基于差分隐私的不确定数据频繁项集挖掘算法*

丁 哲, 秦 臻, 秦志光
(电子科技大学 信息与软件工程学院, 成都 610054)

摘 要: 基于不确定数据的频繁项集挖掘算法已经得到了广泛的研究。对于记录用户敏感信息的不确定数据,攻击者可以利用自己掌握的背景信息,通过分析基于不确定数据的频繁项集从而获得用户的敏感信息。为了从不确定的数据集中挖掘出基于期望支持度的前 K 个最频繁的频繁项集,并且保证挖掘结果满足差分隐私,提出了 FIMUDDP(frequent itemsets mining for uncertain data based on differential privacy)算法。FIMUDDP 算法利用差分隐私的指数机制和拉普拉斯机制确保从不确定数据中挖掘出的基于期望支持度的前 K 个最频繁的频繁项集和这些频繁项集的期望支持度满足差分隐私。通过对 FIMUDDP 进行理论分析和实验评估,验证了 FIMUDDP 算法的有效性。

关键词: 差分隐私; 不确定数据的频繁项集; 截断期望支持度
中图分类号: TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2018)07-1942-05
doi:10.3969/j.issn.1001-3695.2018.07.004

Frequent itemsets mining for uncertain data based on differential privacy

Ding Zhe, Qin Zhen, Qin Zhiguang
(*Shool of Information & Software Engineering, University of Electronic Science & Technology of China, Chengdu* 610054, *China*)

Abstract: Frequent itemsets mining for uncertain data has been studied extensively. For uncertain data which recorded private information of users, an attacker made use of background knowledge to obtain private information of users by analyzing frequent itemsets mined from uncertain data. In this regard, this paper proposed a new algorithm, denoted as FIMUDDP (frequent itemsets mining for uncertain data based on differential privacy), to mine the top K most frequent itemsets based on expected support from uncertain data and satisfy differential privacy. FIMUDDP algorithm applied exponential mechanism and Laplace mechanism in differential privacy to ensure differential privacy for the top K most frequent itemsets based on expected supports of these frequent itemsets respectively. Finally, through analyzing FIMUDDP from theory and experiment evaluation, the results demonstrate the effectiveness of FIMUDDP.

Key words: differential privacy; frequent itemsets for uncertain data; truncated expected support

0 引言

随着互联网技术的不断发展,网络运营商获取用户个人数据的手段不断增多。很多机构掌握了大量的用户个人数据,而这些被机构所掌握的用户个人数据中往往都记录着用户的隐私信息。如何从机构所掌握的用户个人数据中获取对机构有用的信息,与此同时确保用户的隐私信息不会泄露,成为了学界研究的热点[1]。本文聚焦于如何从不确定的数据集中挖掘前 K 个基于期望支持度的最频繁的频繁项集,与此同时确保挖掘出的结果满足差分隐私。

频繁项集挖掘是数据挖掘技术的一个重要分支,并且频繁项集挖掘技术已经被应用到了多个领域,诸如推荐系统、生物信息分析等。频繁项集挖掘所面临的一个重要的挑战就是数据的不确定性。导致数据不确定性的原因很多,如实验误差、噪声和数据的不完整性等。传统的基于确定数据的频繁项集挖掘算法很难直接被应用于挖掘不确定数据的频繁项集,因此,基于不确定数据的频繁项集挖掘算法越来越受关注[2]。对于记录用户隐私信息的不确定数据,攻击者利用自己所掌握的背景信息,通过分析从不确定数据中挖掘出的频繁项集就可以窃取用户的敏感信息[3]。例如,病人的病例都记录着这个病人所患的病和该病人的临床症状,为了研究临床症状和疾病之间的关系,在病人的病例中,病人的临床症状会关联一个概率,这个概率表示这个症状对于病人来说属于身体不正常的可能性。大量这种病例数据就组成了一个不确定数据集[4]。病人的病例对于每个病人都属于敏感信息,攻击者可以利用自己所掌握的背景信息,通过对从病例数据集中挖掘出的基于期望支持度的频繁项集进行分析,从而窃取病人的敏感信息。因此,在从不确定的数据集中得到有用信息的同时,确保用户的隐私不被泄露是非常重要的。

当一个随机算法的输入数据集中的一条记录发生改变时,差分隐私可以保证这种改变对于该随机算法的输出是不敏感的,所以攻击者就不可能通过分析算法的输出而窃取记录在数据集中的用户敏感信息。差分隐私不需要对攻击者所掌握的信息和对所保护的敏感信息有任何前提的基础上,对用户的隐私数据进行保护。

目前对基于差分隐私的确定数据频繁项集挖掘算法有了很多研究成果,但基于差分隐私的不确定数据频繁项集挖掘算法的研究还很少。在文献[3]中,截断频率的方法被用来提高基于差分隐私的确定数据频繁项集挖掘算法的效率,通过分析发现截断频率的方法也可以适用于提高基于差分隐私的不确定数据频繁项集挖掘算法的效率。基于确定数据的频繁项集挖掘算法和基于不确定数据频繁项集挖掘算法最大的不同在于,用哪种方法确定项集是否属于频繁项集。对于确定数据的频繁项集挖掘算法,利用项集出现的频率来确定项集是否是频繁的;对于基于期望支持度的不确定数据的频繁项集挖掘算法,利用表示项集出现频率的随机变量的数学期望来确定项集

收稿日期: 2017-03-23; **修回日期**: 2017-04-30 **基金项目**: 国家自然科学基金资助项目(61672135,61370026);国家"863"计划资助项目(2015AA016007);四川省科技计划资助项目(2015GZ0095,2016JZ0020);国家自然科学基金委员会—广东省人民政府自然科学联合基金重点项目(U1401257)

作者简介: 丁哲(1982-),男,博士研究生,主要研究方向为机器学习、信息安全(dingzhe0301@hotmail.com);秦臻(1983-),男,副教授,博士,主要研究方向为网络测量、无线传感器网络、移动社交网络;秦志光(1956-),男,教授,博士,主要研究方向为信息和网络安全、电子商务、智能交通系统.

是否是频繁的[2]。本文的贡献包括:a)提出了一个新的基于差分隐私的不确定数据频繁项集挖掘算法,即 FIMUDDP 算法;b)通过理论上的分析,得到 FIMUDDP 算法满足差分隐私;c)利用六个现实世界的数据集和两个动态生成数据集,验证了 FIMUDDP 算法的有效性。

1　相关工作

为了从确定数据集挖掘频繁项集,Agrawal 等人提出了 Apriori 算法,但 Apriori 算法存在两个缺陷:a)在 Apriori 算法运行过程中,会生成大量的候选项集;b)在 Apriori 算法运行过程中,需要多次扫描数据集。为了提高 Apriori 算法的效率,Han 等人提出了 FP-tree 算法。然而数据的不确定性对于很多应用来说是固有的,传统的基于确定数据的频繁项集挖掘算法,如 Apriori 算法和 FP-tree 算法,不能被直接应用于基于不确定数据的频繁项集挖掘算法[5],目前基于不确定数据的频繁项集挖掘研究一共分为两类:一类是基于期望支持度的频繁项集挖掘算法。在 2007 年,为了从不确定数据中挖掘频繁项集,Chui 等人[4]提出了期望支持度的概念,并且在此基础上提出了 U-Apriori 算法;2008 年 Leung 等人[6]提出了基于期望支持度的 UF-Growth 算法,UF-Growth 算法将分而制之和深度优先的策略引入了基于不确定数据的频繁项集挖掘算法。另一类是概率频繁项集挖掘。在 2009 年,Bernecker 等人[7]在 Apriori 算法的基础上提出了动态地挖掘基于不确定数据的频繁项集;2010 年 Calders 等人[8]利用正态分布来评估不确定数据中频繁项集的频繁度;在文献[9~11]中提出了基于不确定数据的闭频繁项集挖掘算法;2016 年,Lin 等人[12]提出了基于权重的不确定数据频繁项集挖掘算法;2016 年,Ahmed 等人[13]将 WUIP-tree 引入了挖掘不确定数据的频繁项集算法,并且提出了新的实验评估方法;2016 年,Lin 等人[14]提出了 PHUI-UP 算法,该算法可以在不确定的数据中挖掘出高实用性而且存在概率高的项集。

2006 年,为了防止攻击者利用自己所掌握的背景信息对用户敏感信息进行窃取,Dwork 等人[15]提出了差分隐私的概念,并且在此基础上他们还提出了差分隐私的拉普拉斯机制。为了确保算法的非实数值的输出满足差分隐私,2007 年 McSherry 等人[16]提出了差分隐私的指数机制。为了确保从确定数据集中挖掘长度为 l 的前 K 个最频繁项集满足差分隐私,Bhaskar 等人[3]提出了基于截断频率的确定数据频繁项集挖掘算法,该算法利用基于差分隐私的拉普拉斯机制和指数机制来确保其结果满足差分隐私。2012 年 Li 等人[17]将最大格的概念引入了挖掘基于确定性数据的频繁项集挖掘算法;2015 年 Su 等人[18]提出了 PFP-Growth 算法,PFP-Growth 算法利用对交易数据集中的交易进行分裂,从而降低了噪声对从确定数据中挖掘出的频繁项集的影响;2015 年,Maruseac 等人[19]在差分隐私理论和抽样理论的基础上提出了挖掘高置信度关联规则算法;2016 年,Xu 等人[20]提出了基于差分隐私的频繁项集挖掘算法。本文提出了一种新的基于差分隐私不确定数据的频繁项集挖掘算法,即 FIMUDDP 算法。FIMUDDP 算法利用截断频率的方法对挖掘过程进行优化。

2　基本概念

表 1 给出一些在本文后续所用到的符号和这些符号的含义。在本文中将基于期望支持度的频繁项集简称为频繁项集。

2.1　基于不确定数据的频繁项集挖掘

令 $V=\{v_1,v_2,\cdots,v_m\}$ 是 n 个项的集合,$T=\{t_1,t_2,\cdots,t_m\}$ 是由 m 个记录组成的不确定数据。T 中的每一条记录 $t_i(1\leqslant i\leqslant n)$ 是由元组组成的集合,t_i 中每个元组可以表示为$\langle v_j(1\leqslant j\leqslant m):P(v_j\in t_i)\rangle$。该元组的第一个元素是集合 V 中的元素 v_j,第二个元素表示 v_j 属于 t_i 的概率。表 2 表示一个不确定的数据集。其中 $V=\{\text{football},\text{basketball},\text{baseball},\text{golfball}\}$,$T=\{t_1,t_2\}$;从 t_1 中可以知道,football 属于 t_1 的概率是 0.3,basketball 属于 t_1 的概率是 1,V 中其他元素属于 t_1 的概率为 0。

表 1　符号表示

符号	符号含义
V	项的集合
n	V 中元素的个数
T	不确定的数据集
m	T 中的记录个数
ε	隐私保护预算
l	前 K 个最频繁的频繁项集中,项集的最大长度
S_K	第 K 个最频繁项集的期望支持度
W	不确定数据集的可能世界
$S_e(X)$	项集 X 的期望支持度
$\hat{S}_e(X)$	项集 X 的噪声期望支持度
$f_e(X)$	项集 X 的截断期望支持度

表 2　不确定数据集

ID	不确定数据的记录
t_1	(football: 0.3),(basketball: 1)
t_2	(basketball: 1),(football: 0.4),(golfball: 0.8)

根据不确定数据记录中每个项所附带的概率值,可以推测出该不确定数据集的可能世界。设不确定数据集的可能世界为 $W=\{w_1,w_2,\cdots,w_d\}$。表 3 是由表 2 所示的不确定数据集推测出的可能世界集。

表 3　不确定数据集的可能世界

W	可能世界	$P(w_i)$
w_1	{basketball}, {baseball}	0.084
w_2	{basketball}, {baseball, football}	0.056
w_3	{basketball}, {baseball, golfball}	0.336
w_4	{basketball}, {baseball, football, golfball}	0.224
w_5	{football, basketball}, {baseball}	0.036
w_6	{basketball, football}, {baseball, football}	0.024
w_7	{basketball, football}, {football, golfball}	0.144
w_8	{basketball, football}, {baseball, football, golfball}	0.096

假设不确定数据集中所有的记录相互独立,相同记录中不同的项相互独立,则可能世界 $w_g(1\leqslant g\leqslant d)$ 的概率 $P(w_g)$ 和项集 X 的期望支持度 $S_e(X)$ 通过式(1)(2)得到[4]。

$$P(w_g)=\prod_{i=1}^{n}\Big(\prod_{x\in I(w_g,i)}P(x\in t_i)\times\prod_{y\notin I(w_g,i)}(1-P(y\in t_i))\Big)\tag{1}$$

$$S_e(X)=\sum_{i=1}^{d}P(w_i)\times S(X,w_i)\tag{2}$$

其中:$I(w_g,i)$ 表示不确定记录 t_i 中在可能世界 w_g 中存在的项集;$S(X,w_i)$ 表示项集 X 在可能世界 w_g 中的支持度计数。在表 3 的 w_2 中可以得到

$$S(\text{football},w_2)=1$$

$$I(w_2,1)=\{\text{basketball}\},I(w_2,1)=\{\text{baseball},\text{football}\}$$

$$P(w_2)=1\times(1-0.3)\times1\times0.4\times(1-0.8)=0.056$$

2.2　差分隐私

两个数据集是一对邻居数据集,当且仅当两个数据集有且仅有一个记录不同。

定义 1　差分隐私[15]。对于任意随机算法 A,设 range(A)是算法 A 的输出域。算法 A 满足 ε-差分隐私,当且仅当对于任意一对邻居数据集 D、D',不等式(3)被满足。

$$P[A(D)=S]\leqslant e^{\varepsilon}\times P[A(D')=S]\tag{3}$$

其中:$S\in\text{range}(A)$;ε 表示差分隐私的隐私保护预算。

定义 2　敏感度[15]。给定函数 $f:D^n\to\mathrm{R}^u$,Δf 表示函数 f 的敏感度并可以利用式(4)得到。

$$\Delta f=\max_{D,D'}\|f(D)-f(D')\|_1\tag{4}$$

其中:D 和 D'是任意一对邻居数据集。

定义 3　拉普拉斯机制[15]。给定多维实数查询函数 q:$D\times O\to\mathrm{R}^u$,设 q 的敏感度为 Δq,算法 A 满足 ε-差分隐私。

$$A(D)=q(D)+\text{Lap}\left(\frac{\Delta q}{\varepsilon}\right)$$

其中:Lap(λ)表示噪声数据,服从拉普拉斯分布(位置参数为0,尺度参数为λ)。

定义 4 指数机制[16]。给定数据集 D 和算法 A 的输出 O,令 $q:D\times O\to \mathrm{R}$ 为算法 A 的输出 O 的积分函数,Δq 是 q 的敏感度,如果算法 M 满足等式(5),则算法 M 满足 ε-差分隐私。

$$M(D,q)=\left\{ r \left| P(r\in O)\propto \exp\left(\frac{\varepsilon\times q(D,r)}{2\Delta q}\right)\right.\right\} \tag{5}$$

性质 1[3] 给定序列算法 $f=f_1,f_2,\cdots,f_u$,如果每个算法 f_i $(1\leqslant i\leqslant u)$ 满足 ε_i-差分隐私,那么 f 满足 $\sum_{i=1}^{u}\varepsilon_i$-差分隐私。

性质 2[3] 对于不确定数据集 T,正实数 λ 和 $\rho(0\leqslant\rho\leqslant 1)$,令 S 是基于不确定数据集的频繁项集挖掘算法 A 在数据集 T 上运行的结果,l 是 S 中项集的最大长度,S_K 是第 K 个最频繁的频繁项集的期望支持度。A 是(λ,ρ)-有用的,当且仅当 S 满足如下两个性质:a)所有期望支持度大于等于 $S_K+\lambda$ 的项集都在 S 中;b)所有期望支持度小于 $S_K-\lambda$ 的项集都不在 S 中。

3 FIMUDDP 算法

这部分将会介绍 FIMUDDP 算法,并证明 FIMUDDP 算法满足差分隐私。FIMUDDP 算法可以从不确定数据集中挖掘前 K 个最频繁的频繁项集,并且确保结果满足差分隐私。FIMUDDP 算法一共分为两个阶段,在第一阶段中,FIMUDDP 算法会得到不确定数据集的基于期望支持度的前 K 个最频繁的频繁项集,在这一步差分隐私的指数机制被应用去确保输出结果满足差分隐私;在第二阶段,利用差分隐私的拉普拉斯机制得到不确定数据的前 K 个频繁项集的噪声期望支持度,从而确保基于不确定数据的前 K 个最频繁的频繁项集的期望支持度满足差分隐私。

令 FIMUDDP 算法的总隐私保护预算为 ε,$\alpha\varepsilon$ 和 $(1-\alpha)\varepsilon$ 是分配给 FIMUDDP 算法第一阶段和第二阶段的隐私保护预算。设 $\alpha=0.5$。

3.1 FIMUDDP 算法描述

1)FIMUDDP 算法第一阶段过程描述

在 FIMUDDP 算法的第一阶段,将会获得不确定数据集中的前 K 个最频繁的频繁项集。在这一阶段利用截断频率的方法[3]来提高算法的效率,具体的步骤如下:

a)利用传统的不确定数据的基于期望支持度的频繁项集挖掘算法,得到最大长度为 l 的项集和这些项集的期望支持度。其中 l 为不确定数据集中前 K 个最频繁的频繁项集中项集的最大长度。

b)设挖掘出的项集中,第 K 个最频繁的频繁项集的期望支持度为 S_K,则挖掘出的项集 X 的截断期望支持度为 $f_e(X)$,可以通过式(6)获得。

$$f_e(X)=\begin{cases} S_e(X) & S_e(X)\geqslant\varphi=S_K-\lambda \\ \varphi=S_K-\lambda & S_e(X)<\varphi=S_K-\lambda \end{cases} \tag{6}$$

其中:λ 可以通过式(7)得到,其中 $0\leqslant\rho\leqslant 1$。

$$\lambda=\frac{4K}{\varepsilon}\left(\ln\left(\frac{K}{\rho}\right)+l\ln(n)\right) \tag{7}$$

c)利用差分隐私的指数机制,从挖掘出的频繁项集中无放回地抽取 K 个,抽取出的 K 个项集为不确定数据集的前 K 个最频繁的频繁项集。令挖掘出的项集的截断期望支持度为差分隐私指数机制的积分函数,项集 X 被无放回抽样的概率表示为 $P_S(X)$,满足式(8)。

$$P_S(X)\propto\exp\left(\frac{\varepsilon}{4K}f_e(X)\right) \tag{8}$$

2)FIMUDDP 算法第二阶段过程描述

在 FIMUDDP 算法的第一阶段,FIMUDDP 算法得到了不确定数据的前 K 个最频繁的频繁项集;在第二阶段,将会生成这些频繁项集的噪声期望支持度。生成的噪声期望支持度可以确保挖掘出的频繁项集的期望支持度满足差分隐私。

假设 $H=\{h_1,h_2,\cdots,h_K\}$ 是 FIMUDDP 算法第一阶段挖掘出的不确定数据的前 K 个最频繁的频繁项集,项集 $h_u(1\leqslant u\leqslant K)$ 的噪声期望支持度可以通过式(9)得到。

$$\widehat{S_e}(h_u)=S_e(h_u)+\xi_u \tag{9}$$

其中:$\xi_1,\xi_2,\cdots,\xi_K$ 是噪声数据,这些噪声数据独立同分布于拉普拉斯分布(位置参数为0,尺度参数为 $2K/\varepsilon$)。

3.2 FIMUDDP 算法的隐私分析

在这一部分将证明 FIMUDDP 算法满足 ε-差分隐私。

性质 3 任意项集 X 的截断期望支持度的敏感度等于 1。

证明 根据式(2),可以通过另一种方法计算项集 X 的期望支持度[4]。

$$S_e(X)=\sum_{i=1}^{m}\prod_{x\in X}P(x\in t_i) \tag{10}$$

其中:m 表示不确定数据集中记录的数目;$t_i(1\leqslant i\leqslant m)$ 表示不确定数据集中的一条数据。令 T 和 T' 是任意一对邻居数据集,$D=\{t\mid t\in T\cap t\in T'\}$ 是 T 和 T' 的交集;$S_e^T(X)$ 和 $S_e^{T'}(X)$ 是项集 X 在 T 和 T' 中的期望支持度;S_K^T 和 $S_K^{T'}$ 是数据集 T 和 T' 中的第 K 个最频繁项集的期望支持度,$d_1=\{t\mid t\in T\cap t\notin T'\}$,$d_2=\{t\mid t\notin T\cap t\in T'\}$。

$$\left.\begin{aligned} S_K^T&=S_K^D+\beta_1\\ S_K^{T'}&=S_K^D+\beta_2\end{aligned}\right\}\Rightarrow\|S_K^T-S_K^{T'}\|_1=\|\beta_1-\beta_2\|_1\leqslant 1$$

根据式(10),项集 X 在不确定数据集 T 和 T' 中的期望支持度的计算方式如下:

$$S_e^T(X)=\sum_{j=1}^{|T|}\prod_{x\in X}P(x\in t_j)=\sum_{j=1}^{|D|}\prod_{x\in X}P(x\in t_j)+\prod_{x\in X}P(x\in d_1)$$

$$S_e^{T'}(X)=\sum_{j=1}^{|T'|}\prod_{x\in X}P(x\in t_j)=\sum_{j=1}^{|D|}\prod_{x\in X}P(x\in t_j)+\prod_{x\in X}P(x\in d_2)$$

当 $S_e^T(X)\geqslant S_K^T-\lambda$,$S_e^{T'}(X)\geqslant S_K^{T'}-\lambda$ 时,

$$\|f_e^T(X)-f_e^{T'}(X)\|_1=\left\|\prod_{x\in X}P(x\in d_1)-\prod_{x\in X}P(x\in d_2)\right\|_1\leqslant 1$$

当 $S_e^T(X)\geqslant S_K^T-\lambda$,$S_e^{T'}(X)<S_K^{T'}-\lambda$ 时,

$$f_e^T(X)-f_e^{T'}(X)=S_e^T(X)-S_e^{T'}(X)+\lambda\leqslant S_K^T-\lambda+S_K^{T'}+\lambda\leqslant 1$$

$$f_e^T(X)-f_e^{T'}(X)=S_e^T(X)-S_e^{T'}(X)\geqslant -1\Rightarrow$$

$$\|f_e^T(X)-f_e^{T'}(X)\|_1\leqslant 1$$

当 $S_e^T(X)<S_K^T-\lambda$,$S_e^{T'}(X)<S_K^{T'}-\lambda$ 时,

$$\|f_e^T(X)-f_e^{T'}(X)\|_1=\|S_K^T-\lambda-S_K^{T'}+\lambda\|_1\leqslant 1$$

综上所述,项集 X 的截断期望支持度的敏感度为 1。

定理 1 FIMUDDP 算法的第一阶段满足 $(\varepsilon/2)$-差分隐私。

证明 根据性质 3,得到单个项集的截断期望支持度的敏感度为 1,那么获得前 K 个最频繁项集的截断期望支持度的敏感度等于 K。根据差分隐私的指数机制,本文定义项集的截断支持度为积分函数,则无放回地抽取项集的概率满足如下条件:

$$P_S(X)\propto\exp\left(\frac{(\varepsilon/2)\times f_e(X)}{2K}\right)=\exp\left(\frac{\varepsilon\times f_e(X)}{4K}\right)$$

性质 4 令 H 是由 FIMUDDP 算法第一阶段所抽取的项集所组成的集合,l 为 H 中项集的最大长度。如果 H 中项集的期望支持度都大于 $S_K-\lambda$ 的概率大于等于 $1-\rho$,那么

$$\lambda=\frac{4K}{\varepsilon}\left(\ln\left(\frac{K}{\rho}\right)+l\ln(n)\right)$$

证明 如果在一次抽样中,截断期望支持度为 f 的项集没有抽中,那么抽中截断期望支持度小于 $f-\lambda$ 的项集的概率小于 $\frac{e^{\frac{\varepsilon(f-\lambda)}{4K}}}{e^{\frac{\varepsilon f}{4K}}}=\exp\left(-\frac{\varepsilon\lambda}{4K}\right)$[3]。

截断期望支持度小于 $f-\lambda$ 的项集最多有 n^l 个,所以抽样过程中,抽到截断期望支持度小于 $f-\lambda$ 的项集最大为 $n^l\times\exp\left(-\frac{\varepsilon\lambda}{4K}\right)$[3]。

在 FIMUDDP 算法的第一阶段一共要抽取 K 个项集,那么这些项集的期望支持度小于等于 $S_K-\lambda$ 的概率最大为[3,14]

$$K\times n^l\times\exp\left(-\frac{\varepsilon\lambda}{4K}\right)$$

$$\rho \geqslant K \times n^{l} \times \exp\left(-\frac{\varepsilon\lambda}{4K}\right) \Leftrightarrow \lambda \geqslant \frac{4K}{\varepsilon}\left(\ln\left(\frac{K}{\rho}\right) + l\ln(m)\right)$$

定理 2　FIMUDDP 算法的第二阶段满足($\varepsilon/2$)-差分隐私。

证明　根据性质 3,得到单个项集的截断期望支持度的敏感度为 1,那么得到不确定数据集中前 K 个最频繁的频繁项集的期望支持度的敏感度等于 K。根据差分隐私的拉普拉斯机制,对于查询出的期望支持度加入噪声。所加入的噪声服从拉普拉斯分布,其位置参数为 0,尺度参数为 $\frac{K}{\varepsilon/2}=\frac{2K}{\varepsilon}$。

FIMUDDP 算法的第一阶段和第二阶段都满足($\varepsilon/2$)-差分隐私。根据性质 1,可以得到 FIMUDDP 算法满足 ε-差分隐私的结论。

4　实验及分析

目前关于基于差分隐私的确定数据频繁项集挖掘的研究很多,但基于差分隐私的不确定数据频繁项集挖掘的研究很少,为了验证 FIMUDDP 算法的有效性,本文利用六个真实数据集和两个动态生成数据集来进行实验验证。这些数据集可以从 http://fimi.ua.ac.be/data/下载,其参数如表 4 所示。在表 4 中,N 表示公开数据集项的数目;M 表示公开数据集中交易的数目;average|t| 表示公开数据集中交易的平均长度。为了向公开数据集中引入不确定性,在本文进行的所有实验中,给公开数据集中每个交易中的每个项引入概率值,这些概率值服从均值为 0.5、方差为 0.125 的高斯分布。

表 4　公开数据集的参数

数据集	N	M	average|t|
chess	3 196	75	37
kosarak	990 002	41 270	8.1
accidents	340 183	468	33.81
mushroom	8 123	119	24
pumsb	49 046	2 113	75
pumsb_star	49 046	2 088	51.48
T10I4D100K	100 000	871	11.1
T40I10D100K	100 000	943	40.61

4.1　评估标准

本文利用准确度和相对错误(RE)来对实验结果进行评估。令 F 是 FIMUDDP 算法从不确定的数据集中挖掘出的前 K 个最频繁的频繁项集;$\widehat{F}$ 是不确定数据集中,真实的前 K 个频繁项集。准确率可以由式(11)得到。$S_e(X)$ 表示项集 X 的期望支持度,$\widehat{S}_e(X)$ 表示项集 X 的噪声期望支持度,相对错误(RE)可以由式(12)得到。

$$\text{precision} = \frac{|F \cap \widehat{F}|}{K} \tag{11}$$

$$\text{RE} = \text{median}_{X \in \widehat{F}} \frac{|\widehat{S}_e(X) - S_e(X)|}{S_e(X)} \tag{12}$$

4.2　实验结果分析

FIMUDDP 算法可以从不确定的数据集中挖掘出前 K 个最频繁的频繁项集,并且对挖掘出的结果确保差分隐私。为了评估参数 ρ 对 λ/S_K 的影响,设置 $K=30$,隐私保护预算($\varepsilon/2$) = 0.8。图 1 显示的是随着参数 ρ 的增加,λ/S_K 的变化。从图 1 可以看出,随着参数 ρ 的增加,对于所有公共数据集来说,λ/S_K 都略微地下降,当 $\rho \geqslant 0.3$ 时,λ/S_K 的变化趋势相对稳定。所以在本文实验中 ρ 的值设置为 0.3。

为了评估隐私保护预算($\varepsilon/2$)对 λ/S_K 的影响,设置 $K=30$,$\rho=0.3$。随着隐私保护预算的增加,λ/S_K 的变化如图 2 所示。从图 2 中可以看出,随着隐私保护预算的增加,λ/S_K 单调下降。对于不同的数据集,S_K 越大,则 λ/S_K 越小。对于数据集 accidents、pumsb、pumsb_star、kosarak 和 T40I10D100K,当隐私保护预算($\varepsilon/2$) $\geqslant 0.5$ 时,λ/S_K 下降的趋势明显减缓;对于其他三个数据集,当隐私保护预算($\varepsilon/2$) $\geqslant 1.6$ 时,λ/S_K 的下降趋势稳定。

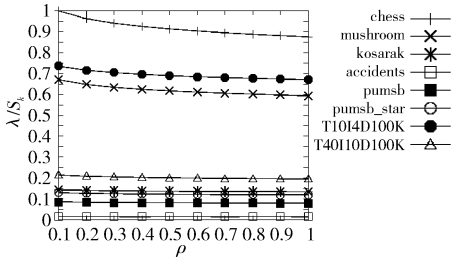

图 1　随着 ρ 的增加 λ/S_K 的变化趋势

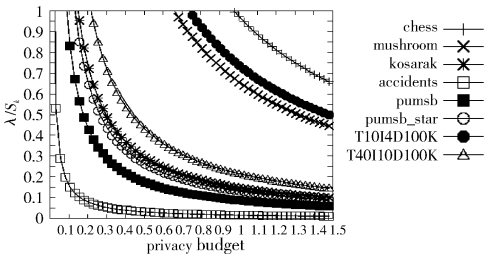

图 2　随着隐私保护预算($\varepsilon/2$)的增加 λ/S_K 的变化趋势

为了评估隐私保护预算($\varepsilon/2$)对准确率的影响,设置 $K=30$,$\rho=0.3$。随着隐私保护预算($\varepsilon/2$)的增加,准确率的变化如图 3 所示。从图 3 中可以看出,随着隐私保护预算($\varepsilon/2$)的增加,所有准确率趋近于 1。根据式(8),随着隐私保护预算($\varepsilon/2$)的增加,前 K 个最频繁的频繁项集被抽中的概率就越大,所以准确率就不断地提高。但是对于不同的数据集,前 K 个频繁项集的期望支持度越小,收敛的速度越慢。根据式(6),期望支持度越大,截断期望支持度就越大;根据式(8),如果截断期望支持度越大,自然会影响到抽取前 K 个最频繁的频繁项集的概率值。

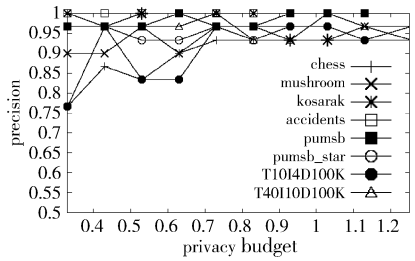

图 3　随着隐私保护预算($\varepsilon/2$)的增加准确率的变化趋势

为了评估 K 的值对准确率的影响,本文设置隐私保护预算($\varepsilon/2$) = 0.8,$\rho=0.3$。随着 K 值的增加,准确率的变化如图 4 所示。从图 4 中可以看出,随着 K 值的增加,对于不同数据集的准确率都会下降。这是因为根据式(7),随着 K 的增加,λ 会增加,这就导致了期望支持度低的频繁项集具有较高的抽样概率,从而影响了算法的准确率。但是对于不同的不确定数据集,如果不确定数据集中项集的期望支持度较小,那么随着 K 的增加,准确率的下降幅度比较大。例如 chess 数据集,它的项集的期望支持度相对其他的数据集的项集的期望支持度小,所以下降的幅度最大。

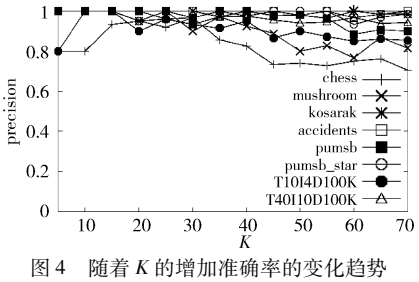

图 4　随着 K 的增加准确率的变化趋势

为了评估隐私保护预算($\varepsilon/2$)对相对错误(RE)的影响,本文设置 $K=30$,$\rho=0.3$。随着隐私保护预算 $\varepsilon/2$ 的增加,相对错误(RE)的变化如图 5 所示。从图 5 中可以看出,随着隐私保护预算($\varepsilon/2$)的增加,相对错误(RE)单调下降并趋近于

0。这是因为根据式(9),随着隐私保护预算($\varepsilon/2$)的增加,加入噪声所服从的拉普拉斯分布的尺度参数会变小,所以噪声值会随着隐私保护预算($\varepsilon/2$)的增加越来越小。对于不同的不确定数据集,如果前 K 个最频繁的频繁项集的期望支持度越小,那么其 RE 的收敛速度越慢。例如数据集 chess 的前 K 个最频繁的项集的期望支持度相对其他数据集最小,所以它的收敛速度越慢。

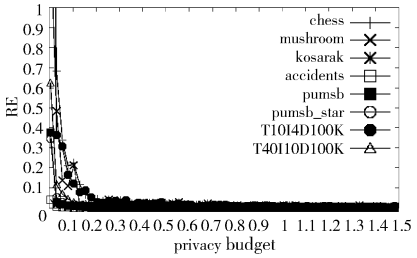

图5 随着隐私保护预算($\varepsilon/2$)的增加相对错误(RE)的变化

5 结束语

为了在不确定的数据中挖掘频繁项集,并利用差分隐私的方法保护用户的隐私数据不被泄露,本文提出了 FIMUDDP 算法。通过理论分析,证明 FIMUDDP 算法满足 ε-差分隐私,并且利用六个真实数据和两个动态生成数据验证了 FIMUDDP 算法的有效性。

参考文献:

[1] Friedman A, Schuster A. Data mining with differential privacy [C]//Proc of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010: 493-502.

[2] Tong Yongxin, Chen Lei, Cheng Yurong, *et al*. Mining frequent itemsets over uncertain databases [J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1650-1661.

[3] Bhaskar R, Laxman S, Smith A, *et al*. Discovering frequent patterns in sensitive data [C]//Proc of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010: 503-512.

[4] Chui C K, Kao Ben, Hung E. Mining frequent itemsets from uncertain data [C]//Proc of the 11th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Berlin: Springer-Verlag, 2007: 47-58.

[5] Han Jiawei, Kamber M, Pei Jian. Data mining: concepts and techniques [M]. 3rd ed. San Francisco: Morgan Kaufmann Publishers Inc, 2011.

[6] Leung C K S, Mateo M A F, Brajczuk D A. A tree-based approach for frequent pattern mining from uncertain data [C]//Proc of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Berlin: Springer-Verlag, 2008: 653-661.

[7] Bernecker T, Kriegel H P, Renz M, *et al*. Probabilistic frequent itemset mining in uncertain databases [C]//Proc of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009: 119-128.

[8] Calders T, Garboni C, Goethals B. Approximation of frequentness probability of itemsets in uncertain data [C]// Proc of IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2010: 749-754.

[9] Peterson E A, Tang Peiyi. Fast approximation of probabilistic frequent closed itemsets [C]//Proc of the 50th Annual Southeast Regional Conference. New York: ACM Press, 2012: 214-219.

[10] Tang Peiyi, Peterson E A. Mining probabilistic frequent closed itemsets in uncertain databases [C]//Proc of the 49th Annual Southeast Regional Conference. New York: ACM Press, 2011: 86-91.

[11] Tong Yongxin, Chen Lei, Ding Bolin. Discovering threshold-based frequent closed itemsets over probabilistic data [C]//Proc of the 28th International Conference on Data Engineering. Washington DC: IEEE Computer Society, 2012: 270-281.

[12] Lin C W J, Gan Wensheng, Fournier-Viger P, *et al*. Weighted frequent itemset mining over uncertain databases [J]. Applied Intelligence, 2016, 44(1): 232-250.

[13] Ahmed A U, Ahmed C F, Samiullah M, *et al*. Mining interesting patterns from uncertain databases [J]. Information Sciences, 2016, 354(8): 60-85.

[14] Lin C W J, Gan Wensheng, Fournier-Viger P, *et al*. Efficient algorithms for mining high-utility itemsets in uncertain databases [J]. Knowledge-Based Systems, 2016, 96(3): 171-187.

[15] Dwork C, McSherry F, Nissim K, *et al*. Calibrating noise to sensitivity in private data analysis [C]//Proc of the 3rd Conference on Theory of Cryptography. Berlin: Springer-Verlag, 2006: 265-284.

[16] McSherry F, Talwar K. Mechanism design via differential privacy [C]//Proc of the 48th Annual IEEE Symposium on Foundations of Computer Science. Washington DC: IEEE Computer Society, 2007: 94-103.

[17] Li Ninghui, Qardaji W, Su Dong, *et al*. PrivBasis: frequent itemset mining with differential privacy [J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1340-1351.

[18] Su Sen, Xu Shengzhi, Cheng Xiang, *et al*. Differentially private frequent itemset mining via transaction splitting [J]. IEEE Trans on Knowledge and Data Engineering, 2015, 27(7): 1875-1891.

[19] Maruseac M, Ghinita G. Differentially-private mining of moderately-frequent high-confidence association rules [C] //Proc of the 5th ACM Conference on Data and Application Security and Privacy. New York: ACM Press, 2015: 13-24.

[20] Xu Shengzhi, Cheng Xiang, Su Sen, *et al*. Differentially private frequent sequence mining [J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28(11): 2910-2926.

(上接第 1941 页)

[6] Li Ruohui. The exploration of modes for college design workshops based on cloud platform [C]//Proc of the 8th International Conference on Measuring Technology and Mechatronics Automation. Washington DC: IEEE Computer Society, 2016: 778-781.

[7] 郑镁,罗磊,江平宇. 基于语义 Web 的云设计服务平台及关键技术[J]. 计算机集成制造系统, 2012, 18(7): 1426-1434.

[8] Valilai O F, Houshmand M. A collaborative and integrated platform to support distributed manufacturing system using a service-oriented approach based on cloud computing paradigm [J]. Robotics and Computer-Integrated Manufacturing, 2013, 29(1): 110-127.

[9] Weinhardt C, Anandasivam A, Blau B, *et al*. Cloud computing: a classification, business models, and research directions [J]. Business & Information Systems Engineering, 2009, 1(5): 391-399.

[10] 尹翰坤, 尹超, 龚小容, 等. 汽摩零部件新产品开发云制造平台总体框架及关键技术[J]. 计算机集成制造系统, 2013, 19(9): 2332-2339.

[11] Chen Jian, Mo Rong, Wu Linjian, *et al*. A method of collaborative task allocation for cloud Service platform of industrial design [C]// Proc of the 8th International Conference on Intelligent Human-Machine Systems and Cybernetics. Piscataway, NJ: IEEE Press, 2016: 484-487.

[12] 潘浩. 联盟能力、产业链整合与战略性新兴产业发展的关系研究[D]. 杭州: 浙江工业大学, 2013.

[13] 刘金彪. 面向设计制造服务的业务协同平台研究与应用[D]. 重庆: 重庆大学, 2015.

[14] 宋庭新, 张成雷, 李成海, 等. 中小企业云制造服务平台的研究与开发[J]. 计算机集成制造系统, 2013, 19(5): 1147-1154.

[15] 杜方, 陈跃国, 杜小勇. RDF 数据查询处理技术综述[J]. 软件学报, 2013, 24(6): 1222-1242.

[16] 徐易. 智能答疑系统关键技术的研究与实现[D]. 南京: 东南大学, 2003.

[17] 刘群, 李素建. 基于知网的词汇语义相似度计算[C]//第三届汉语词汇语义学研讨会论文集. 2002: 59-76.