隐私保护关联规则挖掘算法的研究

王 锐,刘 杰
WANG Rui,LIU Jie

哈尔滨工程大学 计算机科学与技术学院,哈尔滨 150001
College of Computer Science and Technology,Harbin Engineering University,Harbin 150001,China
E-mail:liujie@hrbeu.edu.cn

WANG Rui,LIU Jie.Research of privacy preserving association rules mining algorithm.Computer Engineering and Applications,2009,45(26):126-130.

Abstract: In view of the insufficiency of MASK algorithm,randomized response technology and association rule mining algorithm are integrated and a multi-parameters randomized disturb algorithm is proposed,which is called MRD algorithm.When the data sets are processed with different random parameters,the original data can be disturbed and hidden,and the defects of the simplex using of data diturb and data hiding strategy are solved,and the privacy-preserving degree of the algorithm is improved effectively.On this basis,the algorithm of generating frequent items from transformed data sets is proposed.Finally,through specific certification of examples,it can be proved that when the random parameters are choosen suitably,the privacy and accuracy of MRD algorithm are both better than the original algorithm.

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

摘 要:针对 MASK 算法的不足,将随机响应技术与关联规则挖掘算法相结合,提出一个多参数随机扰动算法—MRD 算法。当以不同的随机参数对数据集进行处理时,可以实现对原始数据的干扰或隐藏,解决了单一使用数据干扰策略和数据隐藏策略的缺陷,有效地提高了算法的隐私保护度。在此基础上,给出了在伪装后的数据集上生成频繁项集的挖掘算法。最后,通过具体实例验证,证明了当随机参数选择合适时,MRD 算法的隐私性和准确性均优于原算法。

关键词:数据挖掘;关联规则;频繁项集;隐私保护;随机响应
DOI:10.3778/j.issn.1002-8331.2009.26.037 **文章编号:**1002-8331(2009)26-0126-05 **文献标识码:**A **中图分类号:**TP301

1 引言

随着数据库技术和网络技术的飞速发展,各行各业都积累了大量有用的数据。如何从这些数据中提取出对决策有价值的知识,成为当务之急。数据挖掘[1]作为一个强有力的数据分析工具,可以发现数据中潜在的模式和规律。然而,由于被挖掘的资料或数据还包含许多敏感数据,必须受到保护,因此,数据挖掘应该在隐私保护的前提下开展。尤其是在现在,数据挖掘和知识发现技术不断地进步,使用这些技术可以在海量的信息中提取出隐藏的、有用的数据和知识,从而更增加了当资料公开给外界时所存在的风险,会对隐私和信息安全构成威胁。因此,进行数据挖掘的同时保护用户数据的隐私是未来数据挖掘的一个极其重要而富有挑战性的课题。将随机响应技术与关联规则挖掘算法相结合,提出一个多参数随机扰动算法—MRD 算法。在此基础上,给出了在伪装后的数据集上生成频繁项集的挖掘算法。最后,通过实验分析,表明了算法的有效性,同时研究了算法中若干参数的取值问题。

2 基本概念

设 $I=\{i_1,i_2,\cdots,i_m\}$是项的集合。设任务相关的数据 D 是数据库事务的集合,其中每个事务 T 是项的集合,$T\subseteq I$。每个事务有唯一的标识符,记为 TID。设 A 是一个项集,事务 T 包含 A 当且仅当 $A\subseteq T$。关联规则是形如 $A\Rightarrow B$ 的蕴涵式,其中 $A\subset I$,$B\subset I$,且 $A\cap B=\phi$。规则 $A\Rightarrow B$ 在事务集 D 中成立,具有支持度 s,其中 s 是事务集 D 中包含 $A\cup B$ 的百分比,即:$support(A\Rightarrow B)=\frac{|A\cup B|}{|D|}\times 100\%$。规则 $A\Rightarrow B$ 在事务集 D 中具有置信度 c,它是事务集 D 中包含 A 的事务同时也包含 B 的百分比,即:$confident(X\Rightarrow Y)=\frac{support(X\cup Y)}{support(X)}$。挖掘关联规则就是产生支持度和置信度分别不小于最小支持度和最小置信度阈值的规则。

关联规则挖掘的过程分为以下两步:

步骤 1 发现支持度不低于用户给定的最小支持度阈值的频繁项集。

基金项目:黑龙江省自然科学基金(the Natural Science Foundation of Heilongjiang Province of China under Grant No.F0310);哈尔滨工程大学基础研究基金项目(No.HEUF04091)。

作者简介:王锐(1982-),女,硕士研究生,主要研究方向:数据挖掘;刘杰(1965-),男,教授,主要研究方向:人工智能、数据库、数据挖掘。

收稿日期:2008-05-14 **修回日期:**2008-09-01

步骤 2 根据步骤 1 发现的频繁项集，产生置信度不低于用户给定的最小置信度阈值的关联规则。

3　多参数随机扰动算法

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和查询限制两大类[2]。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。

3.1　随机扰动技术

MASK 算法是由 Rizvi[3]学者利用贝努利概率模型提出,主要应用于购物篮事务数据集。该数据集的列由商品名组成,行表示每位顾客购物行为,是 1 和 0 的字符串。其中 1 表示购买,0 表示未购买。即一个顾客元组是一个随机向量 $\boldsymbol{X}=\{X_i\}$,$X_i=1$ 或 0,从该元组中产生变换向量 $\boldsymbol{Y}_i=\text{Distortion}(X_i)$。这里定义的变换函数为 $\boldsymbol{Y}_i=\boldsymbol{X}_i$ XOR $\bar{r}_i$,r_i是一个贝努利随机变量，满足二项式分布,$\bar{r}_i$是 r_i的补集。算法的主要思想用概率方法改变数据的原始值,使得项目值以概率 p 保持不变,以 $1-p$ 的概率取反。若项目从 1 变成 0,则相当于删除项目;反之,则为添加噪声项目。其实质是对数据集中的项目以一定概率进行增删或保持不变,从而对原数据集的信息进行了保护。由于发现关联规则必须首先获得频繁项目集,因此需对项目集的支持度进行重构(并非重构项目的实际值),估算项目实际支持度,从而发现频繁项目集。

MASK 算法采用的基本策略是数据干扰,该方法通过数据干扰和支持度重构实现了隐私保护的关联规则挖掘。但 MASK 方法也存在数据干扰策略的不足,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 MASK 算法使用唯一的参数 p 对数据集进行干扰,不可避免地使隐私性和准确性成为一对矛盾。例如:当概率 p 接近 0 或 1 时,隐私保护度接近于 0,方法的隐私性很差;在概率 p 从 0 或 1 逐渐接近于 0.5 过程中,隐私保护度在不断地提高,但挖掘结果的准确性却显著的降低。

3.2　多参数随机扰动算法的提出

针对 MASK 算法的不足,提出一个改进算法—多参数随机扰动(Multi-parameters Randomized Disturb,MRD)算法,实现数据的变换和隐藏。对数据集中的数据以多个参数实施随机扰动,从而尽可能阻止隐私泄露,提高隐私保护度。

设 D 是布尔型数据集,每一行代表一个事务。D 中 1 代表相应的项目在事务中出现,0 则相反。为了保护 D 中的具体项目值,对 D 进行随机扰动,生成相应的数据集 D'。通过对 D'中项目的支持度重构,从而发现原始数据集 D 中的关联规则。

算法思想如下:

给定随机化参数 $0\leqslant p_1,p_2,p_3\leqslant 1$,$p_1+p_2+p_3=1$,对于项 $t\in\{0,1\}$,设 $f_1=t$,$f_2=1-t$,$f_3=0$,则随机化函数 $f(t)$以 p_i的概率选择取值为 f_i,$i=1,2,3$。设项的总数为 k,则对于用 0-1 序列表示的事务 $T=\{t_1,t_2,\cdots,t_k\}$,干扰后的事务 $T'=\{t_1',t_2',\cdots,t_k'\}$,可以通过 $T'=F(T)$计算得到,其中 $t_i'=f(t_i)$。t_i'以 p_1的概率取值为 t_i,以 p_2的概率取值为 $1-t_i$,以 p_3的概率取值为 0。

当随机选择概率 p_1或 p_2进行取值时，实现了数据的干扰策略;而当随机选择概率 p_3进行取值时,数据值变为 0,相应的事务被隐藏起来,实现了查询限制策略的数据隐藏方法。

以 p_3的概率取“0”的原因是数据“1”是数据集信息的体现,是用户需要保护的值,希望通过变换把它隐藏起来,即“1”→“0”,当以 p_3的概率变换时达到了此目的,这正是对查询限制策略的应用。

算法伪代码如下:

输入:原事务集 D,随机参数 p_1,p_2。

输出:使用 MRD 算法处理后的事务集 D'。

```
(1)  扫描事务集 D,for each transaction t∈D{
(2)  for(k=0;k<N;k++)   //N 是事务集中的事务数
(3)    for each item i∈I{
(4)      生成一个随机数 θ;
(5)      if(θ≤p1),then
(6)        t_{k,i}←t_{k,i}   //项 t_{k,i}以概率 p1取正
(7)      else
(8)        if(p1≤θ ≤p1+p2),then
(9)          t_{k,i}←1-t_{k,i}   //项 t_{k,i}以概率 p2取反
(10)       else
(11)         t_{k,i}←0   //项 t_{k,i}以概率 p3=1-p1-p2取 0
(12)   }
(13) }
(14) 输出处理后的事物集 D';
```

4　隐私保护的关联规则挖掘算法

在整个挖掘过程中,对于挖掘算法而言,扰动后的数据集与随机扰动参数 p_1,p_2,p_3是已知的。以下通过分析单个项目 i 支持度被重构的过程,从而推导得到 n 项集支持度的重构。

4.1　1-项集支持度重构

随机考虑一个项目 i,C_1^T和 C_0^T分别表示原始数据集 T 中第 i 列 1 和 0 的个数,C_1^D和 C_0^D分别表示变换后的数据集 D 中第 i 列 1 和 0 的个数。根据这个概念，使用下面的公式重构 T 中 i 的支持度:

$$\boldsymbol{C}^T=\boldsymbol{M}^{-1}\boldsymbol{C}^D \tag{1}$$

其中

$$\boldsymbol{M}=\begin{bmatrix}p_1 & p_2\\ p_2+p_3 & p_1+p_3\end{bmatrix},\boldsymbol{C}^T=\begin{bmatrix}C_1^T\\ C_0^T\end{bmatrix},\boldsymbol{C}^D=\begin{bmatrix}C_1^D\\ C_0^D\end{bmatrix}$$

求解如下:

$$c_1^T=\frac{c_1^D-p_2}{p_1-p_2}(p_1\neq p_2) \tag{2}$$

4.2　k-项集支持度重构

所提出的随机扰动对于各项目之间是独立操作的,因此根据公式(1)容易扩展为 k-项集的支持度区间重构公式。

设 $I=\{i_1,i_2,\cdots,i_k\}$是一个 k–项集,当所有项都使用相同的随机化参数时,T 中的项集 j 经过变换后,转变为 D 中的项集 i 的概率 $m_{i,j}$ 是相等的。

$$m_{i,j}=\prod_{x\in I}F_x \tag{3}$$

其中,F_x 计算如下:

$$F_x=\begin{cases}p_1+p_3 & i_x=0,j_x=0\\ p_2+p_3 & i_x=0,j_x=1\\ p_2 & i_x=1,j_x=0\\ p_1 & i_x=1,j_x=1\end{cases} \tag{4}$$

于是有:

$$\boldsymbol{C}^D=\boldsymbol{M}_k\boldsymbol{C}^T \tag{5}$$

其中

$$\boldsymbol{C}^D=\begin{bmatrix}c^D_{2^k-1}\\ \vdots\\ c^D_1\\ c^D_0\end{bmatrix}\qquad \boldsymbol{C}^T=\begin{bmatrix}c^T_{2^k-1}\\ \vdots\\ c^T_1\\ c^T_0\end{bmatrix}$$

$\boldsymbol{M}_k=[m_{ij}]$是一个 $2^k\times 2^k$ 的矩阵,m_{ij} 表示 T 中的项集 j 经过变换后,转变为 D 中的项集 i 的概率。

当 $\boldsymbol{M}_k$ 可逆时,$\boldsymbol{M}_k^{-1}=[a_{ij}]$,$\boldsymbol{C}^T=\boldsymbol{M}_k^{-1}\boldsymbol{C}^D$,则 k–项集的支持度为:

$$c^T_{2^k-1}=a_{2^k-1,0}c^D_0+a_{2^k-1,1}c^D_1+\cdots+a_{2^k-1,2^k-1}c^D_{2^k-1}$$

首先能算出项 j 在 $\boldsymbol{C}^D$ 中的支持度 $c^D_{2^j-1}$,并利用 $\boldsymbol{M}_k$ 求出 $a_{k,j}$,再计算出 k–项集支持度。

4.3 完整的挖掘算法

关联规则挖掘的本质是发现大于最小支持度的频集,MRD 算法和 MASK 算法均是基于 Apriori 算法的[4]。Apriori 是一个逐层迭代的算法,多次扫描事务集,在第 k 次扫描时,得到所有候选 k–项集,根据其支持度得到频繁 k–项集。和 Apriori 算法不同的是,MRD 算法对于计算某候选项目集的实际支持度时,需要考虑候选项集对应项的各种 0–1 组合。例如,k=2 时,需要分别考虑 00、01、10、11 的个数。因此在算法实现过程中需要对每个子集进行计数,每个候选项集的计数器也从 Apriori 算法中的一个变成了 2^k 个。下面给出经过 MRD 算法处理后的数据生成频繁项集的具体算法。伪代码如下:

输入:被扰动的事务集 D',最小支持度阈值 minsup。

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) $L_1=(\{i\}|i\in I,((i.\text{count}/N)-p_2)/(p_1-p_2)\geqslant minsup)$;

(3) for(k=2;$L_{k-1}\neq\phi$;k++){

(4) C_k=aproiri_gen(L_{k-1}) //生成候选 k–项集 C_k

(5) for(i=0;$i<2^{k-1}$;i++){ //由(k–1)–项集变换矩阵,求出 k–项集的变换矩阵

(6) for(j=0;$j<2^{k-1}$;j++){

(7) $M_k[i][j]=(p_1+p_3)*M_{k-1}[i][j]$;

(8) $M_k[i][j+2^{k-1}]=(p_2+p_3)*M_{k-1}[i][j]$;

(9) $M_k[i+2^{k-1}][j]=p_2*M_{k-1}[i][j]$;

(10) $M_k[i+2^{k-1}][j+2^{k-1}]=p_1*M_{k-1}[i][j]$;

(11) }

(12) }

(13) for each candidate $c_i\in C_k$ //对候选集的支持度进行重构

(14) $re\_supp(c_i)=\sum_{i=0,j=2^k-1}M_k^{-1}[i][j]*supp(c_i)$;

(15) $L_k=\{\{c\}|c\in C_k,re\_supp(c)\geqslant minsup\}$;

(16) }

(17) return $L=\cup_k L_k$

5 算法性能衡量

从隐私保护度和挖掘结果准确度两个方面对算法的性能进行衡量[5]。

5.1 隐私保护度的衡量

隐私保护度的衡量是对隐私保护挖掘算法关于用户隐私信息保护程度的量化。假设某客户 U 买了商品 i,则 i 的初始值即为 1,在挖掘前能够从变换后的数据集中精确重构的概率。设 s_i 是第 i 项的支持度,这表示一个随机顾客购买第 i 项的概率为 s_i。设原始的真实分量为 U_i,与之对应的扰动后分量为 V_i。由全概率公式可知正确重构原始值的概率为:

$$R_1(s_i)=P(V_i=1|U_i=1)P(U_i=1|V_i=1)+P(V_i=0|U_i=1)P(U_i=1|V_i=0) \tag{6}$$

由条件概率的定义可知:

$$P(U_i=1|V_i=1)=\frac{P(U_i=1\cap V_i=1)}{P(V_i=1)}=\frac{P(U_i=1)P(V_i=1|U_i=1)}{P(V_i=1)}=\frac{s_i\times p_1}{P(U_i=1)P(V_i=1|U_i=1)+P(U_i=0)P(V_i=1|U_i=0)}$$

则有:

$$P(U_i=1|V_i=1)=\frac{s_i\times p_1}{s_i\times p_1+(1-s_i)\times p_2} \tag{7}$$

同理:

$$P(U_i=1|V_i=0)=\frac{s_i\times(1-p_1)}{s_i\times(1-p_1)+(1-s_i)\times(1-p_2)} \tag{8}$$

所以:

$$R_1(S_i)=\frac{p_1^2s_i}{p_2(1-s_i)+p_1s_i}+\frac{(1-p_1)^2s_i}{1-p_2+(p_2-p_1)s_i} \tag{9}$$

原始值为 1 的总的重构概率为:

$$R_1=\frac{\sum_i s_iR_1(s_i)}{\sum_i s_i} \tag{10}$$

当数据库中所有的项取相同的支持度时,R_1 的值最小。作为一个估计值,这里用 s_0 代替每个具体项的支持度,R_1 简化为:

$$R_1=\frac{p_1^2s_0}{p_2(1-s_0)+p_1s_0}+\frac{(1-p_1)^2s_0}{1-p_2+(p_2-p_1)s_0} \tag{11}$$

使用相同的方法,可以推导出真值为 0 的重构概率为:

$$R_0=\frac{p_2^2(1-s_0)}{p_2(1-s_0)+p_1s_0}+\frac{(1-p_2)^2(1-s_0)}{1-p_2+(p_2-p_1)s_0} \tag{12}$$

总的重构概率为:

$$R=\alpha R_1+(1-\alpha)R_0 \tag{13}$$

其中,α 为 0–1 数据库中1 的个数的权重。

根据重构率定义,隐私性可简单定义为 $P=(1-R)\times100$。

5.2　挖掘准确度的衡量

隐私保护关联规则挖掘的目标在于保护数据隐私同时允许挖掘者发现精度范围内的频集。但是随机扰动的隐私保护技术是采用概率的方法,实际情况下是不能做到重构支持度的值与实际的值保持一致。这就意味着对于估算频集支持度存在着误差,即大于或小于实际的支持度。支持度估算错误估算比错误统计频集支持度更加有害,因为它们能导致识别频集的错误。因此,关联规则挖掘错误可以按照两种标准来衡量:支持度的错误率(Support Error)和频繁项目集错误率(Identity Error)。

(1)Support Error(ρ)

以 f 表示重建的频繁项目集里真正的频繁项目集所构成的集合。ρ 指 f 中项集支持度的估算值与实际值的平均偏差,这个标准影响重构支持度的平均相对误差,定义如下:

$$\rho=\frac{1}{|f|}\sum\frac{|rec\_sup\_act\_sup|}{act\_sup}\times100\% \qquad (14)$$

(2)Identity Error(σ)

R 表示重建的频繁项目集所构成的集合,F 表示实际的频繁项目集所构成的集合。σ 用来衡量 R 的错误率,包含 σ^+(高估频集,指那些本不是频集的而错误地认为是频集)和 σ^-(低估频集,指那些实际的频集误认为不是频集)两部分,定义如下:

$$\sigma^+=\frac{|R-F|}{|F|}\times100 \qquad \sigma^-=\frac{|F-R|}{|F|}\times100$$

6　实验及结果分析

通过实验来对比 MASK 算法和 MRD 算法的性能。以数据隐私性和挖掘结果准确性作为算法性能的衡量标准,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。

利用 IBM Almaden Research Center 所开发的人工数据集生成器(Synthetic Data Generator)生成事务集 D,参数是 T10I4D100KN1K(T:事务平均长度,I:频繁项目集的平均长度,D:事务数,N:项目数)。

由于篇幅有限,以最小支持度为 0.3%的实验结果进行分析。最小支持度为 0.3%相对于数量庞大的频繁项集来说足够低,因此更能体现 MRD 算法的性能。

表 1 显示了最小支持度为 0.3%,p_1、p_2、p_3 取不同随机数时,MRD 算法的隐私保护度情况。

表 1　隐私保护度的示例

p_1	p_2	p_3	隐私保护度
0.3	0.6	0.1	92.6
0.1	0.6	0.3	92.5
0.4	0.2	0.4	92.8
0.4	0.1	0.5	92.4
0.1	0.3	0.6	92.6
0.2	0.1	0.7	91.6

表 2~表 4 显示随机参数 p_3 取 0.1,p_1/p_2 分别取 3、7、10 时,各层挖掘结果错误的比较。最小支持度为 0.3%,生成的频繁项集长度到 8 为止。结果表明,挖掘结果均保持较低的错误率,而且当 p_1/p_2 的比值较大时,错误率更低。

表 2　p_1/p_2=3 时各层挖掘结果错误率

频繁项集长度	频繁 k–项集个数	σ^+/(%)	σ^-/(%)	ρ/(%)
1	765	1.52	1.40	2.24
2	2 034	5.90	5.29	2.08
3	1 563	3.87	4.84	1.68
4	945	6.34	7.01	2.01
5	467	5.76	6.24	2.65
6	129	4.01	4.56	2.45
7	34	3.23	3.24	3.03
8	3	0	0	2.65

表 3　p_1/p_2=7 时各层挖掘结果错误率

频繁项集长度	频繁 k–项集个数	σ^+/(%)	σ^-/(%)	ρ/(%)
1	765	1.51	1.43	1.91
2	2 034	4.69	3.93	1.82
3	1 563	3.47	3.64	1.63
4	945	5.24	5.01	2.02
5	467	4.64	5.74	1.82
6	129	3.87	4.05	1.73
7	34	3.02	2.11	1.64
8	3	0	0	1.69

表 4　p_1/p_2=10 时各层挖掘结果错误率

频繁项集长度	频繁 k–项集个数	σ^+/(%)	σ^-/(%)	ρ/(%)
1	765	1.42	1.31	1.56
2	2 034	4.62	3.81	1.49
3	1 563	3.40	3.57	1.45
4	945	5.16	4.93	1.58
5	467	4.69	5.62	1.63
6	129	3.81	3.89	1.46
7	34	2.98	2.03	1.32
8	3	0	0	1.35

图 1、图 2 分别显示最小支持度阈值为 0.3,p_3 分别取0.1、0.2、0.3 时,支持度误差与参数的关系。可以看出当 p_1、p_2、p_3 选择合适时,支持度误差较小。

图 3 显示 MASK 算法最小支持度分别取 0.3 和 0.7 时,支持度误差与参数的关系。

从表 1 的实验结果,可以看出当 MRD 算法的最小支持度取 0.3%(这也是大多数数据集选择的最小支持度阈值)时,p_1、p_2、p_3 的几种不同取值组合下的隐私保护度都大于 90%,具有很好的隐私性。

从表 2~表 4 的实验结果,可以看出当 MRD 算法的最小支持度取 0.3%,p_3=0.1,p_1/p_2 分别取 3、7、10 时,各层挖掘结果的项集误差和支持度误差都很小。而且当 p_1/p_2 的比值较大时,误差较小。

从图 1、图 2 的实验结果,可以看出当 MRD 算法最小支持度分别取 0.3 和 0.7,p_3 分别取 0.1、0.2、0.3 时,支持度误差与 p_1/p_2 比值的关系。当最小支持度为 0.3 时,支持度误差随着 p_1/p_2 比值的增大,逐渐降低;当最小支持度为 0.7 时,支持度误差随着 p_2/p_1 比值的增大,逐渐降低。且支持度误差都很小。图 3 是 MASK 算法的最小支持度分别取 0.3 和 0.7 时,支持度误差与随机参数的关系。可以看出随机参数的选择对支持度误差有较大的影响。尤其在参数 p 取 0.49 和 0.51 时,支持度的误差很大。

最后,没有一种参数组合设置是在各个方面最优,要根据

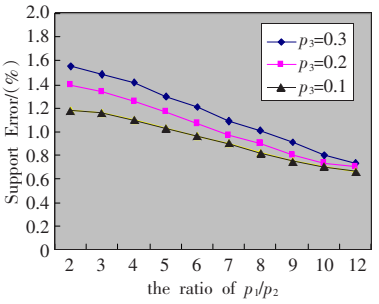

图1 支持度误差与参数的关系
(*minsup*=0.3%)

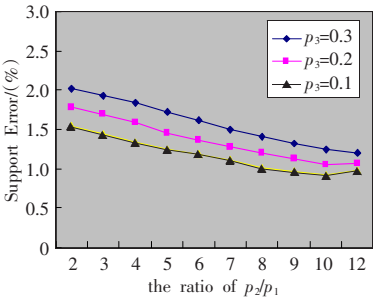

图2 支持度误差与参数的关系
(*minsup*=0.7%)

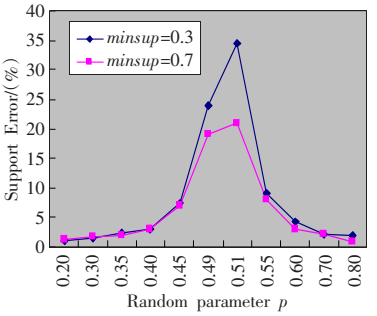

图3 MASK 算法的支持度误差与参数的关系

实际需求进行调整。MRD 算法当 $p_1=p_2$ 时,误差最大;当 p_1-p_2 的绝对值增大时,误差将会减小。并且这个方法的误差率还与数据量的大小有关,当数据量足够大时,受这三个参数的影响减小。在数据量一定的情况下,为求得较小的误差率,应该使这两个绝对值较大,但这两个绝对值的增大,可能使信息隐私性降低,如何在误差率在隐私性之间进行平衡,还需要进一步讨论。

7 结论

针对 MASK 算法的不足,把多参数扰动、随机响应技术和关联规则相结合,提出了 MRD 算法。当以不同的随机参数对数据集进行处理时,可以实现对原始数据的干扰或隐藏,解决了单一使用数据干扰策略和数据隐藏策略的缺陷,有效地提高了算法的隐私保护度。并针对 MRD 算法处理后的数据,给出一个频繁项集生成算法,实现了一种新的隐私保护关联规则挖掘算法。最后通过理论分析和实验结果表明了合理的参数选择,能够同时满足隐私保护和挖掘准确度的要求,性能相对于 MASK 算法均大为提高。

目前对数据集的处理采用的都是统一的扰动概率,扰动概率的单一性容易造成隐私泄露。因此,未来可以对不同的属性根据其保密等级的不同采取不同的参数进行随机扰动。对每个 k-项集,都需要构造一个 $2^k\times 2^k$ 阶的变换矩阵,需要计算 2^k 个等式,将会耗费大量的时间和空间的开销。因此,如何对扰动后的数据集进行重构以及如何通过数据集存储方式的不同减少算法的时间和空间复杂度是未来值得研究的问题。

参考文献:

[1] Han Jia-wei,Kamber M.数据挖掘概念与技术[M].北京:机械工业出版社,2001:100-200.
[2] 张鹏,童云海,唐世渭,等.一种有效的隐私保护关联规则挖掘方法[J].软件学报,2006,17(8):1765-1774.
[3] Rizvi S,Haritsa J.Maintaining data privacy in association rule mining[C]//Proc of the 28th International Conference on Very Large Databases,August 2002.
[4] Agrawal R,Srikant R.Fast algorithms for mining association rules[C]//Proc of 20th International Conference on Very Large Data Bases,September 1994.
[5] 陈芸.隐私保护关联规则挖掘[D].无锡:江南大学,2006:21-23.
[6] Oliveira S R M,Zaiane O R.Privacy preserving frequent itemset mining[C]//Proc of the IEEE International Conference on Privacy,Security,Data mining,2002:43-54.
[7] Oleveira S R M,Zaiane O R.Protecting sensetive knowledge by data sanitization[C]//Proceedings of the Third IEEE International Conference on Data Mining,2003.
[8] Verykios V S,Bertino E,Nai Fovino I.State-of-the-art in privacy-preserving data mining[C]//SIGMOD Record,2004,33(1).
[9] 仲波.基于关联规则的隐私保护算法研究[D].兰州:兰州理工大学,2007:10-18.
[10] 沈中林,崔建国.隐私保护下关联规则挖掘方法[J].中国民航大学学报,2007,25:108-114.

(上接 106 页)

[4] Bauer F,Varma A.Distributed algorithms for multicast path setup in data networks[J].IEEE/ACM Transaction on Networking,2006,1(3):287-293.
[5] Wu J J,Hwang R H.Multicast routing with multiple constraints[J].Information Sciences,2005,124:29-57.
[6] Chen S G.Routing support for providing guaranteed end-to-end quality-of-service[D].Engineering College of Computer Science University of Illinonis at Urbana-Champaign,2004.
[7] 邢文训,谢金星.现代优化计算方法[M].北京:清华大学出版社,2005.
[8] Li La-yuan,Li Chun-lin.QoS multicast routing in networks with uncertain parameter[C]//Proceedings of the International Parallel and Distributed Processing Symposium,2006.

(上接 120 页)

一个语义块切分系统。接下来的工作在于进一步提高语义块切分的效果,主要的研究方向是:更好地利用概念特征;研究语义特征和远距特征的使用;考虑在统计模型框架下加入规则。另外,训练数据不足的问题会随着 HNC 标注语料库的建设逐渐得到解决。

参考文献:

[1] Darroch J N,Ratcliff D.Generalized iterative scaling for loglinear models[J].The Annals of Mathematical Statistics,1972,43(5):1470-1480.
[2] Berger A L,Pietra S A D,Pietra V J D.A maximum entropy approach to natural language process ing[J].Computational Linguistics,1996,22(1):1-36.
[3] 黄曾阳.HNC 理论概要[J].中文信息学报,1997(4):11-20.
[4] 李素建,刘群,杨志峰.基于最大熵模型的组块分析[J].计算机学报,2003,26(12):1722-1727.
[5] 周雅倩,郭以昆,黄萱菁,等.基于最大熵模型的中英文基本名词短语识别[J].计算机研究与发展,2003,40(3):440-446.