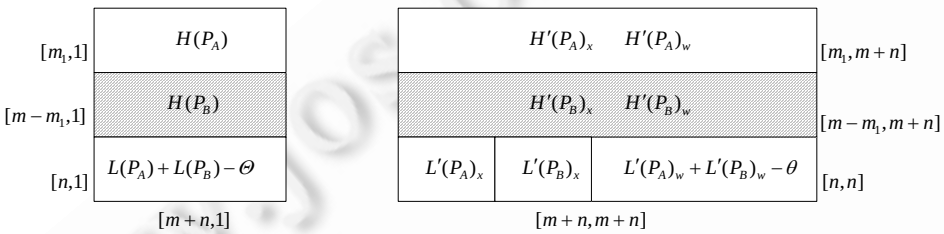
公式(20)表明,A_k 由 5 大部分构成,分布在站点 A 与站点 B 中.令:

$$A_k^A=\begin{bmatrix}\begin{bmatrix}\dfrac{\partial H(P_A)}{\partial x_{[m_1,m]}} & \dfrac{\partial H(P_A)}{\partial w_{[m_1,n]}}\end{bmatrix}\\ [0_{[m-m_1,m]}\ 0_{[m-m_1,n]}]\\ \begin{bmatrix}\dfrac{\partial L(P_A)}{\partial x_{[n,m_1]}}\end{bmatrix}[0_{[n,m-m_1]}]\begin{bmatrix}\dfrac{\partial L(P_A)}{\partial w_{[n,n]}}\end{bmatrix}\end{bmatrix}_{[m+n,m+n]},A_k^B=\begin{bmatrix}[0_{[m_1,m]}\ 0_{[m_1,n]}]\\ \begin{bmatrix}\dfrac{\partial H(P_B)}{\partial x_{[m-m_1,m]}} & \dfrac{\partial H(P_B)}{\partial w_{[m-m_1,n]}}\end{bmatrix}\\ [0_{[n,m_1]}]\begin{bmatrix}\dfrac{\partial L(P_B)}{\partial x_{[n,m-m_1]}}\end{bmatrix}\begin{bmatrix}\dfrac{\partial L(P_B)}{\partial w_{[n,n]}}-\theta\end{bmatrix}\end{bmatrix}_{[m+n,m+n]},$$

A_k 最终可表示为两方的和:

$$A_k=A_k^A+A_k^B \tag{21}$$

上述分析表明,$f^{(k)}$ 与 A_k 在站点 A 与站点 B 的分布情况如图 2 所示.



(a) $f^{(k)}$分布图 (b) A_k 分布图

Fig.2 Distribution of $f^{(k)}$ and A_k

图 2 $f^{(k)}$与 A_k 的分布情况

考虑 Marquardt 算法中 $d^{(k)}=-(A_k^TA_k+\alpha_kI)^{-1}A_k^Tf^{(k)}$,分布式结构下,为保护各站点不泄露隐私信息,需要进行加密操作的计算有 $A_k^Tf^{(k)}$ 与 $A_k^TA_k$.根据公式(19)、公式(21)可得到:

$$\begin{aligned}A_k^Tf^{(k)}&=(A_k^{AT}+A_k^{BT})\cdot(f_A^{(k)}+f_B^{(k)})\\&=(A_k^{AT}\cdot f_A^{(k)}+A_k^{BT}\cdot f_B^{(k)})+(A_k^{AT}\cdot f_B^{(k)}+A_k^{BT}\cdot f_A^{(k)})\end{aligned} \tag{22}$$

同理,

$$\begin{aligned}A_k^TA_k&=(A_k^{AT}+A_k^{BT})\cdot(A_k^A+A_k^B)\\&=(A_k^{AT}\cdot A_k^A+A_k^{BT}\cdot A_k^B)+(A_k^{AT}\cdot A_k^B+A_k^{BT}\cdot A_k^A)\end{aligned} \tag{23}$$

3.4.1.1 数量积和协议

输入:假设站点 A 拥有向量 $a_{[1,n]},b_{[1,m]}$,站点 B 拥有向量 $c_{[1,n]},d_{[1,m]}$;

输出:$R=a\cdot c+b\cdot d$.

(1) 根据已有数量积协议,站点 A 与站点 B 安全计算数量积 $a\cdot c$,站点 A 得到 $a\cdot c+v_1$,站点 B 得到随机数 v_1;

(2) 同理,站点 A 与站点 B 计算数量积 $b\cdot d$,站点 A 得到 $b\cdot d+v_2$,站点 B 得到随机数 v_2;

(3) 站点 A 计算第(1)步与第(2)步结果的和:$r_1=a\cdot c+b\cdot d+v_1+v_2$,站点 B 计算第(1)步与第(2)步结果的和:

$$r_2=v_1+v_2.$$

(4) 站点 B 将计算结果 r_2 发给站点 A,站点 A 计算 $R=r_1-r_2$.

(5) 站点 A 将计算结果 R 发给站点 B.

公式(22)与公式(23)的前半部分可在本地进行而无需进行任何加密操作;后半部分的计算涉及到站点 A 与站点 B 的信息,为防止计算过程中站点 A 与站点 B 的信息泄露,必须加密处理.公式(22)与公式(23)的后半部分均为矩阵与向量或矩阵与矩阵的乘积,涉及到的操作仅有数量积和,为安全执行数量积和操作,本文提出一种安全两方数量积和协议.

3.4.1.2 数量积和协议的安全性分析

数量积和协议安全地计算两个数量积的和 $a\cdot c+b\cdot d$,计算过程中,各站点无法得到 $a\cdot c$ 与 $b\cdot d$ 的值.数量积和协议基于已有的高效安全的数量积协议[39−41].数量积和协议的第(1)步与第(2)步,站点 A 与站点 B 分别执行数量积操作,各站点并不共享得到的结果,$a\cdot c$ 与 $b\cdot d$ 的值是安全的;第(3)步,站点 A 与站点 B 分别对第(1)步与第(2)步的结果求和;第(4)步,站点 B 将随机数的和发给站点 A,站点 A 计算数量积和为:$R=r_1-r_2$,该步骤中,由于站点 A 从站点 B 得到的是两个随机数的和,并不知道 v_1 与 v_2 的值,因此,站点 A 无法知道 $a\cdot c$ 与 $b\cdot d$ 的值,从而推断不出站点 B 的信息;第(5)步中,站点 B 得到数量积和 $a\cdot c+b\cdot d$,同样无法知道 $a\cdot c$ 与 $b\cdot d$ 的值,从而推断不出站点 A 的信息.总之,本文提出的数量积和协议是安全的.

3.4.1.3 全局解决方案(GS)

全局共享:目标函数 $f(X)$、参数(ε,δ)、每次迭代中的 X^k,W^k.

输入:站点 A:D_A;站点 B:D_B;

输出:各项最优支持计数 X.

(1) 非线性规划问题(5)的 K-T 条件等价写为方程组(7)

(2) 设 A 为主站点,使用 Marquardt 法求方程组(7)

(3) Marquardt 法每次迭代:

(4) 站点 A 计算 $f_A^{(k)}$ 与 A_k^A,站点 B 计算 $f_B^{(k)}$ 与 A_k^B.$f^{(k)}$与 A_k 均由两部分组成:

$$f^{(k)}=f_A^{(k)}+f_B^{(k)},A_k=A_k^A+A_k^B.$$

(5) 站点 A 使用数量积和协议计算 $A_k^T f^{(k)}$ 与 $A_k^T A_k$.

(6) 站点 A 计算: $d^{(k)}=-(A_k^T A_k+\alpha_k I)^{-1}A_k^T f^{(k)}$

(7) 站点 A 计算:$x^{(k+1)}=x^{(k)}+d^{(k)}$

(8) 直到 Marquardt 收敛

基于数量积和协议,本文提出全局解决方案安全地求解非线性规划问题(5):第(1)步与第(2)步首先将非线性规划问题(5)的 K-T 条件等价写为方程组(7);其次,从第(3)步开始,采用 Marquardt 法求解方程组,分布式结构下,每次迭代过程为安全计算 $x^{(k+1)}$,第(5)步中使用数量积和协议进行加密计算.全局解决方案基于全局数据求解分布式非线性规划问题,为保证各站点在计算过程中不泄露各自的机密信息,如,项的真实支持计数等,采用两方数量积和协议有效地保证了通信与计算安全.

3.4.2 局部解决方案

全局解决方案同时考虑所有站点拥有项的条件,而局部解决方案中,各站点基于本地所拥有项,首先独立求解优化问题,得到本地所拥有项的最优计数;然后,各站点抽样得到本地事务数据集;最后,通过事务 ID 连接得到集成数据集.定理 2 表明,局部解决方案满足差分隐私要求.

定理 2. 事务数据集 X 被垂直划分为两个部分:X_1,X_2,分别拥有项集 U_1,U_2,全局项集 $U=U_1\cup U_2$.存在两个随机抽样算法 $M_1(X_1)$与 $M_2(X_2)$,分别满足(ε_1,δ_1)与(ε_2,δ_2)-概率差分隐私,则 $M_1(X_1)$与 $M_2(X_2)$的组合 $M(X)$满足$(\varepsilon_1+\varepsilon_2,\delta_1+\delta_2-\delta_1\cdot\delta_2)$-概率差分隐私.

证明:

(1) 第 1 种情况,$\forall O\in\Omega_1$:

因为 M_1 满足(ε_1,δ_1)-概率差分隐私约束,则必然满足 $\Pr[M_1(X_1)\in\Omega_1]\leqslant\delta_1$,所以:

$$\prod_{\forall item_j\in U_1}\left(1-\frac{item_j.x}{item_j.count}\right)\geqslant 1-\delta_1 \tag{24}$$

同理,M_2 满足(ε_2,δ_2)-概率差分隐私约束,则必然满足 $\Pr[M_2(X_2)\in\Omega_2]\leqslant\delta_2$,得到:

$$\prod_{\forall item_j\in U_2}\left(1-\frac{item_j.x}{item_j.count}\right)\geqslant 1-\delta_2 \tag{25}$$

公式(24)与公式(25)左右两端相乘:

$$\prod_{\forall item_j \in U_1}\left(1-\frac{item_j.x}{item_j.count}\right)\cdot\prod_{\forall item_j \in U_2}\left(1-\frac{item_j.x}{item_j.count}\right)=\prod_{\forall item_j \in U_1 \cup U_2}\left(1-\frac{item_j.x}{item_j.count}\right)$$
$$\geqslant(1-\delta_1)\cdot(1-\delta_2) \tag{26}$$
$$=1-(\delta_1+\delta_2)+\delta_1\cdot\delta_2$$

最终得到:

$$\Pr[M(X)\in\Omega_1]=1-\prod_{\forall item_j \in I_1 \cup I_2}\left(1-\frac{item_j.x}{item_j.count}\right)\leqslant(\delta_1+\delta_2)-\delta_1\cdot\delta_2 \tag{27}$$

(2) 第 2 种情况,$\forall O\in\Omega_2$:

$$\Pr[M(X)=O]=\Pr[M_1(X_1)=O]\times\Pr[M_2(X_2)=O]$$
$$\leqslant\Pr[M_1(X_1')=O]\times\Pr[M_2(X_2')=O]\times e^{\varepsilon_1+\varepsilon_2} \tag{28}$$
$$=\Pr[M(X')=O]\times e^{\varepsilon_1+\varepsilon_2}$$

3.4.3 GS 与 LS 对比

LS 中,各点独立完成整个发布策略.然而需要注意的是,本地非线性规划问题(5)的求解过程并不满足差分隐私要求.因为任意事务的改变都会影响最终的结果,导致隐私的泄露.文献[8]存在同样的问题,并提出合理的解决方案.本文可采用相同的方法,对结果添加服从拉普拉斯分布的噪音,具体方法见文献[8].本文主要讨论抽样过程及分布式结构下非线性规划求解,不详细讨论噪音的添加量.

但 GS 中,各站点协作完成发布过程,非线性规划问题(5)的求解不会存在该问题.从全局来看,任意事务中,任意项的改变都会影响结果.分布式结构下,各站点拥有项的支持计数为隐私信息,求解过程中不会泄露给对方,即使知道项的改变发生在哪个事务,也推断不出发生改变的项.因此,基于 GS 的整个发布过程满足差分隐私要求.

4 实验结果与分析

本节我们在实验环境中评价本文提出的差分隐私发布策略.首先考查了在不同参数设置的情况下,(ε,δ)的不同值对 GS 与 LS 结果数据量的影响;其次,对比 GS 与 LS 对 Top-K 项的保留百分比;然后,分析事务数与项集数对 LS 执行时间的影响;最后,从理论上分析 GS 的计算复杂度与通信代价.

4.1 实验设置

- 数据集

本文使用公共基准事务数据集:T40I10D100K,该数据集是人造数据集.由于低支持计数的项在频繁项集挖掘或分类等数据挖掘任务中意义不大,通常会在剪枝步骤中被删除,因此在预处理阶段,移除所有支持计数小于等于 0.1%|T|的项,|T|为事务数.然后,从预处理后的数据集中随机抽取 20 000 个事务、50 个项作为最终的实验数据集,事务数为|T|=20000,项集数|U|=50,数据量为|D|=195913.将事务数据集表示为关系表,站点 A 拥有前 25 个项,站点 B 拥有后 25 个项.表 2 描述实验数据集的基本信息.

Table 2 Experiment data set

表 2 实验数据集

Dataset		#Transactions |D|	Average length	# Items |U|	# Items occurrences	Data size (KB)
T40I10D100K	Original	100 000	39.595 4	924	3 959 538	4 902
	Selected	**20 000**	**9.80**	**50**	**195 913**	**133**

- 运行环境与参数设置

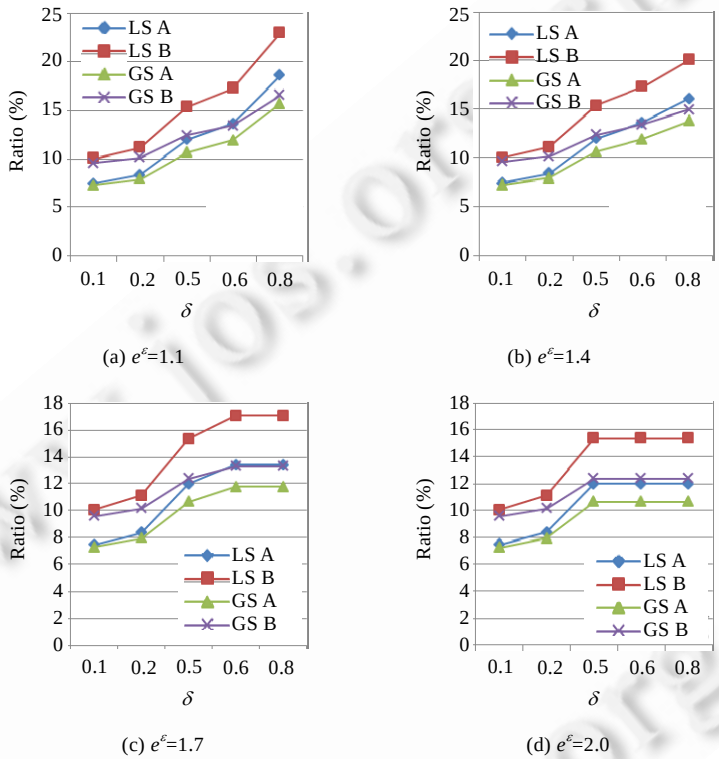
运行环境为 2.50GHz Core™ i5-2450M CPU,6.00GB 内存,64 位 Win7 操作系统.

实验环境为 MATLAB 7.11.0(R2010b).为观察不同参数值对结果的影响,实验过程中参数设置为

δ={0.1,0.2,0.5,0.6,0.8},e^{ε}={1.01,1.1,1.4,1.7,2.0}.

4.2 项支持计数

项支持计数优化问题保证结果数据集的数据量|D'|尽可能大,以保留更多数据信息.根据不同的参数设置,对 T40I10D100K 数据集进行实验.令 *ratio*=|D'|/|D|×100%,表示结果数据量占原数据量的百分比.图 3 给出站点 A 与站点 B 采用 GS 与 LS 求解非线性规划问题的结果.



(a) e^{ε}=1.1 (b) e^{ε}=1.4

(c) e^{ε}=1.7 (d) e^{ε}=2.0

Fig.3 Size of the result data of LS and GS vs. e^{ε}, δ

图 3 LS 与 GS 对不同参数的结果数据量

图 3 表明,结果数据量比较小.主要是因为约束条件与未知变量太多,共有|D|=195913 个约束条件,|U|=50 个变量;又由于事务数据的稀疏性,导致结果数据量较小.这是合理的,与理论分析相符,为差分隐私约束下的最优解.具体来说,对于站点 A,LS 的结果范围是 7.45%~18.58%,GS 的结果范围是 7.24%~15.72%;对于站点 B,LS 的结果范围 10.04%~23.03%,GS 的结果范围是 9.57%~16.58%.从整体来看,LS 所得的结果比 GS 要好.这是由于 GS 所处理的变量总比 LS 要多.图 3 还表明:当δ增大时,结果数据量会增大,与理论分析相符,δ越大时,隐私强度越小,个人隐私越容易泄露,因此结果数据量越大是合理的;而当 e^{ε}增大时,结果数据量反而会随之减小,这也与理论分析相符,e^{ε}越大,隐私强度越大,个人隐私越不容易泄露,因此结果数据量越小是合理的.最后,站点 B 的结果数据量总比站点 A 的结果数据量大,这与站点所拥有事务的平均事务长度有关,站点 A 的事务平均长度为 5.66,站点 B 的事务平均长度为 4.13,因此站点 B 的每个约束条件所含项数平均来说比站点 A 的要少,站点 B 的结果数据量总会比站点 A 的多是合理的.

图 4 给出了当参数 e^{ε}与δ指定时,事务数|T|与项集数|U|对 GS 执行结果的影响.实验令 e^{ε}=1.1,δ=0.8.图 4 表明:事务数|T|较小时对结果数据量有一定的影响,而|T|到一定值时不再是影响结果的主要因素;项集数|U|对结果影响较大,随着项集数|U|的增大,得到的结果数据量减少.因为|U|越大,约束条件越复杂,最优解越小,与理论分析

相符,是合理的.

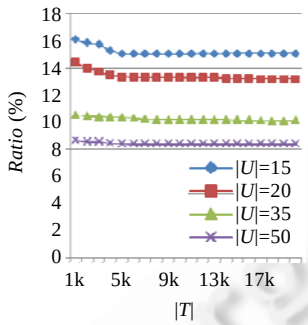
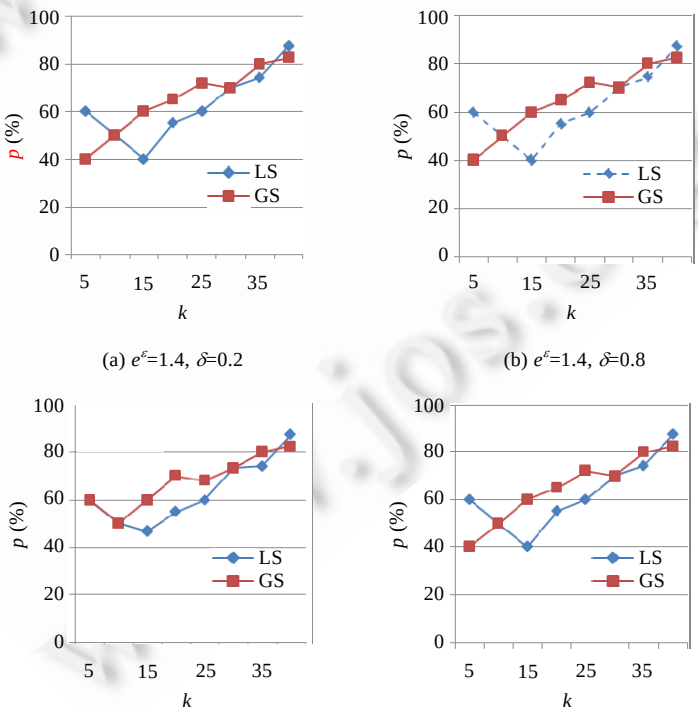


Fig.4 Result of GS vs. |*T*|, |*U*|

图 4 事务数|*T*|、项集数|*U*|对 GS 执行结果的影响

4.3 Top-*K*项

图 5 给出了全局解决方案与局部解决方案输出数据集的效用性.考查指标为 Top-*K* 项的保留百分比,即 $p=|FI\cap FI'|/k$,其中,*FI* 为原事务数据集支持计数排在前 *k* 位的项的集合,*FI'* 为输出数据集支持计数排在前 *k* 位的项的集合.图 5 表明,GS 的执行效果比 LS 要好.这是因为 GS 基于全局的项集,对项的支持计数的顺序保持良好.而 LS 基于局部的项集,尽管局部项的顺序能有很好的保持,但组合后这种保持被打破,因此,GS 的 Top-*K* 项保持效果比 LS 好是合理的.



(a) e^{ε}=1.4, δ=0.2 (b) e^{ε}=1.4, δ=0.8

(c) e^{ε}=2.0, δ=0.2 (d) e^{ε}=2.0, δ=0.8

Fig.5 Retention Top-*K* items of LS and GS

图 5 LS 与 GS 对 Top-*K* 项的保持性

4.4 LS执行时间分析

图 6 显示了 LS 求解优化问题的执行时间,实验令 e^{ε}=1.4,δ=0.5.图 6 表明:当项集数|U|较小时,LS 的执行时间随着事务数|T|的增加而增长,但增长速率较缓慢,导致曲线较平缓;而当项集数|U|较大时,执行时间的增长速率较大,导致曲线较陡峭.可见,项集数|U|是影响 LS 执行时间的主要因素.另有实验表明:当|U|>60 时,在预设定时间内没有执行完成.

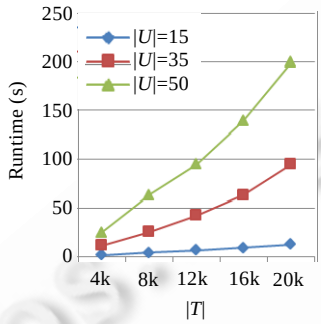


Fig.6 Runtime of LS vs. |T|, |U|

图 6 事务数|T|、项集数|U|对 LS 执行时间的影响

4.5 GS复杂度分析

GS 的计算与通信代价发生在 Marquardt 算法的每次迭代中,因此只要分析 GS 第(4)步~第(7)步的通信代价与计算复杂度.表 3 中假设数量积和协议的通信代价为 $O(\tau'(z))$,其中,z 为向量长度,$\tau'(z)$为所使用的安全数量积协议的复杂度表达式;假设数量积和协议的计算复杂度为 $O(\tau(z))$,其中,z 为向量长度,$\tau(z)$为所使用的安全数量积协议的复杂度表达式.

Table 3 Communication cost and computational complexity of GS

表 3 GS 的通信代价与计算复杂度

	通信代价	计算复杂度
第(4)步	**0**	$O((m+n)^2)$
第(5)步	$O((m+n)\times\tau'(m+n))$	$O((m+n)\times\tau(m+n))$
第(6)步	**0**	$O((m+n)^2)$
第(7)步	**0**	$O(m+n)$
Marquardt 迭代 N 次	$O(N\times(m+n)\times\tau'(m+n))$	$O(N\times((m+n)\times\tau(m+n)+O((m+n)^2)))$

5 结束语

本文针对分布式结构下事务数据发布的隐私保护问题,提出一种差分隐私发布策略.基于差分隐私,构造分布式非线性规划模型,最优化目标函数使结果效用性最大化.设计两种解决方案求解该模型:第 1 种是基于安全数量积协议的全局解决方案;第 2 种是基于组合定理的局部解决方案.理论分析与实验结果证明:本文提出的发布策略在保护隐私的同时最大化结果效用,且计算与通信是安全的.下一步工作中,将讨论水平分布环境下事务数据发布问题,以及针对不同的应用场景提出更多的效用性目标函数.

References:

[1] 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]

[2] Savasere A, Omiecinski ER, Navathe SB. An efficient algorithm for mining association rules in large databases. In: Proc. of the 21th Int'l Conf. on Very Large Data Bases (VLDB). San Francisco: Morgan Kaufmann Publishers, 1995. 432−444.

[3] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc. of the 20th Int'l Conf. on Very Large Data Bases (VLDB). San Mateo: Morgan Kaufmann Publishers, 1994. 487−499.

[4] Adar E, Weld DS, Bershad BN, Gribble SD. Why we search: Visualizing and predicting user behavior. In: Proc. of the 16th Int'l Conf. on World Wide Web. Banff: ACM Press, 2007. 161−170. [doi: 10.1145/1242572.1242595]

[5] Sweeney L. *K*-Anonymity: A model for protecting privacy. Int'l Journal of Uncertainty Fuzziness and Knowledge Based Systems, 2002,10(5):557−570. [doi: 10.1142/S0218488502001648]

[6] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. *L*-Diversity: Privacy beyond *k*-anonymity. ACM Trans. on Knowledge Discovery from Data (TKDD), 2007,1(1):3. [doi: 10.1145/1217299.1217302]

[7] Fung BCM, Wang K, Chen R, Yu PS. Privacy preserving data publishing: A survey of recent developments. ACM Computing Surveys, 2010,42(4):1−53. [doi: 10.1145/1749603.1749605]

[8] Hong Y, Vaidya J, Lu HB, Wu MR. Differentially private search log sanitization with optimal output utility. In: Proc. of the 15th Int'l Conf. on Extending Database Technology. Berlin: ACM Press, 2012. 50−61. [doi: 10.1145/2247596.2247604]

[9] He Y, Naughton JF. Anonymization of set-valued data via top-down, local generalization. Proc. of the VLDB Endowment, 2009, 2(1):934−945. [doi: 10.14778/1687627.1687733]

[10] Xu YB, Wang K, Fu AWC, Yu PS. Anonymizing transaction databases for publication. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD). New York: ACM Press, 2008. 767−775. [doi: 10.1145/1401890.1401982]

[11] Ghinita G, Kalnis P, Tao YF. Anonymous publication of sensitive transactional data. IEEE Trans. on Knowledge and Data Engineering, 2011,23(2):161−174. [doi: 10.1109/TKDE.2010.101]

[12] Loukides G, Gkoulalas-Divanis A, Malin B. COAT: COnstraint-Based anonymization of transactions. Knowledge and Information Systems, 2011,28(2):251−282. [doi: 10.1007/s10115-010-0354-4]

[13] Terrovitis M, Mamoulis N, Kalnis P. Local and global recoding methods for anonymizing set-valued data. The VLDB Journal, 2011,20(1):83−106. [doi: 10.1007/s00778-010-0192-8]

[14] Terrovitis M, Mamoulis N, Kalnis P. Privacy-Preserving anonymization of set-valued data. Proc. of the VLDB Endowment, 2008, 1(1):115−125. [doi: 10.14778/1453856.1453874]

[15] Xu YB, Fung BCM, Wang K, Fu AWC, Pei J. Publishing sensitive transactions for itemset utility. In: Proc. of the IEEE Int'l Conf. on Data Mining (ICDM). Piscataway: IEEE, 2008. 1109−1114. [doi: 10.1109/ICDM.2008.98]

[16] Gkoulalas Divanis A, Loukides G. Utility-Guided clustering-based transaction data anonymization. Trans. on Data Privacy, 2012, 5(1):223−251.

[17] Cao J, Karras P, Raïssi C, Tan KL. ρ-Uncertainty: Inference-proof transaction anonymization. Proc. of the VLDB Endowment, 2010, 3(1-2):1033−1044. [doi: 10.14778/1920841.1920971]

[18] Kifer D. Attacks on privacy and deFinetti's theorem. In: Proc. of the 35th SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2009. 127−138. [doi: 10.1145/1559845.1559861]

[19] Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: Proc. of the IEEE Symp. on Security and Privacy. New York: IEEE, 2008. 111−125. [doi: 10.1109/SP.2008.33]

[20] Ghinita G, Tao Y, Kalnis P. On the anonymization of sparse high-dimensional data. In: Proc. of the Int'l Conf. on Data Engineering (ICDM). Piscataway: IEEE, 2008. 715−724. [doi: 10.1109/ICDE.2008.4497480]

[21] Dwork C. Differential privacy in new settings. In: Proc. of the Annual ACM-SIAM Symp. on Discrete Algorithms. New York: ACM Press, 2010. 174−183.

[22] Dwork C. Differential privacy: A survey of results. Lecture Notes in Computer Science, Heidelberg: Springer-Verlag, 2008: 1−19. [doi: 10.1007/978-3-540-79228-4_1]

[23] Dwork C, Mcsherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. Lecture Notes in Computer Science, Heidelberg: Springer-Verlag, 2006. 265−284. [doi: 10.1007/11681878_14]

[24] Xiao X, Wang G, Gehrke J. Differential privacy via wavelet transforms. IEEE Trans. on Knowledge and Data Engineering (ICDE), 2011,23(8):1200−1214. [doi: 10.1109/TKDE.2010.247]

[25] Hay M, Li C, Miklau G, Jensen D. Accurate estimation of the degree distribution of private networks. In: Proc. of the IEEE Int'l Conf. on Data Mining (ICDM). Piscataway: IEEE, 2009. 169−178. [doi: 10.1109/ICDM.2009.11]

[26] Mcsherry F, Mironov I. Differentially private recommender systems: Building privacy into the net. In: Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2009. 627−636. [doi: 10.1145/1557019.1557090]

[27] Chen R, Mohammed N, Fung BCM, Desai BC, Xiong L. Publishing set-valued data via differential privacy. Proc. of the VLDB Endowment, 2011, 4(11):1087−1089.

[28] Bayardo RJ, Agrawal R. Data privacy through optimal *k*-anonymization. In: Proc. of the Int'l Conf. on Data Engineering. Piscataway: IEEE, 2005. 217−228. [doi: 10.1109/ICDE.2005.42]

[29] Lefevre K, Dewitt DJ, Ramakrishnan R. Mondrian multidimensional *k*-anonymity. In: Proc. of the Int'l Conf. on Data Engineering (ICDE). Piscataway: IEEE, 2006. 25. [doi: 10.1109/ICDE.2006.101]

[30] Ghosh A, Roughgarden T, Sundararajan M. Universally utility-maximizing privacy mechanisms. In: Proc. of the Annual ACM Symp. on Theory of Computing. New York: ACM Press, 2009. 351−360. [doi: 10.1145/1536414.1536464]

[31] Jiang W, Clifton C. A secure distributed framework for achieving k-anonymity. The VLDB Journal, 2006,15(4):316−333. [doi: 10.1007/s00778-006-0008-z]

[32] Mohammed N, Fung BC, Debbabi M. Anonymity meets game theory: Secure data integration with malicious participants. The VLDB Journal, 2011,20(4):567−588. [doi: 10.1007/s00778-010-0214-6]

[33] Jurczyk P, Xiong L. Distributed anonymization: Achieving privacy for both data subjects and data providers. In: Lecture Notes in Computer Science, Canada: Springer-Verlag, 2009. 191−207. [doi: 10.1007/978-3-642-03007-9_13]

[34] Mohammed N, Fung BCM, Hung PCK, Lee CK. Centralized and distributed anonymization for high-dimensional healthcare data. ACM Trans. on Knowledge Discovery from Data (TKDD), 2010,4(4):18. [doi: 10.1145/1857947.1857950]

[35] Alhadidi D, Mohammed N, Fung BCM, Debbabi M. Secure distributed framework for achieving e-differential privacy. Privacy Enhancing Technologies, 2012:120−139. [doi: 10.1007/978-3-642-31680-7_7]

[36] Machanavajjhala A, Kifer D, Abowd J, Vilhuber L. Privacy: Theory meets practice on the map. In: Proc. of the Int'l Conf. on Data Engineering. Piscataway: IEEE, 2008. 277−286. [doi: 10.1109/ICDE.2008.4497436]

[37] Goldreich O. Foundations of Cryptography: Vol.2, Basic Applications. New York: Cambridge University Press, 2004. 615−626.

[38] Yao AC. Protocols for secure computations. In: Proc. of the 23rd Annual Symp. on Foundations of Computer Science. New York: IEEE, 1982. 160−164. [doi: 10.1109/SFCS.1982.38]

[39] Du W, Atallah MJ. Privacy-Preserving cooperative statistical analysis. In: Proc. of the Annual Computer Security Applications Conf. Piscataway: IEEE, 2001. 102−110. [doi: 10.1109/ACSAC.2001.991526]

[40] Goethals B, Laur S, Lipmaa H, Mielikainen T. On private scalar product computation for privacy-preserving data mining. In: Proc. of the Information Security and Cryptology (ICISC 2004). Springer-Verlag, 2005. 23−25. [doi: 10.1007/11496618_9]

[41] Vaidya J, Clifton C. Privacy preserving association rule mining in vertically partitioned data. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2002. 639−644. [doi: 10.1145/775047.775142]

[42] Rao SS. Engineering Optimization: Theory and Practice. Hoboken: Wiley, 2009. 422−425.

附中文参考文献:

[1] 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.

欧阳佳(1986－),男,湖南新化人,博士,主要研究领域为数据挖掘与隐私保护,机器学习.

刘少鹏(1984－),男,博士,主要研究领域为文本挖掘,主题模型.

印鉴(1968－),男,博士,教授,博士生导师,主要研究领域为数据库,数据挖掘,人工智能.