基于差分隐私的多源数据关联规则挖掘方法

崔一辉 宋 伟 彭智勇 杨先娣
(武汉大学计算机学院 武汉 430072)

摘 要 随着大数据时代的到来,挖掘大数据的潜在价值越来越受到学术界和工业界的关注。但与此同时,由于互联网安全事件频发,用户越来越多地关注个人隐私数据的泄露问题,用户数据的安全问题成为阻碍大数据分析的首要问题之一。关于用户数据的安全性问题,现有研究更多地关注访问控制、密文检索和结果验证,虽然可以保证用户数据本身的安全性,但是无法挖掘出所保护数据的潜在价值。如何既能保护用户的数据安全又能挖掘数据的潜在价值,是亟需解决的关键问题之一。文中提出了一种基于差分隐私保护的关联规则挖掘方法,数据拥有者使用拉普拉斯机制和指数机制在数据发布的过程中对用户数据进行保护,数据分析者在差分隐私的FP-tree上进行关联规则挖掘。其中的安全性假设是:攻击者即使掌握了除攻击目标以外的所有元组数据信息的背景知识,仍旧无法获得攻击目标的信息,因此具有极高的安全性。所提方法是兼顾安全性、性能和准确性,以牺牲部分精确率为代价,大幅增加了用户数据的安全性和处理性能。实验结果表明,所提方法的精确性损失在可接受的范围内,性能优于已有算法的性能。

关键词 隐私保护的数据挖掘,差分隐私,拉普拉斯机制,指数机制
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.06.006

Mining Method of Association Rules Based on Differential Privacy

CUI Yi-hui SONG Wei PENG Zhi-yong YANG Xian-di
(Computer School, Wuhan University, Wuhan 430072, China)

Abstract With the advent of the era big data, the potential value of mining big data has attracted more and more attention from academia and industry. However, at the same time, due to frequent Internet security incidents, users are increasingly concerned about the disclosure of personal privacy data, and user data security issues become one of the most important obstacles to big data analysis. With regard to the study of user data security, the existing researches more focus on access control, ciphertext retrieval and result verification. The above researches can guarantee the security of user data itself, but can not dig out the potential value of protected data. Therefore, how to protect the security and dig the potential value of the data in the meantime is one of the key issues that need to be addressed. This paper proposed an association rules mining method based on differential privacy protection. Data owners use Laplacian mechanism and exponential mechanism to protect user data during data release. Data analysis is associated with differential privacy FP-tree Rule mining. The experimental results show that the performance and accuracy of the proposed method are superior to the existing methods.

Keywords Privacy preserving data mining, Differential privacy, Laplace mechanism, Exponential mechanism

1 引言

隐私保护的数据挖掘在保护所收集的用户数据的前提下专注于挖掘大数据潜在的隐含价值。随着大数据挖掘的发展,安全事件频发,导致用户数据被泄露,因此该研究具有很重要的现实意义。隐私保护的数据挖掘有着广泛的应用,如银行金融领域、网络舆情预警领域、生物医疗领域等。

本文以医疗机构为例进行研究。对于医疗机构而言,由于其本身并不具备关联规则挖掘的能力,往往需要将数据发布给专业的数据分析人员进行分析,因此会带来病患隐私信息的泄露问题。病患隐私信息的泄露会给病患带来很多意想不到的麻烦,例如乔布斯患癌的消息造成了苹果股价的大跌。如何在保证用户隐私的条件下进行关联规则的挖掘是亟须解决的问题。以医疗行业数据为例,进行隐私保护的关联规则

到稿日期:2017-03-11 返修日期:2017-05-19 本文受国家自然科学基金(61232002,61572378,61202034),湖南省自然科学基金面上项目(2017CFB420),CCF中文信息技术开放课题基金(CCF2014-01-02),武汉市创新团队项目(2014070504020237),武汉大学自主科研项目(2042016gf0020,2016-2017)资助。

崔一辉(1981-),男,博士生,主要研究方向为可信数据管理;**宋 伟**(1978-),男,博士,副教授,主要研究方向为可信数据管理,E-mail:songwei@whu.edu.cn(通信作者);**彭智勇**(1963-),男,博士,教授,主要研究方向为海量数据管理;**杨先娣**(1974-),女,博士,副教授,主要研究方向为社区数据管理。

挖掘。表 1 列出了 HELQ 基因缺失与癌症的关联规则。

表 1　医疗机构隐私保护数据的挖掘

Table 1　Data mining of privacy preserving in medical institutions

TID	基因 HELQ	属性 2	Disease
1000	缺失	* * * *	cancer
2000	缺失	* * * *	cancer
3000	缺失	* * * *	cold
4000	正常	* * * *	cold
…	…	…	…

据此，数据分析者把分析的关联规则返回给医疗机构，医疗机构可以对拥有该基因的病人进行预警，相应的病人可以增加癌症检测的次数，进而在早期发现癌症并及时治疗，以提升存活率。事实上，关联规则的挖掘需要整合大数据资源，而现实中医院往往针对自己拥有的局部数据进行挖掘，不能得到完整的结果，因此召回率较高。本文致力于解决多数据源的隐私保护数据挖掘问题（见图 1），在保证隐私的前提条件下整合多家机构的数据资源，进而降低召回率。

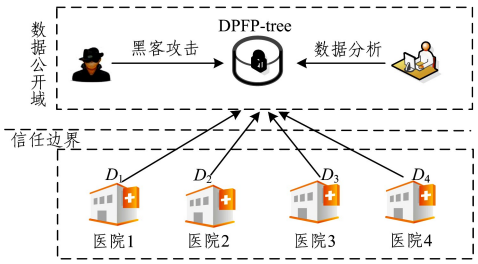

图 1　多数据拥有者的隐私保护数据挖掘问题

Fig. 1　Data mining problem of privacy preserving for multiple data owners

本文的主要贡献如下：

1）提出了一种基于差分隐私的数据挖掘方法。数据为数值型时采用拉普拉斯机制，数据为顺序的字符型时采用指数机制，一次性将关联规则挖掘所需要的信息发布成为差分隐私 F-tree。该方式为非交互式方式，不需要数据拥有者一直在线。

2）数据发布采用 FP-tree 方式，FP-tree 保留了支持度信息和频繁项组合信息；在此基础上设计了从 FP-tree 恢复项集的递归算法，没有对原始数据集进行整体发布，因此具有空间占用率低的优点。

3）与数据外包挖掘模式和分布式挖掘模式相比，所提出的基于差分隐私的关联规则挖掘方法具有处理效率高的优势。这是因为数据外包和分布式挖掘大多都需要对数据进行加密处理，而差分隐私的处理开销小于加密处理的开销。

4）本文针对多数据源的数据进行挖掘，设计了 FP-tree 的数据融合算法，具有比单数据源的数据挖掘更低的召回率。

2　相关工作

随着大数据的发展，越来越多的学者开始关注大数据潜在价值的挖掘和商业应用[2-4]。Agrawal 在 SIGMOD 2000 会议上首次提出了隐私保护的数据挖掘问题[1]。近年来，由于用户隐私数据泄漏问题高发，隐私保护的数据挖掘问题成为热点研究领域之一[6-10,16-18]。隐私保护的关联规则挖掘的相关研究较多，按照数据处理的隐私保护模式可以分为 3 种：分布式处理模式、数据外包处理模式和基于隐私保护数据发布的处理模式[6-10]。

分布式处理模式是指多个数据拥有者在不泄露自己所拥有的数据的同时，进行协作关联规则的挖掘。Wong 提出了一种基于加法同态的安全多方计算方法，该方法可以进行频繁项的挖掘，进而实现关联规则的挖掘[11]。Nanavati 提出了一种针对水平分隔的对称加密的同态算法来进行隐私保护的求和运算，并在此基础上进行频繁项的挖掘。安全多方计算往往要求挖掘的多方同时在线来协同进行数据分析，在其过程中可以隐藏本机构的数据。该算法的优点在于安全性较高，但受网络传输的影响较大[12]。

数据外包处理模式是指数据存储外包给云服务提供商，用户与云服务提供商交互对数据进行挖掘。Giannotti 提出了基于加密的外包数据隐私保护关联规则挖掘方法，其设计的算法保证了 k-匿名，但数据发布针对原始数据集进行，处理开销较大[15]。

基于隐私保护进行数据发布的处理模式是指数据拥有者对数据进行隐私保护处理后发布给第三方，由第三方进行数据挖掘。差分隐私的提出使得隐私保护的数据发布得到了快速的发展。Johnson 提出了一种针对基因数据的关联规则挖掘方法来解决现有基因关联规则大多针对给定基因的问题。该方法的不足在于它是一种交互式的差分隐私方法，灵活性较低，且受制于网络带宽，因此挖掘效率较低[5]。

针对上述分布式处理模式安全多方计算过程中加密处理开销大、外包模式因对整体数据集进行处理而存储开销大，以及现有的数据发布方法因使用交互式差分隐私而通信开销大的问题，本文提出了一种基于差分隐私的非交互式数据发布方法。所提方法不仅解决了用户数据在发布过程中的隐私泄露问题，避免了加密的时间开销和存储开销，也解决了交互式隐私保护的通信开销以及需要数据源长期在线的问题。

3　基础知识

本节主要介绍本文研究所涉及的基础知识，特别是差分隐私的两种机制[13,19]：拉普拉斯机制和指数机制。

3.1　基本定义

定义 1（个人身份标示 ID）　个人身份标识 ID 可以唯一确定数据集中个人的身份。

定义 2（准表示符 QI）　准表示符 QI 可以通过连接攻击获取用户身份的一组属性信息。

定义 3（ε-差分隐私）　对于数据集 D 和 D'，它们只有一条记录不同，算法 f 是差分隐私的，需要满足如下条件：

$$\Pr(f(D)=\hat{D})\leqslant e^{\varepsilon}\times\Pr(f(D')=\hat{D})$$

定义 4（敏感度）　对于任意数据集 D 和 D'，它们只有一条记录不同，算法 f 的敏感度被定义为[14]：

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

定义 5（Laplace 机制）　给定数据集 D，设函数 $f:D\to R^d$，其敏感度为 Δf，那么随机算法 $M(D)=f(D)+Y$ 提供 ε-差分隐私保护，其中 $Y\sim\mathrm{Lap}(\Delta f/\varepsilon)$ 服从尺度参数为 $\Delta f/\varepsilon$ 的 Laplace 分布。

定义 6(指数机制) 设随机算法 M 的输入为数据集 D,输出为实体对象 $R\in Range$,$q(D,r)$为可用性函数,Δq 为函数 $q(D,r)$的敏感度,若函数以正比于 $\exp(\varepsilon q(D,r)/2\Delta q)$的概率从 Range 中选择输出 r,那么算法提供 ε-差分隐私保护。

定义 7(FP-tree) 频繁模式树保持数据集中频繁项的信息,通过树中的指针连接相同的元素。

3.2 FP-growth 算法

FP-growth 算法最早将 FP-tree 用于频繁项挖掘。首先,第一轮遍历构建项集的支持度降序序列;然后,进行第二轮遍历,删除支持度低于阈值的元素,并在此基础上构建 FP-tree;最后,在 FP-tree 上进行频繁项挖掘。本文在 FP-growth 算法的基础上,结合差分隐私,最终发布差分隐私的 FP-tree[18]。

在详细描述隐私保护的关联规则挖掘算法之前,先给出一些基本的符号解释和定义,文中使用的符号如表 2 所列。

表 2 符号

Table 2 Notations

符号	描述
D	数据集 $D=\{t_1,t_2,\cdots,t_n\}$
I	属性集 $I=\{i_1,i_2,\cdots,i_n\}$
ε	差分隐私的隐私预算
$sensitivity$	数据集上算法 f 的敏感度
$Supp(X)$	支持度
$Con()$	置信度

4 基于差分隐私的 FP-tree 数据发布方法

在发布隐私数据时,FP-tree 中项集的顺序和支持度都可能泄露用户隐私。例如,数据集中没有某病患时支持度为 100,当数据集中有某病患时支持度为 101,由此攻击者就可以得到增加的这个病患的信息。因此,数据发布过程中必须对病患进行基因支持度保护。与此同时,项集的降序也会暴露用户隐私,特别是可用性函数值比较相近的元组,若增加一个病患数据,项集的降序发生了变化,就可以获取病患的信息,因此还需要对数据发布之后 FP-tree 中项集的支持度降序进行保护。差分隐私可以保证攻击者在掌握了除病患本人其他所有病患的背景知识信息的情况下,仍旧不会泄露目标病患的信息,因此具有很强的安全性。

本文在 Han 等提出的 FP-growth[18] 的基础上进行数据处理。差分隐私 FP-tree 发布总体上需要经历 3 步,原始数据集如表 3 所列。

表 3 原始数据

Table 3 Initial data

TID	Item
100	$\{f,a,c,d,g,i,m,p\}$
200	$\{a,b,c,f,l,m,c\}$
300	$\{b,f,h,j,o\}$
400	$\{b,c,k,s,p\}$
500	$\{a,f,c,e,l,p,m,n\}$
…	…

第 1 步 对原始数据集进行第一轮遍历,生成频繁项降序数据集序列。

第 2 步 对数据集进行第二轮遍历,删除支持度低于阈值的元素,并对每一项进行排序存储,结果如表 4 所列。由于 f 出现的频率最高,因此在每个元组中 f 都排在最前面。在此基础上生成 FP-tree,如图 2 所示。

表 4 删除支持度低于阈值属性的元素后的结果

Table 4 Element results after deleting attributes whose support degree is below threshold

TID	Ordered Item
100	$\{f,c,a,m,p\}$
200	$\{f,c,a,b,m\}$
300	$\{f,b\}$
400	$\{c,b,p\}$
500	$\{f,c,a,m,p\}$
…	…

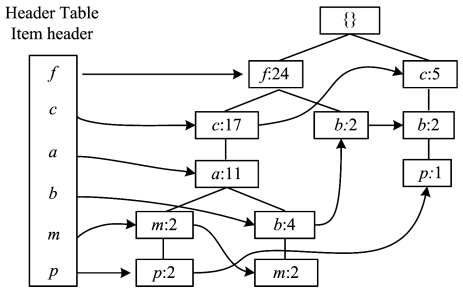

图 2 加噪音之前的 FP-tree

Fig. 2 FP-tree before adding noise

图 2 中的 Header Table 数据结构存储了数据集中的频繁项信息及频繁项的组合信息。原始 FP-tree 的支持度和 Head Table 的属性排序均会泄露用户的隐私。

第 3 步 计算数据集的敏感度并选择隐私预算,隐私预算越小则发布数据隐私保护的程度越大,数据的可用性越差。对发布的 FP-tree 进行差分隐私保护,最终生成差分隐私 FP-tree 发布或共享,如图 3 所示。

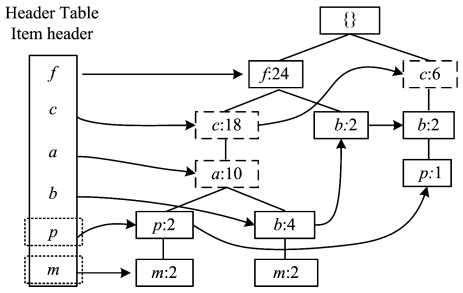

图 3 加噪音之后的 FP-tree

Fig. 3 FP-tree after adding noise

发布数据的支持度计算函数的敏感度为 1,我们选择隐私预算 $\varepsilon=0.01$,其中,指数机制的隐私预算为 $\varepsilon/2$,拉普拉斯机制的隐私预算为 $\varepsilon/2$。使用生成拉普拉斯噪音的 lapace 变量的计算式为:

$$X=\mu-b*\mathrm{sgn}(U)\ln(1-2|U|)$$

其中,变量 U 为已知区间 $(-1/2,1/2]$ 中的均匀分布。经过变换,虚线方框属性的支持度发生了变化,如 c 和 a;其他的属性没有发生变化。由于噪音是由随机算法生成的,因此同一 FP-tree 多次发布时变化的属性会不同。与此同时,需要对 Head Table 的项集顺序进行指数机制的差分隐私,可用性函数为支持度计算函数,可用性函数的敏感度 $\Delta q=1$。经过指数机制的混淆,p 和 m 的顺序发生了改变,从而保护了顺

序隐私。由于存在拉普拉斯噪音，攻击者无法确认是否交换了顺序。指数机制处理过程中，将完整的项集序列划分成若干数据块 $D_1, D_2, \cdots, D_n$，然后在其上进行指数机制混淆，而不是在全局进行混淆，进而在保证安全的同时最大可能地保证数据的可用性。

基于差分隐私的 FP-tree 数据发布算法如算法 1 所示。

算法 1　基于差分隐私的 FP-tree 数据发布算法

输入：原始数据 D

输出：差分隐私处理后的差分隐私 FP-tree$_{DF}$

1. FList＝scanOnceForFList(D);/＊ 扫描数据集生成支持度降序的频繁序列 ＊/
2. FP-tree＝buildFP-tree(FList,D);
3. Sentivity＝computeSentivity(D);
4. ε＝choosePrivacyBudget();/＊ 选择隐私预算 ＊/
5. for each node
6. noise＝conputeLaplace(sentivity/);/＊ 为每个 FP-tree 节点计算拉普拉斯噪音 ＊/
7. weight＝weight＋noise;
8. end for
9. expoMech();使用指数机制混淆顺序;
10. return FP-treeDF;

5　隐私保护的规则挖掘方法

数据拥有者将进行隐私保护的数据发布，分析者可以收集来自多个数据拥有者的 FP-treeDF，在此基础上构建全局的差分隐私 FP-tree。构建过程为：首先在 FP-tree 的基础上重新构建项集支持度降序集，然后根据新的降序集构建排序元组，将重复的元组作为一个元组。事实上，部分属性集出现的次数和实际属性值相差 1，这是拉普拉斯噪音造成的。由于在数据发布过程中拉普拉斯噪音的绝对值不超过 1，因此该差值对关联规则挖掘结果的影响不大。转换算法为：从根节点递归遍历所有的孩子节点，直到遍历的节点为叶子节点，输出根节点到叶子节点的路径以及叶子节点的支持度，删除叶子节点，并将从根节点到叶子节点的支持度依次减去叶子节点的支持度。

数据分析者在得到每个机构的 FP-tree(见图 3)后，调用算法 2 来进行转换。首先遍历得到$\{f,c,a,m,p\}$的支持度是 2，因此将从根节点到叶子节点的权重减 2，并删除支持度为 0 的节点；然后得到$\{f,c,a,b,m\}$的支持度是 2，再将从根节点到叶子节点的支持度减 2，并删除权重为 0 的节点，此时 b 的支持度为 2，因此$\{f,c,a,b\}$的支持度为 2。递归执行算法，最终得到如表 5 所列的属性集。

算法 2　由 FP-tree 生成属性集的算法 transfer()

输入：FP-treeDF

输出：全局模式下数据集中的组合以及支持度

1. 初始化遍历从根节点开始
2. for each child
3. if(child is leaf)/＊ 如果遍历到叶子节点 ＊/
4. tList.add(root_to_leaf);)/＊ 将根节点到叶子节点的路径和叶子节点的支持度记录下来 ＊/
5. delete(leaf);
6. subtract(support);
7. end if
8. else
9. FP-tree_transfer();/＊ 递归调用算法本身 ＊/
10. end for
11. return 项集组合和支持度;

表 5　转换 FP-tree 得到的属性集

Table 5　Attribute sets after converting FP-tree

TID	Ordered Item	Count
1001	$\{f,c,a,m,p\}$	2
2001	$\{f,c,a,b,m\}$	2
3001	$\{f,c,a,b\}$	2
4001	$\{f,c\}$	8
5001	$\{f,c\}$	2
6001	$\{f\}$	4
…	…	…

将表 5 所列的属性集合并成一个大的属性集，执行 FP-tree 生成算法，最终得到多方数据整合之后的全局 FP-tree，如图 4 所示。

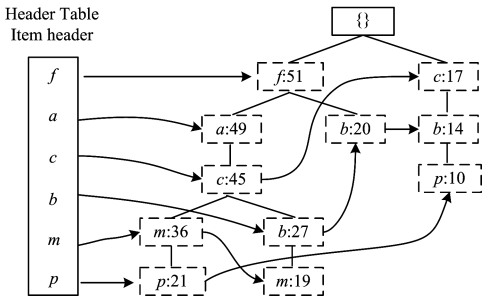

图 4　多方数据整合之后的全局 FP-tree

Fig. 4　Global FP-tree after multi-party data integration

发布的最终数据由于存在拉普拉斯噪音，因此取值为真实的频繁值和敏感度之和，由此会带来部分假阳性率，但由于支持度往往比较大，因此在关联规则的挖掘过程中，该敏感度和假阳性率可以忽略不计。

基于差分隐私的 FP-tree 关联规则挖掘算法如算法 3 所示。

算法 3　基于差分隐私的 FP-tree 关联规则挖掘算法

输入：全局 FP-treeDF

输出：数据集中的频繁项和关联规则

1. FList＝scanOnceForFList(D);/＊ 扫描数据集生成支持度降序的频繁序列 ＊/
2. FP-tree＝buildFP-tree(FList,D); /＊ 由 FP-tree 生成 ordered Item 算法 ＊/
3. 重新构建全局 fp_tree;
4. 调用 FP-growth 算法计算频繁项集;
5. 计算置信度，求出满足置信的关联规则;
6. return 关联规则

6　隐私保护的规则挖掘方法

6.1　性能分析

在隐私保护的数据发布过程和数据挖掘过程中，本文采用了一次性发布的非交互差分隐私机制，该机制的总体性能比基于加密和基于交互式的差分隐私的性能更优。数据发布

的时间复杂度为 FP-tree 生成的时间复杂度、拉普拉斯机制的时间复杂度和指数机制的时间复杂度的总和，即 $T(n)=O(n+\log_2^n)$。事实上，主要的时间开销为 FP-tree 的生成过程，差分隐私处理的时间复杂度相对较小。数据挖掘的时间复杂度主要包括整合不同机构的 FP-tree 时间复杂度、FP-growth 的时间复杂度和寻找关联规则的时间复杂度，即 $T(n)=O(n+\log_2^n)$。

6.2 安全性分析

性质 1 设有算法 $M_1,M_2,\cdots,M_n$，其隐私保护预算分别为 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$，那么对于同一数据集 D，由这些算法构成的组合算法 $M(M_1(D),M_2(D),\cdots,M_n(D_n))$ 提供的差分隐私保护预算为 $\sum_{i=1}^{n}\varepsilon_i$。

性质 2 设有算法 $M_1,M_2,\cdots,M_n$，其隐私保护预算分别为 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$，那么对于不相交的数据集 $D_1,D_2,\cdots,D_n$，由这些算法构成的组合算法 $M(M_1(D_1),M_2(D_2),\cdots,M_n(D_n))$ 提供的差分隐私保护预算为 $\max(\varepsilon_i)$。

本文算法使用拉普拉斯机制对每个属性增加噪音，每个属性都是部分数据集。根据定理 2，拉普拉斯机制使用了 $\varepsilon/2$ 的隐私预算进行数值隐私保护。利用指数机制对顺序进行扰动的过程中，我们对整个队列进行了分区，然后在每个分区上进行了扰动，由性质 2 可以得出，整个顺序扰动过程中的隐私预算为 $\varepsilon/2$。根据性质 1，首先对同一数据集进行拉普拉斯差分隐私，隐私预算为 $\varepsilon/2$；然后进行指数机制的拉普拉斯隐私保护，隐私预算为 $\varepsilon/2$。因此，本文的方法满足 ε-差分隐私。

7 实验与结果

本节使用本文提出的算法进行了隐私保护的数据发布以及多源数据的融合，并且在 2 个数据集(乳腺癌的威斯康辛州(诊断)数据集[20]和模拟数据集)上对所提算法进行了测试。

7.1 度量标准

隐私保护关联规则挖掘方法的判断标准一般包括数据发布性能、数据挖掘性能和准确率。为了更加直观地衡量隐私保护的数据挖掘对挖掘精度的影响，以明文相同算法挖掘的结果作为参照基准，重新定义了准确率的计算方法。

$$OutPrecision=\frac{|NumCipher|}{|NumPlain|}\times 100\%$$

其中，$NumPlain$ 是明文状态下挖掘出的关联规则数据集的数目，$NumCipher$ 是密文状态下挖掘出的关联规则存在于明文下挖掘出的关联规则集合中的数目。

7.2 数据集

实验环境：Window 7 操作系统；CPU intel B950 2.10 GHz；4.00 GB 内存。

选用加州大学尔湾分校的机器学习数据库中的乳腺癌的威斯康辛州(诊断)数据集作为真实数据[20]，针对真实数据集规模有限的问题，依据癌症真实数据集的维度生成了模拟数据集，并使用模拟数据集进行较大规模的实验。实验中随机选择了 10000，20000，30000 和 40000 条记录规模分别进行实验分析。

7.3 实验和结果

本节通过 3 个实验来验证所提出的隐私保护的关联规则挖掘的可行性和高效性。

实验 1 通过变化数据集的大小(分别在 10000，20000，30000 和 40000 数据集的情况下)来比较数据发布的时间开销。如图 5 所示，显而易见，本文提出的算法具有很好的性能优势。主要原因在于，Giannotti 的算法需要将数据集整体加密，加密算法的时间复杂度较高。本文的差分隐私算法都是明文处理运算，拉普拉斯机制和指数机制的处理过程均为随机数运算，时间复杂度较低；且本文方法不需要对每一条数据进行处理，而是对经过数据压缩的 FP-tree 进行处理，大大降低了时间开销。

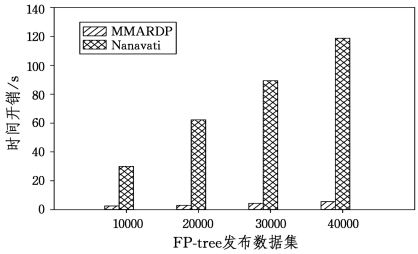

图 5 数据发布过程的性能对比

Fig. 5 Performance comparison of data release process

实验 2 在发布后的 FP-tree 上进行树挖掘。如图 6 所示，与现有的算法相比，本文算法也具有明显的优势。原因在于，分布式多方计算的方法涉及到加密处理和网络传输开销。而本文的方法在挖掘过程中，不管是数据整合的迭代算法，还是 FP-growth 算法，都不包含加密和解密的处理。

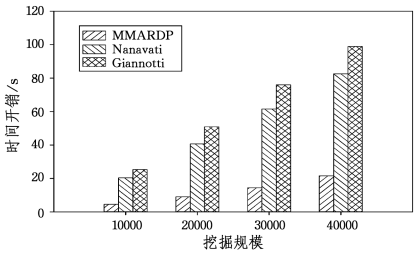

图 6 数据挖掘过程的性能对比

Fig. 6 Performance comparison of data mining process

实验 3 主要进行精确性的验证。本文提出的精确性的衡量指标与明文的衡量指标略有不同，本文提出了以明文挖掘结果为参照系的准确率评价方法，更直观地衡量了隐私保护对数据挖掘的影响。如图 7 所示，由于噪音的存在，与现有隐私保护数据挖掘的方法相比，本文所提方法的精确率略低，但基本处于可接受的范围。因此，本文方法是性能与准确率的一个均衡。

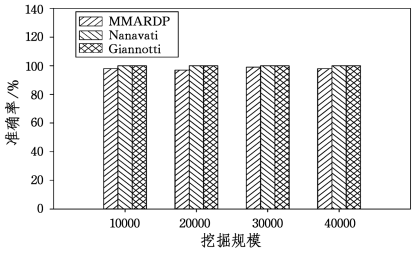

图 7 挖掘结果的精确率对比

Fig. 7 Comparison of precision rate of mining results

(下转第 56 页)