第 36 卷 第 5 期 计 算 机 工 程 2010 年 3 月
Vol.36 No.5 Computer Engineering March 2010
·软件技术与数据库· 文章编号：1000—3428(2010)05—0079—02 文献标识码：A 中图分类号：TP309

基于分组无关问题模型的隐私保护算法

阚莹莹，曹天杰
(中国矿业大学计算机学院，徐州 221116)

摘 要：针对分组无关问题模型存在隐私泄露的问题，提出一种改进的分组无关问题模型，采用随机响应的方法，通过对原始数据进行伪装变换处理，实现具有隐私保护的关联规则挖掘。实验结果表明，改进后的模型在伪装变换后的数据集上挖掘出的规则与原始数据规则相比，保证了低误差，具有较好的隐私保护性。

关键词：分组无关问题模型；随机响应；关联规则挖掘

Privacy-preserving Algorithm Based on Grouping Unrelated-question Model

KAN Ying-ying, CAO Tian-jie
(School of Computers Science and Technology, China University of Mining and Technology, Xuzhou 221116)

【Abstract】 Aiming at the problem of privacy leak exists in the grouping unrelated-question model, this paper presents an improved grouping unrelated-question model to achieve the Privacy-preserving association rules through disguising and changing the original data. Experimental results show that the rule which algorithm gets has a lower error and better privacy compared with the original rule.

【Key words】 grouping unrelated-question model; random response; association rules mining

1 概述

随着网络和数据挖掘技术的发展，企业通过在网上提供调查问卷的方式收集数据进行数据挖掘。如果调查问卷涉及用户的隐私，用户可能提供虚假数据，则基于虚假数据的数据挖掘会导致错误决策。如何能让用户在不威胁隐私的情况下提供正确的数据成为研究数据挖掘关键的问题。

随机响应技术[1]是在统计学中为保护被调查者的隐私而设计的数据隐藏技术。随机响应的方法是：为了解一群人中具有属性 A 的百分比，需要对这些人进行调查，由于属性 A 涉及个人隐私，因此被调查者可能不愿意回答或做出错误的回答。在相关问题模型[2]中，首先对属性 A 设计 2 个互为否定的问题：(1)具有属性 A；(2)不具有属性 A。对每个问题的回答均为“是”或“否”。调查者首先确定一个实数 $\theta\in[0,1]$，被调查者通过随机函数产生一个 0~1 之间的随机实数 r，若 $r\leqslant\theta$，则回答问题(1)，否则，回答问题(2)，即被调查者将以概率 θ 回答问题(1)，以概率 $1-\theta$ 回答问题(2)。尽管调查者知道被调查者的答案，但并不知被调查者回答的是哪个问题，从而保护了被调查者的隐私。假设分别用 $P\times(A=y)$ 和 $P\times(A=n)$ 来表示被调查者回答为“是”和“否”的概率。通过式(1)求出被调查者中具有属性 A 和不具有属性 A 的概率的近似值 $P(A=y)$ 和 $P(A=n)$：

$$\begin{cases}P\times(A=y)=P(A=y)\,\theta+P(A=n)\,(1-\theta)\\P\times(A=n)=P(A=y)\,(1-\theta)+P(A=n)\,\theta\end{cases}\tag{1}$$

当 $\theta\neq0.5$ 并且被调查人数很多时，这 2 个近似值的误差就足够小。

为进一步保护被调查者的隐私，文献[3]提出无关问题模型，与相关问题模型的区别在于提出的 2 个问题是没有关系的：(1)有敏感属性 A；(2)有属性 Y，Y 是一个不含敏感信息的属性。文献[4]提出分组无关问题模型，解决相关问题模型和无关问题模型的缺陷，但是存在一定的隐私泄漏，本文在分组无关问题模型的基础上做了改进。

2 分组无关问题模型的缺陷

设 π_0 是无关问题模型中敏感属性概率，从 $\lambda_0=p\pi_0+(1-p)\theta$ 可以得出：

$$\hat{\pi}=\frac{\lambda_0-(1-p)\theta}{p}\tag{2}$$

其中，p 是回答敏感属性的概率；θ 是属性 Y 的概率；λ_0 是回答为“是”的概率，$0\leqslant\lambda_0\leqslant1$。

Cranor 的调查表明 27%的人愿意提供他们的隐私数据[2]，现有的数据挖掘没有考虑到不同层次的人对隐私的看法不同。因此，文献[4]提出分组无关模型，每个人可以选择诚实的或随机的进行回答问题，提高收集数据的准确度。如果对被调查者进行分组，设 $k(0<k<1)$ 的被调查者不考虑敏感属性，π 是敏感属性的概率，λ 是回答为“是”的概率，则：

$$\lambda=k\times\pi+(1-k)\times p\times\pi+(1-k)\times(1-p)\times\theta\tag{3}$$

从式(3)可以得出：

$$\hat{\pi}=\frac{\lambda-(1-k)(1-p)\theta}{k+(1-k)p}\tag{4}$$

设 Ei 是一个变量，$Ei=1$ 表示回答为“是”，$Ei=0$ 表示回答为“0”，则 Ei 的分布概率：

$$P(Ei)=\begin{cases}k\pi+(1-k)p\pi+(1-k)(1-p)\theta & Ei=1\\k(1-\pi)+p(1-k)(1-\pi)+(1-k)(1-p)(1-\theta) & Ei=0\end{cases}\tag{5}$$

基金项目：江苏省自然科学基金资助项目(BK2007035)
作者简介：阚莹莹(1983－)，女，硕士，主研方向：数据挖掘，隐私保护算法；曹天杰，教授、博士生导师
收稿日期：2009-12-10 **E-mail：**yyingkan@126.com

从式(3)可得：

$$\lambda = \sum_{i=1}^{n} Ei / n$$

其中，n 是被调查者的人数，同时得出：

$$E(\hat{\pi})=\pi,\quad Var(\hat{\pi})=\frac{\lambda(1-\lambda)}{n[k+(1-k)p]^2} \tag{6}$$

文献[4]提出分组无关问题模型，设 Q 表示分组标志，当被调查者采用随机响应回答问题时，Q=1，否则 Q=0。当 Q=1 时，设 $p(0<p<1)$已知，通过随机函数产生一系列随机数 αi ($\alpha i \in [0,1]$)。如果 $\alpha i < p$，则被调查者回答所有敏感属性的问题；如果 $\alpha i \geqslant p$，则随机产生一系列随机数 βi ($\beta i \in [0,1]$)。如果 $\beta i < \theta$，则被调查者对所有无关问题回答“1”，否则回答“0”。但这样会破坏被调查者的隐私。如被调查者所有问题的答案都为“0”或“1”时，不会破坏隐私，因为有无关问题的干扰，但是如果所有问题的答案既有“0”又有“1”时，调查者就知道，被调查者回答的都是敏感属性的问题，会破坏被调查者的隐私。

3 对分组无关问题模型的改进

由于分组无关问题模型存在隐私泄漏，因此本文做出如下改进：设 Q 表示分组标志，当被调查者采用随机响应回答问题时，Q=1，否则 Q=0。当 Q=1 时，假设 $p(0<p<1)$已知，通过随机函数产生一系列随机数字 αi ($\alpha i \in [0,1]$)。如果 $\alpha i < p$，则被调查者回答所有敏感属性的问题；如果 $\alpha i \geqslant p$，则对于每个问题随机产生一系列随机数 βi ($\beta i \in [0,1]$)。如果 $\beta i < \theta$，则被调查者对该无关问题回答“1”，否则对该无关问题回答“0”。

设数据集有 N 个属性，E 是表示 N 个属性组合的一个逻辑表达式。$P(E)$为原始数据满足 E 的一个概率，是伪装后的数据满足 E 的概率，它可以从伪装后的数据直接计算获得，但是 $P(E)$才是挖掘时真正需要的数据，因为原始的数据无法直接得到，所以必须通过某种方法，由已知的 $P\times(E)$ 来估算 $P(E)$。

(1)当 N=1 时，E=1, $\overline{E}$ =0，则：

$$P\times(E)=k\times P(E)+(1\text{-}k)\times P(E)\times p+(1\text{-}k)(1\text{-}p)\times\theta \tag{7}$$

从式(7)可得：

$$P(E)=\frac{P\times(E)-(1-k)(1-p)\theta}{k+(1-k)\times p} \tag{8}$$

$$P(\overline{E})=\frac{P\times(\overline{E})-(1-k)(1-p)(1-\theta)}{k+(1-k)\times p} \tag{9}$$

其中，P,θ 是已知的，k 可以从 Q=1 的被调查者数计算得出。

(2)当 N=2 时，设有 2 个属性 E_1，E_2，$P\times(E_1E_2)$ 代表 2 个属性都为真，且

$$P\times(E_1E_2)=kP(E_1E_2)+(1-k)pP(E_1E_2)+(1-k)(1-p)\theta^2 \tag{10}$$

同理，

$$\begin{bmatrix} P\times(E_1E_2) \\ P\times(E_1\overline{E_2}) \\ P\times(\overline{E_1}E_2) \\ P\times(\overline{E_1E_2}) \end{bmatrix}=[k+p(1-k)]\begin{bmatrix} P(E_1E_2) \\ P(E_1\overline{E_2}) \\ P(\overline{E_1}E_2) \\ P(\overline{E_1E_2}) \end{bmatrix}+\begin{bmatrix} (1-k)(1-p)\theta^2 \\ (1-k)(1-p)\theta(1-\theta) \\ (1-k)(1-p)(1-\theta)\theta \\ (1-k)(1-p)(1-\theta)^2 \end{bmatrix} \tag{11}$$

因此，

$$\begin{bmatrix} P(E_1E_2) \\ P(E_1\overline{E_2}) \\ P(\overline{E_1}E_2) \\ P(\overline{E_1E_2}) \end{bmatrix}=\frac{1}{k+p(1-k)}\left(\begin{bmatrix} P\times(E_1E_2) \\ P\times(E_1\overline{E_2}) \\ P\times(\overline{E_1}E_2) \\ P\times(\overline{E_1E_2}) \end{bmatrix}-\begin{bmatrix} (1-k)(1-p)\theta^2 \\ (1-k)(1-p)\theta(1-\theta) \\ (1-k)(1-p)(1-\theta)\theta \\ (1-k)(1-p)(1-\theta)^2 \end{bmatrix}\right) \tag{12}$$

同理，当有 N 个敏感属性时：

$$\begin{bmatrix} P(E_1E_2\cdots E_N) \\ P(E_1E_2\cdots\overline{E_N}) \\ \cdots \\ P(\overline{E_1}\overline{E_2}\cdots E_N) \\ P(\overline{E_1E_2\cdots E_N}) \end{bmatrix}=\frac{1}{k+p(1-k)}\left(\begin{bmatrix} P\times(E_1E_2\cdots E_N) \\ P\times(E_1E_2\cdots\overline{E_N}) \\ \cdots \\ P\times(\overline{E_1}\overline{E_2}\cdots E_N) \\ P\times(\overline{E_1E_2\cdots E_N}) \end{bmatrix}-\begin{bmatrix} (1-k)(1-p)\theta^N \\ (1-k)(1-p)\theta^{N-1}(1-\theta) \\ \cdots \\ (1-k)(1-p)(1-\theta)\theta^{N-1} \\ (1-k)(1-p)(1-\theta)^N \end{bmatrix}\right) \tag{13}$$

下面介绍关联规则的概念，设 $I=\{i_1,i_2,\cdots,i_n\}$是项的集合，D 是数据库事务的集合，其中，每个事务 T 是项的集合，使得 $T\subseteq I$，每个事务有一个标识符，称作 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，$A\subset T$, $B\subset T$，并且 $A\cap B=\Phi$。规则 $A\Rightarrow B$ 在事务集 D 中成立，具有支持度 s，其中，s 是 D 中包含 $A\cup B$ 的百分比，即概率 $P(A\cup B)$。规则 $A\Rightarrow B$ 在事务集 D 中具有置信度 c，它是 D 中包含 A 的事务，同时也包含 B 的百分比，即条件概率 $P(B|A)$。关联规则挖掘是产生支持度和置信度分别大于用户给定的最小支持度和最小置信度阈值的规则，阈值可以由用户或领域专家根据经验设定。

由式(13)可以得出原始数据 N 个敏感属性不同取值的相应概率。同理，可以得出任意属性组合的概率，设任取 n 个属性，当各个属性取值相同且为 1 时，则：

$$P(E_i\cdots E_j)=\frac{1}{k+p(1-k)}[P\times(E_i\cdots E_j)-(1-k)(1-p)\theta^n] \tag{14}$$

同为 0 时，则：

$$P(\overline{E_i}\cdots\overline{E_j})=\frac{1}{k+p(1-k)}[P\times(\overline{E_i}\cdots\overline{E_j})-(1-k)(1-p)(1-\theta)^n] \tag{15}$$

若各个属性取值不同时，假如有 n_1 个属性取值为 1，n_2 属性取值为 0，则：

$$P(\overline{E_i}\cdots E_n)=\frac{1}{k+p(1-k)}[P\times(\overline{E_i}\cdots E_n)-(1-k)(1-p)\theta^{n_1}(1-\theta)^{n_2}] \tag{16}$$

其中，i, j, n 为任意整数，且 $1<i<j<N$。

4 对改进模型的实验分析

改进模型的实验步骤如下：

(1)随机产生有 10 000 个事务的样本数据集 D，假设该数据集有 3 个属性，设第 i 个事务为 $T_i=(A_i,B_i,C_i)$。

(2)当 p 和θ给定时，从原始数据集 D 中计算 π。设 p =0.3, θ=0.6 时，k =0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0。设 1-k 的记录为 Q=0，k 的记录为 Q=1，让 Q=0 的记录保持不变，另外，k 的记录用随机响应方法变换数据，即用随机函数产生一系列随机数 αi ($\alpha i \in [0,1]$)，如果 $\alpha i < p$，则保持记录不变，反之，对该条记录的每个属性随机产生一系列随机数 βi ($\beta i \in [0,1]$)。如果 $\beta i < \theta$，该记录对应属性的数据变为 1，反之为 0，此时原始数据集 D 变为 D^*。

(3)当 k 和 θ 给定时，k=0.2, θ =0.6。

1)取 p=0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0。

2)用关联规则算法求出原始数据集频繁大项集的支持度 S，根据式(13)用 D^* 的数据估算频繁大项集的支持度 S'，误差 $W=|S-S'|$。

由图 1 得出，误差随着 p 的增大而减少，当 p 接近 1 时，误差接近于 0，隐私保护程度最差，误差的大小和隐私保护程度是矛盾的，在实际运用中，权衡两者的关系。改进的模型在保护隐私的同时，仍然保证低误差。

（下转第 83 页）