ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.8, August 2006, pp.1764−1774
DOI: 10.1360/jos171764

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

一种有效的隐私保护关联规则挖掘方法*

张 鹏[1], 童云海[1,2+], 唐世渭[1,2], 杨冬青[1], 马秀莉[1,2]

[1](北京大学 信息科学技术学院,北京 100871)

[2](视觉与听觉信息处理国家重点实验室(北京大学),北京 100871)

An Effective Method for Privacy Preserving Association Rule Mining

ZHANG Peng[1], TONG Yun-Hai[1,2+], TANG Shi-Wei[1,2], YANG Dong-Qing[1], MA Xiu-Li[1,2]

[1](School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)

[2](National Laboratory on Machine Perception (Peking University), Beijing 100871, China)

+ Corresponding author: Phn: +86-10-62756920, E-mail: yhtong@pku.edu.cn, http://www.pku.edu.cn

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. http://www.jos.org.cn/1000-9825/17/1764.htm

Abstract: Privacy preservation is one of the most important topics in data mining. The purpose is to discover accurate patterns without precise access to the original data. In order to improve the privacy preservation and mining accuracy, an effective method for privacy preserving association rule mining is presented in this paper. First, a new data preprocessing approach, Randomized Response with Partial Hiding (RRPH) is proposed. In this approach, the two privacy preserving strategies, data perturbation and query restriction, are combined to transform and hide the original data. Then, a privacy preserving association rule mining algorithm based on RRPH is presented. As shown in the theoretical analysis and the experimental results, privacy preserving association rule mining based on RRPH can achieve significant improvements in terms of privacy, accuracy, efficiency, and applicability.

Key words: privacy preservation; data mining; association rule; frequent itemset; randomized response

摘 要: 隐私保护是当前数据挖掘领域中一个十分重要的研究问题,其目标是要在不精确访问真实原始数据的条件下,得到准确的模型和分析结果.为了提高对隐私数据的保护程度和挖掘结果的准确性,提出一种有效的隐私保护关联规则挖掘方法.首先将数据干扰和查询限制这两种隐私保护的基本策略相结合,提出了一种新的数据随机处理方法,即部分隐藏的随机化回答(randomized response with partial hiding,简称 RRPH)方法,以对原始数据进行变换和隐藏.然后以此为基础,针对经过 RRPH 方法处理后的数据,给出了一种简单而又高效的频繁项集生成算法,进而实现了隐私保护的关联规则挖掘.理论分析和实验结果均表明,基于 RRPH 的隐私保护关联规则挖掘方法具有很好的隐私性、准确性、高效性和适用性.

关键词: 隐私保护;数据挖掘;关联规则;频繁项集;随机化回答

中图法分类号: TP311 文献标识码: A

* Supported by the National Natural Science Foundation of China under Grant Nos.60403041, 60473072 (国家自然科学基金)
Received 2005-01-27; Accepted 2006-01-09

随着信息技术,特别是网络技术、数据存储技术和高性能处理器技术的飞速发展,海量数据的收集、管理和分析变得越来越方便,知识发现和数据挖掘更是在一些深层次的应用中发挥了积极的作用.但与此同时,也带来了隐私保护方面的诸多问题.例如,通过对医院病人的病历数据进行挖掘,可以发现各种疾病之间的关联.但在使用一般方法进行挖掘的过程中,不可避免地会使病例数据暴露,从而造成病人隐私的泄露.还有治安系统中的违法记录、保险公司的理赔情况、银行卡客户的交易行为等信息中的关联关系,都对政府和企业决策具有相当重要的意义,但同时又都是公民非常注重的个人隐私.所以,如何在数据挖掘过程中解决好隐私保护的问题,目前已经成为数据挖掘界的一个研究热点[1,2].

首先需要明确的是,可能泄露隐私的并不是数据挖掘技术本身,而是数据挖掘方法的特定应用和具体过程.数据挖掘有一个重要特征,就是从大量数据中挖掘出来的模式或者规则,通常是针对综合数据而非细节数据.那么,我们是否可以基于非精确的原始数据而抽取出准确的模式与规则呢?实现隐私数据的合理保护和基于统计数据的模式抽取两者兼得,正是隐私保护数据挖掘方法研究的出发点和最终目标[3].

(1) 相关工作

目前,隐私保护的数据挖掘方法按照基本策略主要可以分成数据干扰[4−7]和查询限制[8−12]两大类.数据干扰策略就是首先通过数据变换、数据离散化和在数据中增加噪声等方法对原始数据进行干扰,然后再针对经过干扰的数据进行挖掘,得到所需的模式和规则;查询限制策略则是通过数据隐藏、数据抽样和数据划分等方式,避免数据挖掘者拥有完整的原始数据,而后再利用概率统计的方法或者分布式计算的方法得到所需的挖掘结果.但是,这两种策略本身都存在着一些固有的缺陷.在采用数据干扰策略的方法中,所有经过干扰的数据均与真实的原始数据直接相关;而在采用查询限制策略的方法中,所有提供的数据又都是真实的原始数据,这些都会降低方法对隐私数据的保护程度.

关联规则的挖掘作为数据挖掘中最重要的方法之一,也在隐私保护方面取得了一定的研究成果.在数据干扰策略的应用中,文献[4,5]提出了一种基于随机变换[6]的 MASK(mining associations with secrecy konstraints)方法.该方法通过数据干扰和分布重构实现了隐私保护的关联规则挖掘,并且还在随机化参数的设置和支持度的计数方法等方面进行了优化.文献[7]则提出了“部分支持度”的概念以及相应的计算方法,对项集的支持度进行估算,进而实现了隐私保护的关联规则挖掘.但这些方法都是利用了统计学中的沃纳模型,不仅由于变换后的所有数据均与真实的原始数据直接相关,使得对隐私数据的保护程度并不理想,而且随机化参数的选择也都受到限制,必须偏离 0.5[13].

在查询限制的策略方面,文献[8]提出了通过使用“未知”值来替代部分敏感的原始数据,使得敏感规则不被发现的方法.文献[9]也提出了针对特定的敏感规则对原始数据进行隐藏,降低敏感规则支持度,使其不被发现的方法.在这些数据隐藏方法中,虽然一部分敏感信息得到了很好的保护,但由于所提供的所有数据都是真实的原始数据,所以对整个数据集的隐私保护程度并不高.而且,这类方法必须预先知道需要隐藏或者处理的敏感规则,但在通常情况下,具体的规则在挖掘结果出来以前都是未知的.

数据划分是采用查询限制策略的另一类方法,多个数据挖掘的参与者都只拥有一部分原始数据,然后通过协同的分布式计算得到所需的挖掘结果.文献[10,11]分别提出了基于水平划分和垂直划分的隐私保护关联规则挖掘方法,文献[12]还提供了一个支持隐私保护数据挖掘的分布式计算工具集.但这类方法只能用于分布式数据库,所有的数据提供者都必须参与到计算中来,单点故障就会产生错误的结果,甚至造成挖掘无法进行.而且,当数据提供者的数量很多时,这类方法的性能会由于巨大的通信开销而显著下降.

(2) 本文的工作

可以看出,在应用需求的推动下,隐私保护的数据挖掘方法已受到广泛的关注,并取得了相当大的进展,但仍然存在着很多的不足之处以及尚未解决的问题.在国内,隐私保护的数据挖掘还是一个崭新的领域.

为了提高对隐私数据的保护程度,相互弥补两种策略之间的缺陷,本文将数据干扰和查询限制的策略相结合,提出了一种新的基于部分隐藏随机化回答(randomized response with partial hiding,简称 RRPH)的隐私保护关联规则挖掘方法.该方法对随机化参数的选择没有任何限制,也不需要其他额外的信息,而且可以同时适用于

集中式数据库和分布式数据库.更重要的是,理论分析和实验结果均表明:基于 RRPH 的隐私保护关联规则挖掘方法克服了单一策略中固有的缺陷,在相同时间和空间开销的条件下,可以得到比原有方法更好的隐私性和准确性,并且还具有很好的适用性.

本文的主要贡献在于:(1) 将数据干扰和查询限制策略相结合,提出了一种新的数据随机处理与分布重构方法——RRPH;(2) 以 RRPH 方法为基础,实现了数据的变换和隐藏,并针对经过 RRPH 方法处理的数据给出了一种简单、高效的频繁项集生成算法;(3) 从方法的隐私性、准确性、高效性、适用性等方面进行了详细的分析和实验,并与原有方法进行了比较.

本文第 1 节首先给出问题的描述,然后提出解决方法的总体架构.第 2 节介绍 RRPH 方法以及如何利用 RRPH 方法进行数据的变换和隐藏.第 3 节说明如何面向经过 RRPH 方法处理的数据进行挖掘.对方法的详细分析和评价以及与原有方法的比较在第 4 节中给出.第 5 节展示实验结果.第 6 节给出结论与展望.

1 问题与架构

在本节中,我们首先给出问题的描述,然后提出解决方法的总体架构.

1.1 问题描述

设 $I=\{i_1,i_2,...,i_m\}$是一组项的集合,D 为事务 T 的集合,这里事务 T 是项的集合,并且 $T\subseteq I$.设 X 是一个 I 中项的集合,如果 $X\subseteq T$,则称事务 T 包含 X.一个关联规则是形如 $X\Rightarrow Y$ 的蕴涵式,这里 $X\subseteq I,Y\subseteq I$,并且 $X\cap Y=\varnothing$.规则 $X\Rightarrow Y$ 的支持度是指包含 X 和 Y 的事务数与总事务数之比;置信度是指包含 X 和 Y 的事务数与包含 X 的事务数之比.隐私保护的关联规则挖掘,就是要在不精确访问原始事务集 D 的条件下,尽可能准确地产生支持度和置信度分别不低于用户给定阈值的关联规则.

通常,关联规则挖掘的过程分为以下两步:

第 1 步:发现频繁项集.所谓频繁项集是指支持度不低于预先设定阈值的项集.

第 2 步:根据第 1 步发现的频繁项集,产生置信度不低于预先设定阈值的关联规则.

可以看出,第 2 步是在第 1 步结果的基础上进行的,而与原始事务集无关.所以,隐私保护的关联规则挖掘问题可以归结为:在不精确访问原始事务集的条件下,尽量准确地发现其中的频繁项集,用于产生感兴趣的关联规则.

1.2 总体架构

为了解决上述问题,本文将数据干扰和查询限制的隐私保护策略相结合,提出了一种新的数据随机处理方法——RRPH,并以此为基础来实现隐私保护的关联规则挖掘.总体架构如图 1 所示.

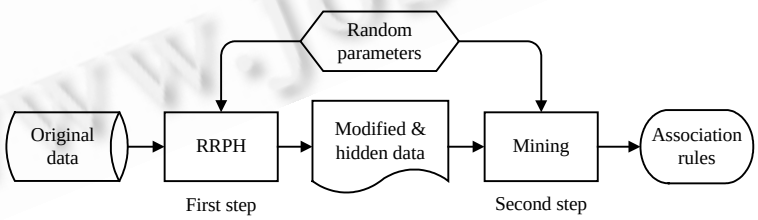

Fig.1 Framework

图 1 总体架构

整个方法分两步进行:第 1 步利用 RRPH 方法对原始数据进行变换和隐藏,具体方法将在第 2 节中给出;第 2 步面向经过 RRPH 方法处理的数据进行挖掘,得到满足支持度条件的频繁项集,进而产生感兴趣的关联规则,详细过程将在第 3 节中加以介绍.在该方法中,连接两步操作的除了经过 RRPH 方法处理的数据以外,还有一组随机化参数.这组参数的作用是在第 1 步操作时描述数据变换和隐藏的规则,在第 2 步操作时指导项集支持度的重构计算.而且,参数值的选择对于隐私保护的程度和挖掘结果的准确性都具有非常重要的影响.

为了更加直观,本文将以购物篮问题为例加以说明.每种商品作为一个项,并拥有一个确定的顺序号,客户的一次购物作为一个事务,并用一个长度为总项数的 0-1 序列来表示:购买了相应序号的商品该位就取值为 1,否则就取值为 0.本文要解决的问题是,在不精确访问原始 0-1 序列事务集的条件下,尽量准确地发现 0-1 序列中不同位上取值同时为 1 的关联程度.需要指出的是,0-1 序列只是一种逻辑结构,并非实际的物理存储,因为事务集通常是非常稀疏的.

2 部分隐藏的随机化回答方法(RRPH)

随机化回答方法的基本思想是:数据提供者依照给定的随机化参数对原始数据进行变换,再提供给数据使用者.在使用者得到的信息中,虽然详细信息被提供者进行了干扰,但在数据量比较大的情况下,统计信息和聚集信息仍然可以被相当精确地估算出来.由于关联规则挖掘是基于一个数据集合的聚集信息值而不是一个详细的数据项,因此,随机化回答的方法可以很好地用于关联规则的挖掘.

但是,现有的隐私保护数据挖掘中使用的随机化回答方法都是单纯地采用数据干扰的策略,而且都是基于沃纳模型建立的,不仅由于所有变换后的数据都与原始数据直接相关而使得隐私保护的程度不高,而且随机化参数的选择也受到很大的限制,必须偏离 0.5.采用查询限制策略的数据隐藏方法,可以克服所有变化后的数据均与原始数据直接相关的不足,但却又存在着所有提供的数据都是真实数据的缺陷.那么,能否将两种策略相结合,使之相互弥补,从而提高对隐私保护的程度呢?正是基于这种想法,本文将数据干扰和查询限制的策略相结合,提出了一种新的随机化回答方法——RRPH,在实施挖掘之前,同时对原始数据进行变换和隐藏,而且对随机化参数的选择也没有任何限制.具体方法如下:

给定随机化参数 $0\le p_1,p_2,p_3\le 1$,且 $p_1+p_2+p_3=1$.

对于项 $x\in\{0,1\}$,设 $r_1=x,r_2=1,r_3=0$,则随机化函数 $r(x)$以 p_j 的概率选择取值为 $r_j,j=1,2,3$.

设项的总数为 k,则对于用 0-1 序列表示的事务 $X=(x_1,x_2,...,x_k)$,干扰后的事务 $Y=(y_1,y_2,...,y_k)$可以通过 $Y=R(X)$计算得到,其中 $y_i=r(x_i)$.也就是说,y_i 以 p_1 的概率取值为 x_i,以 p_2 的概率取值为 1,以 p_3 的概率取值为 0.

这样,对事务集 D 中的每一个事务 X,经过随机化函数 R 的处理得到 Y,而且由于 Y 在形式上仍然是一个与 X 长度相同的 0-1 序列,所以就可以作为一个伪造的事务加入到伪造的事务集 D'中.表面看来,我们仍然只采用了数据干扰的策略来进行数据变换;但实质上,当以 p_2 或 p_3 的概率选择 r_2 或 r_3 来进行随机化回答的时候,相应的事务被完全隐藏起来,这正是查询限制策略的明显体现.对于现有的随机化回答方法,一旦掌握了具体事务的变换方式,就能够完全重构真实的原始数据;而本文提出的方法在最坏情况下也至多泄漏比例为 p_1 的原始数据.通过第 4 节和第 5 节的分析和实验表明:本文所采用的 RRPH 方法将数据干扰和查询限制的策略有机结合,在真实数据比例相同的情况下,通过合理的参数选择可以同时得到比原有方法更高的隐私性和准确性.

原则上,可以对不同的项使用不同的随机化参数.为简单起见,本文中将对所有项使用相同的随机化参数.而且在这种情况下,可以通过优化策略,大幅度降低挖掘算法的时间和空间复杂度,本文第 3 节将给出详细说明.

3 隐私保护的关联规则挖掘算法

事务集 D 经过上述 RRPH 方法的数据变换和隐藏,得到了一个伪造的事务集 D'.本节将介绍针对 D'进行挖掘,生成频繁项集,进而得到感兴趣的关联规则的方法.

在生成频繁项集的过程中,最关键的就是计算出项集的支持度.在下面的讨论中,我们将首先分别介绍 1 项集和 k 项集的支持度计算方法,然后以 Apriori 算法[14]为基础,给出完整的挖掘算法.

3.1 计算1项集的支持度

设 i 是一个项,π表示 D 中 i 的支持度,λ表示 D'中 i 的支持度.

设 D 中的事务 T 经过 RRPH 方法处理,变成 D'中的事务 T',则 T_i 和 T_i' 的取值和对应概率见表 1.

因此,$\lambda=\pi\cdot(p_1+p_2)+(1-\pi)\cdot p_2=\pi p_1+p_2$.于是,

$$\pi=\frac{\lambda-p_2}{p_1} \tag{1}$$

也就是说,我们首先计算出项 i 在 D'中的支持度λ,然后就可以利用式(1)推算出项 i 在 D 中的支持度π.如果π不小于预先给定的支持度阈值,则说明 i 是一个频繁 1 项集.

Table 1 Probabilities of data mapping by RRPH method

表 **1** RRPH 方法的数据映射概率

No.	T_i	T_i'	Probability
1	0	0	p_1+p_3
2	0	1	p_2
3	1	0	p_3
4	1	1	p_1+p_2

3.2 计算k项集的支持度

设 $A=\{i_1,i_2,\ldots,i_k\}$是一个 k 项集,在所有项都使用相同的随机化参数进行处理的情况下,可以利用文献[4]中提出的一些优化策略来降低项集支持度计算的复杂度.这是因为,当所有项都使用相同的随机化参数时,恰好包含 A 中 j 项的 D 中事务经过 RRPH 方法处理,转变成为恰好包含 A 中 i 项的 D'中事务的概率 m_{ij} 是相等的.

$$m_{ij}=\sum_{t=\max(0,i+j-k)}^{\min(i,j)} C_j^t\cdot(p_1+p_2)^t\cdot p_3^{j-t}\cdot C_{k-j}^{i-t}\cdot p_2^{i-t}\cdot(p_1+p_3)^{k-i-j+t}.$$

对于 k 项集 A 和 D 中事务 T,$|T\cap A|$共有 k+1 种可能的取值.我们依次用 $C_0,C_1,\ldots,C_k$ 表示每种取值所代表的事务在 D 中所占的比例.例如,对于一个 3 项集,D 中的所有事务将被划分成{000},{001,010,100},{011,101,110},{111}这 4 类,而 C_2 则是恰好包含 A 中 2 项的事务在 D 中所占的比例.类似地,对于 D'中的事务 T',$|T'\cap A|$也共有 k+1 种可能的取值.我们依次用 C_0', C_1' ,..., C_k' 表示每种取值所代表的事务在 D'中所占的比例.

于是就有 $C'=MC$,其中: $C'=\begin{bmatrix}C_0'\\C_1'\\\vdots\\C_k'\end{bmatrix}$; $C=\begin{bmatrix}C_0\\C_1\\\vdots\\C_k\end{bmatrix}$;$M=[m_{ij}]$是一个$(k+1)\times(k+1)$的矩阵,$m_{ij}$ 表示恰好包含 A 中 j 项的 D 中事务在经过 RRPH 方法处理之后,转变成为恰好包含 A 中 i 项的 D'中事务的概率.

当 M 可逆时,设 $M^{-1}=[a_{ij}]$,$C=M^{-1}C'$,而 C_k 正是我们要计算的 k 项集 A 的支持度.

$$C_k=a_{k,0}C_0'+a_{k,1}C_1'+\ldots+a_{k,k}C_k' \tag{2}$$

可以看出,首先针对 D'计算出 C_j',并利用 M 求解出 $a_{k,j}$,再通过式(2)就能计算出 k 项集 A 的支持度了,计算的时间和空间复杂度为 $O(k)$.

另外,我们还注意到 $C_0'+C_1'+\ldots+C_k'=|D'|=N$,因而在所有的 C_j' 中可以有一项不通过计数来得到.通常情况下, C_0' 的值要大于其他各项,于是可以通过 $C_0'=N-(C_1'+\ldots+C_k')$来计算出 C_0' 的值.

3.3 完整的挖掘算法

有了上述计算公式,我们就可以借助现有的各种频繁项集生成算法来挖掘出感兴趣的关联规则.为了简单起见,本文使用 Apriori 算法,多次扫描事务集,在第 k 次扫描时,通过对候选项集的计数来发现频繁 k 项集.不同之处在于:原算法只需计数包含各候选项集的事务数,即只考虑 0-1 事务序列中候选项集对应项全为 1 的情况;而我们的算法需要考虑 0-1 事务序列中候选项集对应项的各种 0-1 组合,并按照各个事务所包含的候选项集中项的数目来分别计数 $c.count_j(j=1,2,\ldots,k)$,进而计算出支持度.下面给出具体的面向经过 RRPH 方法处理后的数据生成频繁项集的算法.

算法 **1**. 面向经过 RRPH 方法处理的数据生成频繁项集.

输入:RRPH 方法处理后的事务集 D',最小支持度阈值 s.

输出:D 中的频繁项集 L.

1) scan D', for each item $i\in I$ count $i.count$; //对每个项 i 计数

2) $L_1=\{\{i\}|i\in I, ((i.count/N)-p_2)/p_1\geq s\}$;
3) for (k=2;$L_{k-1}\neq\varnothing$;k++) {
4) C_k=*aproiri_gen* (L_{k-1}); //生成候选 k 项集 C_k
5) for each transaction $t\in D'$ {
6) for (j=1; $j\leq k$; j++) {
7) $C_{t,j}$=*partial_subset* (C_k,t,j); //事务 t 中恰好包含 j 项的候选 k 项集
8) for each candidate $c\in C_{t,j}$
9) $c.count_j$++;
10) }
11) }
12) for each candidate $c\in C_k$ {
13) $c.count= a_{k,0}\cdot c.count_0 + a_{k,1}\cdot c.count_1 + ... + a_{k,k}\cdot c.count_k$;
14) $L_k=\{\{c\}|c\in C_k, c.count/N\geq s\}$;
15) }
16) }
17) return $L=\cup_k L_k$

算法 1 与普通的 Aproiri 算法相比,最大的区别在于:步骤 7)中不是生成事务 t 所包含的候选项集,而是分别生成事务 t 中恰好包含 j=1,2,...,k 项的候选项集.针对每个候选项集的计数器,也从原来的 1 个变成了 k 个.所有候选 k 项集真实的支持度都是在第 k 次事务集扫描结束以后集中计算的,因为此时一个候选 k 项集 c 所需的各个计数器取值 $c.count_j$ (j=1,2,...,k)才能最终全部得到,并用于计算出支持度计数 $c.count$.

4 分析与评价

对于隐私保护的数据挖掘方法,目前尚没有一个标准的评价体系.本节将从隐私性、准确性、高效性、适用性这 4 个方面对本文提出的基于 RRPH 的隐私保护关联规则挖掘方法进行详细的分析和评价,并与原有的 MASK 方法[4]进行对比.

4.1 隐私性

顾名思义,隐私保护数据挖掘方法研究的出发点和最终目标,就是要在合理保护隐私数据的前提下进行数据挖掘和知识发现,寻找其中潜在的有用的模式与规则.因而,隐私性程度的高低就成为评价一种方法好坏的首要因素.

在本文分两步进行的问题解决架构中,隐私性主要是在第 1 步,即 RRPH 方法的数据变换和隐藏中实现的.我们突破原有处理方法仅采用的单一策略的模式,将数据干扰和查询限制的策略有机结合,提出了一种新的随机处理方法,在对数据进行随机变换的同时,又将相当一部分原始数据隐藏了起来.这样,一方面克服了原有随机变换的处理方法中所有变化后的数据均与原始数据直接相关的不足,另一方面,又弥补了原有数据隐藏的处理方法中所有提供的都是真实原始数据的缺陷,从而极大地提高了对隐私保护的程度.

为了更好地比较不同方法对隐私的保护程度,我们还提出了一个量化指标——隐私破坏系数 Breach,其定义如下:

$$Breach=P_{\text{真实数据的比例}}\times P_{\text{真实数据被识别出的概率}}+P_{\text{非真实数据的比例}}\times P_{\text{非真实数据被识别出的概率}}\times P_{\text{非真实数据被还原的概率}}.$$

需要说明的是:真实数据一旦被识别出来,隐私就已经遭到了破坏;非真实数据对隐私的破坏不仅要被正确地识别出来,而且还要经过正确的还原.

下面,我们将针对简单隐藏的方法、MASK 方法和 RRPH 方法,分别计算隐私破坏系数.假设真实的原始数据在各种方法中所占的比例 p 都相同,则有:

(1) 简单隐藏的方法,提供比例为 p 的真实数据.$Breach_1=p\cdot 1=p$.

(2) MASK 方法,随机化参数即为 p.$Breach_2=p\cdot p+(1-p)\cdot(1-p)\cdot 1=p^2+(1-p)^2$.

(3) RRPH 方法,随机化参数 $p_1=p$.我们取 $p_2=p_3=\frac{1-p_1}{2}$,因为这样,非真实数据取值为 0 和 1 的概率是完全相等的,无法被还原;否则,非真实数据就有被识别并还原的可能.例如,若取 $p_2=1-p_1,p_3=0$,则在经过处理的数据中,所有取值为 0 的数据均为真实数据,从而会使得对隐私的保护程度大大降低.在实际应用中,这种 0,1 取值均分的方法既简便又有利于保护隐私.此时,$Breach_3=p_1\cdot\frac{p_1}{p_1+p_2}=\frac{2p^2}{p+1}$.

显然,$Breach_1>Breach_3$;而 $Breach_2$ 和 $Breach_3$ 的关系与 p 的取值有关.

$$\Delta_1=Breach_2-Breach_3=\frac{2p^3-2p^2-p+1}{p+1}=\frac{(\sqrt{2}p+1)(\sqrt{2}p-1)(p-1)}{p+1}.$$

于是,当 $0<p<\frac{1}{\sqrt{2}}$ 时,$Breach_2>Breach_3$.而出于隐私保护的考虑,真实的原始数据比例不宜过高.所以,这也是比较理想的随机化参数选择范围.

因此,本文提出的基于 RRPH 的隐私保护关联规则挖掘方法与简单隐藏的方法和 MASK 方法相比,具有更好的隐私性.

4.2 准确性

保护好数据的隐私,是隐私保护数据挖掘方法最基本的要求.但我们的最终目标是要通过挖掘来获取真实、可用的知识与规则.因此,在隐私受到合理保护的前提下,所采取的方法还必须能够得到尽量准确的挖掘结果.

粗想起来,隐私性和准确性似乎是一对矛盾,隐私性的提高势必造成准确性的下降;而要提高挖掘结果的准确性就必定要以牺牲一定的隐私性为代价.然而,事实并非如此.前面的分析已经表明,本文提出的 RRPH 方法与原有的 MASK 方法相比,具有更好的隐私性.下面,我们将通过方差分析来进一步说明:只要参数选择合理,RRPH 方法在准确性方面也同样会优于 MASK 方法.

下面,我们将针对项 i 支持度估计的方差来比较 RRPH 方法和 MASK 方法.设 π 为项 i 在原始事务集中的支持度,λ 为项 i 在经过处理的事务集中的支持度.同样假设真实数据所占的比例 p 相同,则

(1) MASK 方法,$\hat{\pi}_1=\frac{\lambda_1-(1-p)}{2p-1}$,方差 $Var(\hat{\pi}_1)=\frac{p(1-p)}{n(2p-1)^2}$,其中 $p\neq\frac{1}{2}$.

(2) RRPH 方法,$\hat{\pi}_2=\frac{\lambda_2-p_2}{p_1}$,方差 $Var(\hat{\pi}_2)=\frac{p_1(1-p_1)\pi}{np_1^2}+\frac{p_2(1-p_2)-2\pi p_1p_2}{np_1^2}$.

由于 $p_1=p$,且同样取 $p_2=p_3=\frac{1-p_1}{2}$,则 $Var(\hat{\pi}_2)=\frac{p_2(1-p_2)}{np_1^2}=\frac{(1-p)(1+p)}{4np^2}$.

$\hat{\pi}_1$ 和 $\hat{\pi}_2$ 都是 π 的极大似然无偏估计量,而

$$\Delta_2=Var(\hat{\pi}_1)-Var(\hat{\pi}_2)=\frac{1-p}{n}\left[\frac{p}{(2p-1)^2}-\frac{1+p}{4p^2}\right]=\frac{(1-p)(3p-1)}{4np^2(2p-1)^2},\ p\neq\frac{1}{2},$$

所以,当 $\frac{1}{3}<p<1$ 时,$Var(\hat{\pi}_1)>Var(\hat{\pi}_2)$.

再结合前面隐私性的分析可以得到:当 $\frac{1}{3}<p<\frac{1}{\sqrt{2}}$ 时,本文提出的基于 RRPH 的隐私保护关联规则挖掘方法与 MASK 方法相比,同时具有更好的隐私性和准确性,而这也正是实际应用中比较理想的随机化参数选择范围.另外,MASK 方法还有 $p\neq\frac{1}{2}$ 的限制,这一点也在 RRPH 方法中得到了消除.

4.3 高效性

本文提出的基于 RRPH 的隐私保护关联规则挖掘方法,在提高隐私性和准确性特征的同时并没有带来计算复杂度的增长,其时间和空间代价与 MASK 方法基本相同,而且具体的操作还更为简便.

设事务集的大小为 N,每个事务中平均包含 n 个项,则在第 1 步利用 RRPH 方法进行数据变换和隐藏的过程中,与 MASK 方法一样,都要分别对原始事务进行随机化处理,时间复杂度为 $O(Nn)$,但 RRPH 方法对于每一个项都是根据随机化参数直接给出处理结果,并不需要像 MASK 方法那样,在原始数据的基础上进行计算.

在第 2 步挖掘的过程中,第 1 次扫描事务集的时间复杂度为 $O(N)$,但与 MASK 方法相比,计算一个项支持度的运算更加简单;第 2 次扫描时,由于通常候选 2 项集的数量是最多的,所以,可以利用文献[4,5]中的策略进行算法的优化,设频繁 1 项集的个数为 c_1,每个事务中平均包含 c 个频繁 1 项集,则时间复杂度为 $O(Ncc_1)$;第 $k(k\geq 3)$ 次扫描时,设候选 k 项集的个数为 m_k,则时间复杂度为 $O(Nm_k)$.在空间方面,普通的频繁项集生成算法只需要为每个候选项集设置 1 个计数器来计算支持度;而本文的方法与 MASK 方法一样,都需要为每个候选 k 项集设置 k 个计数器来计算支持度.

4.4 适用性

如前所述,本文提出的基于 RRPH 的隐私保护关联规则挖掘方法在同等的时间和空间开销条件下,获得了比原有方法更好的隐私性和准确性.不但如此,该方法还具有很好的适用性.下面我们将从挖掘方法的适用性和数据类型的适用性这两方面分别加以说明.

虽然本文的 RRPH 方法是为了解决关联规则挖掘中的隐私保护问题而提出的,但由于最先设计的总体架构中将整个方法分成了相对独立的两步,所以当我们希望将其用于分类等其他挖掘方法时,第 1 步的 RRPH 处理方法可以直接使用,只需要在第 2 步的挖掘过程中使用面向经过 RRPH 方法处理的数据的相应方法就可以了.

另外,现有的随机处理方法所针对的都是数值类型或者布尔类型的数据,都不适用于分类属性类型的数据.而本文提出的 RRPH 方法可以通过增加随机化参数和随机变换结果取值的数量来支持对分类属性类型数据的处理,这也是我们目前正在进行的工作之一.

综上所述,本文提出的基于 RRPH 的隐私保护关联规则挖掘方法在隐私性、准确性、高效性、适用性等方面都取得了良好的效果.特别是在随机化参数选取适当的情况下,让隐私性和准确性这对看似矛盾的指标,与原有的随机处理方法相比,同时得到了提高.

5 实验结果

在本节中,我们将通过一组实验来对比 RRPH 方法与 MASK 方法在进行隐私保护关联规则挖掘时的效果,并说明数据隐私性和挖掘结果准确性与随机化参数之间的关系.

5.1 实验方法

我们在实验中使用的数据是由 IBM Almaden 生成器得到的事务集 D,参数为 T10,I4,D100k,N100,即每个事务平均包含 10 个项,频繁项集的平均长度为 4,总事务数为 100 000,总项数为 100.

设 F 是 D 中所有频繁项集的集合,在隐私保护的关联规则挖掘中,F'是从 D'中挖掘出来的所有频繁项集的集合,则这种方法的项集误差 $IE=\frac{|F'-F|}{|F|}$.又设 f 是 F 中的一个频繁项集,且它的实际支持度为 s_f,估算出的支持度为 s'_f,则项集 f 的支持度误差 $SE_f=\frac{|s'_f-s_f|}{s_f}$.这种方法总的支持度误差为 $SE=\frac{1}{|F|}\sum_f SE_f$.

在实验中,我们将选取不同的随机化参数 p_1=p=0.1,0.2,0.3,0.35,0.4,0.45,0.49,0.51,0.55,0.6,0.65,0.7,0.8,0.9,p_2=p_3=$\frac{1-p_1}{2}$,并针对不同的最小支持度阈值 s=1‰,2‰,3‰,4‰,5‰,6‰,7‰,8‰,9‰,10‰,分别使用 RRPH 方法

和 MASK 方法生成频繁项集,同时计算出这些频繁项集的支持度,进而得到它们的项集误差和支持度误差,然后进行对比分析.

5.2 实验结果

图 2 给出了 RRPH 方法和 MASK 方法在最小支持度阈值 s=1‰,2‰,3‰,4‰,5‰,6‰,7‰,8‰,9‰,10‰时,平均项集误差随参数 p 变化的情况.

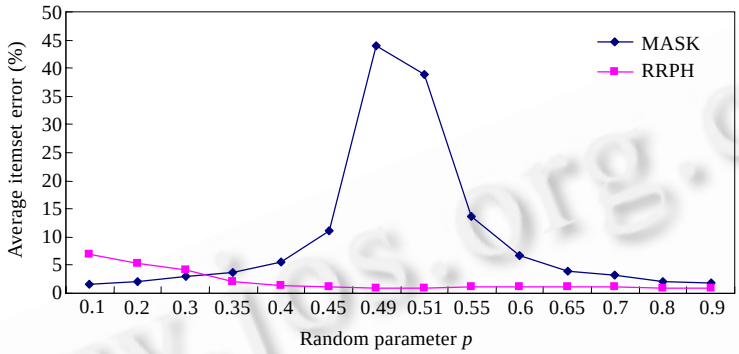

Fig.2 Average itemset error of RRPH method and MASK method

图 2 RRPH 方法和 MASK 方法的平均项集误差

图 3 中则详细给出了当随机化参数 p 分别取值 0.2,0.3;0.35,0.4;0.45,0.49,0.51,0.55;0.6,0.7,0.8 时,RRPH 方法和 MASK 方法在最小支持度阈值 s=1‰,2‰,3‰,4‰,5‰,6‰,7‰,8‰,9‰,10‰情况下的支持度误差比较.

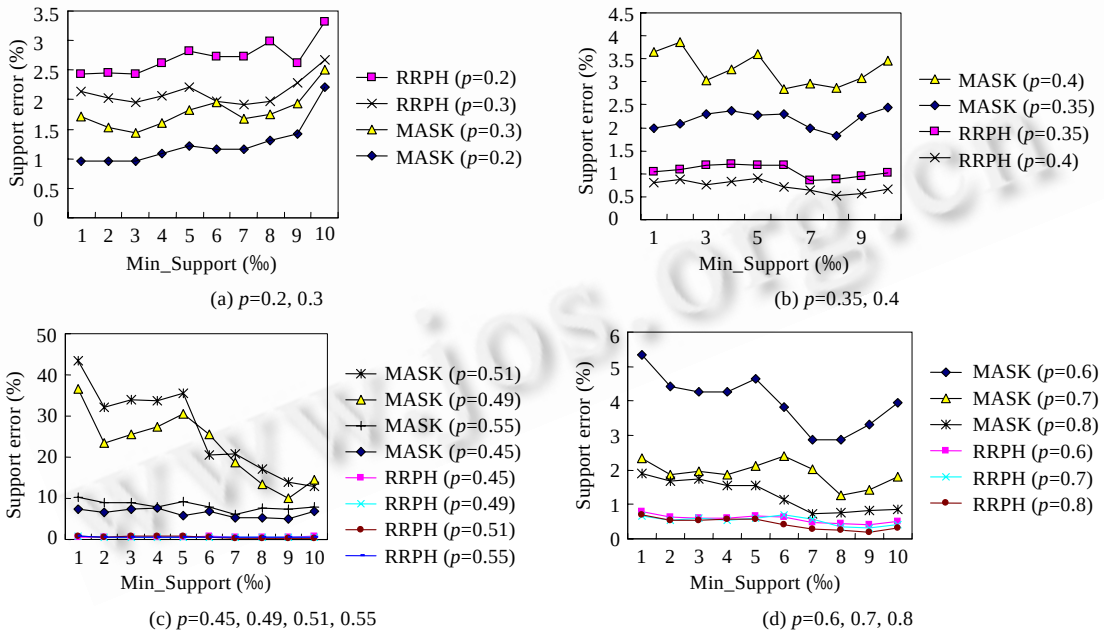

Fig.3 Support error of RRPH method and MASK method

图 3 RRPH 方法和 MASK 方法的支持度误差

5.3 结果分析

首先,从图 2 的结果可以看出:MASK 方法的误差变化比较大,当 p 接近 0 或 1 时,挖掘结果比较准确;但此时的隐私破坏系数接近于 1,方法对隐私的保护程度很差;在 p 从 0 或 1 逐渐接近 0.5 的过程中,隐私破坏系数会逐

渐减小,隐私保护的程度在不断提高,但挖掘结果的准确性将显著下降.而本文提出的 RRPH 方法误差变化相对比较平稳,随着 p 值,也就是真实数据所占的比例从 0 增加到 1,隐私破坏系数也从 0 增长到 1,方法对隐私的保护程度不断下降,而挖掘结果的准确性不断提高.再从图 3 的详细结果可以看出:当 p 的取值较小时,MASK 方法的误差比 RRPH 方法要小,准确性要高;而当 p 的取值超过 0.35 以后,RRPH 方法的误差就要低于 MASK 方法了,特别是当 p 值接近 0.5 时,误差更是相差了数十倍.这些实验结果与第 4 节中关于隐私性和准确性方面的论证是完全一致的.

在理论分析和实验结果的基础上,权衡数据的隐私性和挖掘结果的准确性,我们建议在区间[0.35,0.6]上选取随机化参数 p 的值,使用 RRPH 方法进行隐私保护的关联规则挖掘.

6 结论与展望

在本文中,我们将数据干扰和查询限制的策略相结合,提出了一种新的随机处理方法——RRPH,来进行数据的变换和隐藏;然后针对经过 RRPH 方法处理的数据,给出一种简单而又高效的频繁项集生成算法,从而实现了一种新的隐私保护的关联规则挖掘方法.

我们还通过理论上的分析和实验结果说明了 RRPH 方法中的随机化参数对数据隐私性和挖掘结果准确性的影响.通过合理的随机化参数选择,该方法在相同时间和空间开销的条件下,可以得到比原有方法更好的隐私性和准确性,并且还具有很好的适用性.

在未来的工作中,我们希望能够进一步提高挖掘算法的运行效率,并利用 RRPH 方法良好的适用性,扩充解决问题的范围,包括支持分类属性类型的数据以及实现基于 RRPH 的分类挖掘方法.

References:

[1] Verykios VS, Bertino E, Fovino IN, Provenza LP, Saygin Y, Theodoridis Y. State-of-the-Art in privacy preserving data mining. SIGMOD Record, 2004,33(1):50−57.

[2] Han J, Kamber M. Data Mining: Concepts and Techniques. Beijing: China Machine Press, 2001 (in Chinese).

[3] Agrawal R, Srikant R. Privacy-Preserving data mining. In: Weidong C, Jeffrey F, eds. Proc. of the ACM SIGMOD Conf. on Management of Data. Dallas: ACM Press, 2000. 439−450.

[4] 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. Hong Kong: Morgan Kaufmann Publishers, 2002. 682−693.

[5] Agrawal S, Krishnan V, Haritsa JR. 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. LNCS 2973, Jeju Island: Springer-Verlag, 2004. 113−124.

[6] Evfimievski A. Randomization in privacy preserving data mining. SIGKDD Explorations, 2002,4(2):43−48.

[7] Evfimievski A, Srikant R, Agrawal R, Gehrke J. Privacy preserving mining of association rules. In: Hand D, Keim D, Ng R, eds. Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Edmonton: ACM Press, 2002. 217−228.

[8] Saygin Y, Verykios VS, Clifton C. Using unknowns to prevent discovery of association rules. ACM SIGMOD Record, 2001,30(4): 45−54.

[9] Oliveira SRM, Zaiane OR. Privacy preserving frequent itemset mining. In: Clifton C, EstivillCastro V, eds. Proc. of the IEEE Int'l Conf. on Data Mining Workshop on Privacy, Security and Data Mining. Maebashi: IEEE Computer Society, 2002. 43−54.

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

[11] 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. Edmonton: ACM Press, 2002. 639−644.

[12] Clifton C, Kantarcioglu M, Vaidya J, Lin X, Zhu MY. Tools for privacy preserving distributed data mining. SIGKDD Explorations, 2002,4(2):28−34.

[13] Zhao JK. Theory and Methods of Sampling Design in Statistical Survey. Beijing: China Statistics Press, 2002 (in Chinese).

[14] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Bocca JB, Jarke M, Zaniolo C, eds. Proc. of the 20th Int'l Conf. on Very Large Data Bases. Santiago: Morgan Kaufmann Publishers, 1994. 487−499.

附中文参考文献:

[2] 韩家炜,坎伯.数据挖掘:概念与技术.北京:机械工业出版社,2001.

[13] 赵俊康.统计调查中的抽样设计理论与方法.北京:中国统计出版社,2002.

张鹏(1978－),男,北京人,博士生,主要研究领域为数据仓库,数据挖掘,联机分析处理.

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

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模型,数据库系统,数据仓库,数据挖掘.

杨冬青(1945－),女,教授,博士生导师,CCF 高级会员,主要研究领域为数据模型,数据库系统,Web 环境下的信息集成与共享.

马秀莉(1972－),女,博士,讲师,主要研究领域为数据仓库,数据挖掘,联机分析处理.

****

中国计算机学会优秀博士论文评奖启动通知

为推动中国计算机领域的科技进步，鼓励创新性研究，促进青年人才成长，中国计算机学会（CCF）设立了优秀博士学位论文奖。从 2006 年开始，CCF 每年评选一次 CCF 优秀博士学位论文奖。2006 年度优秀博士学位论文的评选范围为 2003 年 7 月 1 日~2006 年 6 月 30 日在中国获得的计算机科学与技术学科相关专业博士学位的学位论文。CCF 办公室从 2006 年 7 月 10 日起受理本年度申请。受理截止日期为 2006 年 8 月 20 日。参加评选的博士学位论文须经具有计算机科学与技术学科博士点的高校计算机学院（系）或研究机构推荐，或由 3 位（含）以上 CCF 理事推荐。详情请访问 CCF 网站：http://www.ccf.org.cn。评选结果将于 2006 年 11 月 30 日前公布。