中国科学 F辑: 信息科学 2009 年 第 39 卷 第 8 期: 799～808
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE IN CHINA PRESS

支持信息共享的有效的安全数据发布算法

杨晓春 ①②*, 王斌 ①②*, 于戈 ①②*

① 东北大学医学影像计算教育部重点实验室, 沈阳 110004;
② 东北大学信息科学与工程学院, 沈阳 110004
*通信作者 E-mail: yangxc@mail.neu.edu.cn, binwang@mail.neu.edu.cn, yuge@mail.neu.edu.cn

收稿日期: 2008-04-09; 接受日期: 2009-03-18
国家自然科学基金 (批准号: 60828004, 60503036)、教育部新世纪优秀人才支持计划 (批准号: NCET-06-0290) 和霍英东教育基金会青年教师基金优选资助课题 (批准号: 104027) 资助项目

摘要 在许多数据共享的应用中, 要求发布的共享信息可以保护个体的敏感信息, 例如, 患者的疾病、客户的信誉等级、雇员的工资等涉及个人隐私的信息. 同时, 另一些信息也需要被共享. 文中考虑了数据发布应用中同时存在的信息保护与信息共享问题. 攻击者能够使用发布的信息以很高的置信度推断出敏感数据的真实值. 目的是防止在发布数据上利用关联规则进行数据推演. 对攻击框架进行了形式化定义, 并研究了复杂性问题. 证明了计算安全的可发布数据表的计算复杂性为 NP 难. 根据信息发布的要求对问题进行了分类, 提出了相应的计算安全可发布数据表的算法. 基于实际数据集的实验分析表明所提出的算法可以产生近似最多的发布数据, 并且改善了现有算法的运算效率.

关键词
信息共享
数据发布
数据隐私
关联规则
推演攻击

0 引言

由于许多数据库应用都需要在网络上发布数据或者在不同组织间共享数据, 对于数据安全的需求也相应地增加了. 通常, 数据的所有者需要保护涉及到个人的敏感信息, 例如患者的疾病信息、雇员的工资信息等 [1,2]. 然而, 攻击者却能够利用公布的数据来推演出这些敏感的信息.

攻击者可以在获得的共享数据上进行关联规则挖掘, 并推演出被隐匿的敏感信息 [3]. 例如, 某医院在共享表 1 的数据的同时, 希望隐匿患者 Baker 的疾病信息 (表 1 用括号标记的值). 一种简单的方法是完全隐匿该数据项, 例如用 NULL 代替. 但是, 如果攻击者利用在可见数据之上导出的关联规则 "Gender=Male, Smoker=Yes → Disease=Lung Cancer" 发现 Baker 有较高的概率患有肺癌 (置信度是 50%), 攻击者可以推导出 Baker 的实际疾病信息. 因此, 为了可以安全发布数据, 隐匿 Baker 疾病信息的同时, 需要隐匿表中的其他信息, 使挖掘出的关联规则的置信度降低到数据所有者可以接受的范围. 例如, 一种安全数据发布的方案, 是在表 1 中隐匿敏感数据项的同时, 还需要隐匿 Smith 的疾病信息.

因此本文研究了如何保护个人的敏感信息不受推演攻击, 尤其是基于关联规则的推演攻击. 在决定哪些额外的信息需要被隐匿时, 需要考虑以下因素:

1) 隐匿哪些信息依赖于数据所有者对于攻击者能够得到的关于敏感数据项真实值的置信度的容忍程度. 在表 1 的例子中, 如果数据所有者对置信度的容忍程度是 40%, 那么由于攻击者的置信度是 50%, 大于容忍上限, 敏感信息是不安全的或者说发生了敏感信息泄漏. 因此数据所有者需要隐匿额外的信息. 如果数据所有者的容忍上限是 60%, 那么就不认为敏感信息被泄漏, 也就不需要隐匿任何额外信息. 同样, 数据的所有者在判断某条规则是否安全和某

引用格式: 杨晓春, 王斌, 于戈. 支持信息共享的有效的安全数据发布算法. 中国科学 F 辑: 信息科学, 2009, 39(8): 799–808

项敏感信息是否被泄漏时也可以考虑关联规则的最小支持度.

2) 当发布或共享数据时, 还有许多基于不同应用的特殊需求. 有些应用中, 用户可能要求某些信息一定要被发布. 例如, 必须发布关于患者吸烟习惯的信息. 为了满足这种需求, 数据拥有者在考虑如何隐匿信息以保护敏感数据的同时, 还要考虑必须发布哪些信息给使用者. 此外, 在实际应用中通常还要求数据的所有者发布尽可能多的信息.

3) 为了确保发布表中的敏感数据项是安全的, 数据的所有者要确定需要隐匿哪些额外的数据项才能保证敏感数据项的每条关联规则都是安全的. 这样的安全的发布表不是唯一的, 因此需要找到一个保留了最多信息的部分表.

表 1　包含患者信息的关系表

Name	Gender	Smoker	Disease
Baker	Male	Yes	(Lung Cancer)
Belen	Male	No	Asthma
Clark	Male	Yes	Liver Cancer
Foster	Male	Yes	Lung Cancer
Decena	Female	No	Leukemia
Jones	Female	No	Breast Cancer
Lopes	Female	No	Diabetes
May	Male	Yes	Lung Cancer
Moore	Female	No	Breast Cancer
Nicolson	Female	No	Asthma
Nichols	Female	Yes	Leukemia
Nixon	Female	No	Breast Cancer
Parks	Female	No	Breast Cancer
Richard	Male	Yes	Stomach Cancer
Smith	Male	Yes	Lung Cancer
Taylor	Male	No	Lung Cancer
Walker	Male	Yes	Stomach Cancer
Young	Male	Yes	Lung Cancer

上述挑战是研究的动因. 本文的主要贡献如下:

• 形式化了数据所有者保护敏感信息免受基于关联规则的推演攻击的数据隐私框架.

• 研究了此类问题的计算复杂性.

• 根据发布信息的需求对问题进行了分类, 并对每一类子问题给出了相应的算法.

• 在实际数据集上进行实验, 对算法进行评估.

1　相关工作

在数据安全和隐私保护方面的研究有很多, 包括 k−匿名技术、基于隐私保护的数据挖掘技术、关联规则隐匿技术等. k−匿名技术 [4~6] 关注的是隐匿敏感实体的准标识符, 而本文的目标是隐匿个体的敏感信息的真实值, 而不是其标识符. 基于隐私保护的数据挖掘 [7] 研究的是隐匿数据集中的数据分布特性而不暴露每个单独实体的信息. 最常用的方法是数据扰乱, 例如使用随机化函数. 然而, 在本文所考虑的应用中, 数据只能被隐匿而不能被修改. 一个与之密切相关的问题是关联规则隐匿 (ARH, association-rule hiding)[8~10]. 关联规则隐匿的目标是保护敏感的关联规则不被挖掘出来. 例如, 文献 [8] 中证明了要隐匿一组关联规则同时又要求对其他关联规则影响最小是 NP 难的. 这种影响可用在计算过程中损失的规则数和新引进的规则数来衡量. 文献 [8~10] 提出了几种隐匿关联规则的启发式方法. 虽然本文方法也是基于隐匿敏感关联规则的, 但是和以上的关联规则隐匿相比仍有几点基本的区别: 1) 本文的目标是隐匿敏感数据项, 关联规则隐匿只是一个中间步骤. 而 ARH 的目标是隐藏规则. 2) 大多数 ARH 研究处理的是只有 0、1 取值的数据, 而本文处理的是更为一般的关系数据, 其中大部分是分类数据. 3) 大多数 ARH 研究的目标是最小化信息隐藏对其他规则的影响. 本文则关注信息损失的最小化, 这种最小化是用隐藏的数据项的数目衡量的. 4) 本文考虑了一些特殊的应用需求, 这些需求要求某些信息必须被发布. 这是 ARH 研究没有考虑的问题. 为了用实验的方式说明本文算法与 ARH 算法的不同, 实现了其中的一个代表性算法 –CR 算法用作比较. 其他的相关工作如下: 文献 [11,12] 提出了一些依赖于数据的隐私保护方法, 文献 [13] 提出了基于加密的 XML 文档访问控制方法. 另外还有其他一些对于关系数据库推演控制的研究 [14~17], 关于 XML 安全访问控制的研究 [18,19], 以及关于空间数据访问控制的研究 [20].

2　数据隐私框架

本文考虑基于关系数据的安全发布问题. 数据

的所有者将数据存储在关系 $R(A_1,\ldots,A_k)$ 中, 根据信息共享与隐私保护的需求, 将关系 R 中数据项分为 3 类: 共享数据项、敏感数据项、普通数据项.

定义 1 共享数据项. 是数据所有者必须发布的数据项, 在表中由正号 ("+") 标识. 例如, 在表 1 中, 可以将属性 Smoker 对应的数据项标记为 "+", 表示所有这些数据项必须被发布.

定义 2 敏感数据项. 是数据所有者必须隐匿的数据项, 由负号 ("−") 标识, 表示这些数据项是必须被保护的敏感信息. 数据的所有者通过用 NULL 替换等方法隐匿每个敏感数据项. 有敏感数据项的记录叫做敏感记录. 例如, 在表 1 中, 可以将患者 Baker 的疾病信息用 "−" 标识, 表示数据所有者想保护这个患者的疾病信息. 假设所有的负数据项都来自同一个属性, 这种属性叫做敏感属性, 由 S 表示. 本文结果可直接应用于多个敏感属性的情况.

未被标识的数据项被称为普通数据项, 数据所有者既可以发布, 又可以隐匿这些数据项.

2.1 关联规则

数据所有者隐匿了某些数据项后的关系被称为部分表. 攻击者可以基于获得的部分表进行关联规则挖掘 [21], 并利用导出的关联规则进行推演. 对于从关系 R 中导出的关联规则 $r: p \to s$, 其中 p 是前项, 由条件属性分量表示, s 是后项, 由敏感数据项表示. 规则的支持度 $\mathrm{supp}(r)$ 是包含 p 和 s 的元组数 [21], 记做 $\mathrm{supp}(p \cup \{(S,s)\})$. p 的支持度是包含 p 的元组个数 [21]. 关联规则的置信度如 (1) 式所示.

$$\mathrm{conf}(r) = \frac{\mathrm{supp}(p \cup \{(S,s)\})}{\mathrm{supp}(p)}. \tag{1}$$

例如表 1, 关联规则是 {(Gender, Male), (Smoker, Yes)}→(Disease, LungCancer), 其前项的支持度是 8. 为表示简便, 将关联规则简写为 {Male, Black}→LungCancer. 第一条记录 (患者 Baker) 符合这种模式, 且与这条关联规则相关.

2.2 利用关联规则的推演攻击

设 t 是敏感记录, 它在敏感属性 S 上的分量 $t[S]$ 需要被隐匿. 攻击者利用一组条件属性 $C_1,\ldots,C_k$ 来推演敏感数据项 $t[S]$ 的值. 任何非敏感属性都可作为条件属性.

设 $\{D_1,\ldots,D_m\}$ 是敏感记录 t 的条件属性的非空子集. 相应的关联规则如 (2) 式所示:

$$r: \{(D_1,t[D_1]),(D_2,t[D_2]),\ldots,(D_m,t[D_m])\} \to t[S]. \tag{2}$$

用 $t[D_i]$ 表示记录 t 的属性 D_i 的值. 攻击者可能使用这条规则来推演记录中的敏感数据 $t[S]$. 如果这条规则的支持度和置信度足够大 (取决于数据所有者的容忍度), 敏感数据项的真实值就泄漏了.

定义 3 敏感信息泄漏. 给定关系 R 和一组敏感数据项 S. 如果针对 S 存在一条不安全的关联规则, 就认为敏感数据项被泄漏. 反之, 敏感的数据项是安全的.

本文用未被隐匿的数据项的个数来衡量发布数据的信息量. 显然, 隐匿越多的数据项, 就发布越少的信息.

定义 4 最大安全部分表. 如果存在一个安全部分表 T, 使对于任何安全部分表 T', T 包含的数据项不少于 T' 包含的数据项, 则称 T 为最大安全部分表.

3 隐匿非敏感记录的敏感属性分量

设敏感记录的敏感值为 s, 如果对发布表不做安全化处理, 会导出一条不安全的关联规则 $r: p \to s$. 根据定义, 不安全规则的置信度是大于置信度上限 minconf. 若只有非敏感记录的敏感属性分量可以被隐匿, 需要隐匿的最少的数据项的个数p满足不等式:

$$\frac{\mathrm{supp}(p \cup \{s\}) - \beta_p}{\mathrm{supp}(p)} \leqslant \mathrm{minconf}. \tag{3}$$

定义 5 候选前项集合. 对于每个敏感数据项为 s 的非敏感记录 t, S_t 为记录 t 中所有在不安全规则中出现的前项, S_t 叫做 t 的候选前项集合.

定义 6 隐匿非敏感记录中最少的敏感属性分量问题. 基于隐匿非敏感记录的敏感属性分量的前提, 对于表 R, 设 U 为由所有不安全规则的前项构成的集合, L 为所有候选前项集合构成的集合. 目的是找到这些集合的一个覆盖, 使覆盖每个前项 p 至少 β_p 次.

3.1 基本近似算法

提出了一种近似算法. 算法的核心思想是将找到最优的部分表的问题归约为 Set Multi-Cover 问题 [22]. Set Multi-Cover 问题可以被描述为给定一

个通用集合 $U=\{e_1,\ldots,e_n\}$, 和一个由 U 的子集组成的集合 $L=\{S_i|S_i\subseteq U\}$, 目标是找到一个最小的子集合 $C\subseteq L$, 这个子集合覆盖了每个元素 $e\in U$ 至少 β_p 次. 现在问题就由找到一个最优的安全部分表转变为一个 Set Multi-Cover 问题的实例.

算法 1 隐匿非敏感记录中最少的敏感属性分量的近似算法

输入: 待发布关系 R, 部分元组的敏感属性分量被标记为敏感数据项, 所有的条件属性分量被标记为共享数据项, 最小置信度 minconf, 最小支持度 minsupp;

输出: 关系 R 的最大安全部分表的近似解;

```
1.  R' = 在 R 中隐匿敏感数据项后的部分表;
2.  基于 R' 计算 (p, β_p);
3.  L = ∅;
4.  FOR (每个 (p', β_p)) {
      IF (p' ⊆ U)
        β_p = max{β_p, β_p'};
      ELSE U = U ∪ {p'};
    }
5.  FOR (每个敏感记录 r_s) {
    //由 r_s 导出的不安全关联规则的前项集合;
      S_i = {r_s[A] | r_s[A] ∈ U};
      L = L ∪ {S_i};
    }
6.  按 |S_t| 降序排列 L 中的元素;
7.  WHILE (L ≠ ∅) {
      在 L 中选择第一个不安全候选前项集合 S_t;
      FOR (每个 p ∈ S_t) {
        IF (p ∈ U) β_p = β_p − 1;
      }
      L = L − {S_t};
      从 R' 中删除 S_t 对应的敏感属性分量;
    }
8.  返回 R';
```

算法 1 在每次迭代过程中, 选择一个具有最多不安全模式的集合 S_t, 且把它加入到覆盖中. 同时, 隐匿记录 t 的敏感数据项. 因此将 S_t 集合的每个候选前项的 β_p 值减 1.

算法的性能比 (performance ratio) 如 (4) 式所示, 如果 C 是算法中的覆盖, $C*$ 是最优覆盖, 那么性能比的上限可用文献 [23] 中的 H_k 的上限计算. 其中 H_k 是序号为 k 的谐和数 (harmonic number)[22], k 是 L 中最大集合的大小, k 的最大值为 $2(n-1)$, n 是条件属性的个数.

$$\frac{|\mathcal{C}|}{|\mathcal{C}^*|}<H_{2^{(n-1)}}<(n-1)\ln 2+1. \tag{1}$$

3.2 改进的近似算法

定理 1 假设需要隐匿至少 p_1 个符合模式 p_1 的敏感属性分量以保证不会导出不安全的关联规则 $p_1\rightarrow s$, 隐匿至少 p_2 个符合模式 p_2 的敏感属性分量以保证不会导出不安全的关联规则 $p_2\rightarrow s$. 设 $p_2\subset p_1$ 且 $\beta_{p_2}\leqslant\beta_{p_1}$. 如果算法 1 能保证 $p_1\rightarrow s$ 安全, 则一定能保证 $p_2\rightarrow s$ 安全.

根据定理 1, 提出改进的近似算法, 见算法 2. 除了对算法 1 的第 4 步进行修改之外, 其他步骤与算法 1 相同.

算法 2 隐匿非敏感记录中最少的敏感属性分量的改进的近似算法

输入: 待发布关系 R, 部分元组的敏感属性分量被标记为敏感数据项, 所有的条件属性分量被标记为共享数据项, 最小置信度 minconf, 最小支持度 minsupp;

输出: 关系 R 的最大安全部分表的近似解;

```
... // 省略与算法 1 相同的部分
4.  FOR (每个 (p', β_p)) {
      IF (p' ⊆ U)
        β_p = max{β_p, β_p'};
      ELSE IF (p' ⊇ p, p ∈ U && β_p' ⩾ β_p )
        U = U ∪ {p'} − {p};
      ELSE IF (p' ⊉ p, p ∈ U|| β_p' > β_p )
        U = U ∪ {p'};
    }
... // 省略与算法 1 相同的部分
```

4 隐匿非敏感记录的条件属性分量

首先提出了一个新的数据结构 $R-$图用于求解在隐匿非敏感记录的条件属性分量条件下的近似最大发布部分表, 然后基于 $R-$图提出相应的近似算法, 并将算法扩展到具有不同敏感值的情况.

4.1 $R-$图

对于具有 k 个条件属性的敏感记录集合 $H=\{t_1,\ldots,t_m\}$, 可以导出最多 $m(2k-1)$ 条关联规则, 提出一个新的数据结构: $R-$图表示这些关联规则的偏序关系.

定义 7 $R-$图. 是一个有向无环图 $G=(V,E)$, V 是结点的集合, 结点 $v(r)\in V$ 对应从敏感记录导出的关联规则 r,$v(r)$ 由 (Item, MinHide, Exact-Match) 构成. 其中, Item 为关联规则 r 的前项和后项对应的属性分量, MinHide 指需要隐匿的条件值

的数量, 利用 (5) 式可计算 MinHide 的最小整数取值, ExactMatch 指与规则相关且与此结点的任何祖先的规则都不相关的记录的个数. E 是边的集合, 如果 $v(r) \supset v(r')$, 且 $|v(r)| = |v(r')| + 1$, 则图中有一条从结点 $v(r)$ 到结点 $v(r')$ 的有向边 e, $e \in E$. $v(r)$ 叫做 $v(r')$ 的父结点.

$$\frac{\mathrm{supp}(p(r) \cup \{s\}) - x}{\mathrm{supp}(p(r)) - x} \leqslant \mathrm{minconf}. \tag{2}$$

显然, 隐匿记录中与子结点 $v(r')$ 对应的数据项同样会降低符合父结点 $v(r)$ 的规则的置信度. 对于一个部分表, 当且仅当结点 $v(r)$ 的 MinHide 值为 0 时, 该结点对应的规则才是安全的.

4.2 基于 $R-$图的近似算法

基于 $R-$图的近似算法的执行过程是迭代的. 在每一步都找到符合一定模式的一些记录, 然后决定哪些数据项需要被隐匿. 算法在每一轮迭代后更新 $R-$图. 重复执行直到 $R-$图中所有结点都是安全的. 算法包括以下 3 个步骤:

步骤 1 选择要隐匿的记录. 如果某个结点的 ExactMatch 值不为 0, 就将此结点称为主动结点. 可以选择主动结点用来隐匿与其对应的记录以减小它的不安全的子孙结点的置信度. 算法选择顶层主动结点中的一个. 在不只一个顶层主动结点的情况下, 选择有最多的不安全子孙结点的结点. 显然, 隐匿符合这种模式的记录的条件值能够影响更多的不安全结点的置信度.

步骤 2 确定选定记录中要隐匿的数据项. 希望任意一个被隐匿的数据项可以减小尽可能多的不安全子孙结点对应的规则的置信度. 因此, 选择在结点的所有数据项中最频繁出现在子孙结点中的那个数据项进行隐匿.

步骤 3 减少要隐匿的记录的数量. 确定了要隐匿的记录中的数据项 e 后, 需要确定要隐匿 m 个相同记录中的数据项. 对于与结点 $v(r)$ 有相同的 e 值的所有不安全的子孙结点, 设 α 是其中最小的 MinHide 值. 那么 m 取 α 和规则 r 的 ExactMatch 值中较小的一个.

在隐匿了 m 条记录的 m 个数据项后, 需要更新结点的 MinHide 和 ExactMatch 值, 以使一些不安全结点变成安全的. 同时一些结点也可能变为主动结点. 算法重复以上步骤直至所有的结点对应的规则都是安全的.

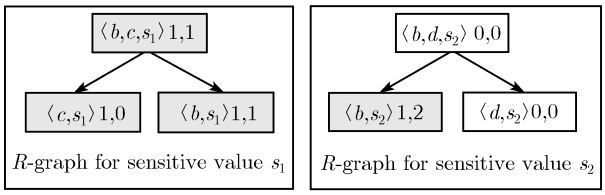
4.3 扩展到多敏感值情况

图 1 的例子显示了当敏感记录有不同的敏感数据情况的微妙性. 图 1(a) 中的表有两个不同的敏感值. 设置信度上限 minconf=0.3. 由于 $b \to s_1$ 和 $b \to s_2$ 的置信度都为 $\frac{2}{6}$, 它们都是不安全的. 如果在第 2 条记录 (b, c, s_1) 中隐匿了数据项 b, 那么规则 $b \to s_2$ 的置信度就增加到 $\frac{2}{5}$. 因此, 通过隐匿来减小一条不安全规则的置信度可能增加其他规则的置信度. 其原因在于在待发布表中存在多个不同敏感值的情况.

对 4.2 小节的算法扩展如下: 对每一个敏感值, 建立相应的 $R-$图. 例如图 1(b) 是基于图 1(a) 的表所构造的两个针对敏感值 s_1 和 s_2 的 $R-$图. 首先选择其中之一运行基于 $R-$图的算法. 要注意的是在每一轮迭代之后都要同时更新另一个 $R-$图中的 MinHide 和 ExactMatch 的值. 在一个 $R-$图上处理完毕后, 在另一个 $R-$图上运行算法. 在一轮中完成对每个 $R-$图的处理, 然后检查是否有哪个 $R-$图还含有不安全结点. 如果仍然存在不安全规则, 就需要重新对这些 $R-$图重复上述步骤. 重复进行直至不存在含有不安全结点的 $R-$图.

(a)

A_1	A_2	A_3
b	c	s_1-
b	c	s_1+
b	$*$	s_1+
b	d	s_2-
b	$*$	s_2+
b	$*$	s_2+



(b)

图 1 有不同敏感数据的多个 $R-$图. 每个*代表唯一值

(a) 表; (b) $R-$图

5 其他情况讨论

5.1 隐匿敏感记录的条件属性分量

设不安全规则 $r : p \to s$, 有 n 个敏感记录含有 s

且包含前项 p. 通过观察, 隐匿 n 条记录中的一个数据项会增加这条不安全规则的置信度. 通常, 使不安全规则 $r : p \to s$ 安全的唯一方法是隐匿至少 n 个条件分量, 每一个都来自含有 s 和前项 p 的敏感记录. 基于以上观察, 可以使用 4.2 小节的算法来求解. 只是规则 $r : p \to s$ 的 MinHide 值要设为 n (即敏感值为 s 且包含前项 p 的敏感记录个数).

5.2 隐匿条件属性分量

类似于隐匿敏感记录的条件属性分量, 对不安全的规则 $r : p \to s$, 设 n 为敏感值为 s 且包含前项 p 的敏感记录个数. 用 (5) 式计算 MinHide 值. 将规则分为 2 类: 如果规则满足 $n \leqslant$MinHide , 则将它和它的不安全规则 $r' : p' \to s$, $p' \subset p$ 归于第 1 类. 将这两者归为一类的原因是要同时减小 (p, s) 及其子集 (p', s) 构成的不安全规则的置信度. 所有其他的不安全规则被划分为第 2 类. 对于第 1 类规则, 运行隐匿敏感记录的条件属性分量的算法, 隐匿敏感记录的条件属性分量. 而对于第 2 类规则, 运行非敏感记录的敏感属性分量的算法 (算法 2), 隐匿非敏感记录的敏感属性分量.

5.3 隐匿非敏感记录的敏感属性分量和敏感记录的条件属性分量

将隐匿条件属性分量的算法作很小地改动就可以应用于这种情况. 使用 (3) 式的 β_p 值代替 MinHide 对不安全规则分类. 同样, 对于第 1 和第 2 类规则分别运行隐匿敏感记录的条件属性分量和隐匿条件属性分量的算法.

5.4 隐匿非敏感数据项或隐匿非敏感记录的数据项

下面引理 1 表明对于这种情况, 不需要隐匿非敏感记录的条件属性分量, 因此适用于非敏感记录的条件属性分量的算法同样适用于隐匿非敏感记录的数据项, 且适用于非敏感记录的敏感属性分量和敏感记录的条件属性分量的算法也适用于隐匿所有非敏感数据项.

引理 1 在隐匿非敏感数据项或隐匿非敏感记录的数据项情况下, 对于每个隐匿了非敏感记录的条件属性分量的安全部分表 T, 总存在另外一个局部表 T', 使 T' 不比 T 隐匿更多的数据项, 且 T' 中没有隐匿任何非敏感记录的条件属性分量.

定理 2 在本文中的每个子情况下, 计算最优的安全部分表的问题都是 NP 难问题.

6 实验

使用两个实际的关系数据集进行实验. 第 1 个是 UC Irvine 大学的 "adult" 数据库 (http://www.ics.uci.edu/ mlearn/MLSummary.html). 其中的记录包含成年人的性别、教育背景、年龄 (以 10 年为年龄段分组)、婚姻状况等信息. 数据集的训练集中包含 32,561 条记录, 在其中随机选取了 30 000 条用于实验. 将其中的 7 个分类属性作为条件属性, 包括性别、婚姻状况、雇员等级、最高学历、学位个数和工资的高低级别. 将职业作为敏感属性. 第 2 个数据集包含银行的客户信息, 有 30 000 条记录. 其中的 7 个分类属性作为条件属性, 分别是背景 (嘉奖或者坏记录)、贷款类型 (购房贷款或者教育贷款)、储蓄额 (低, 中或者高)、职业、婚姻状况、公寓/住房数、居民类型, 而将信用率作为敏感属性.

实验评估了各种因素对算法的效率和数据发布程度的影响. 使用 "隐匿率" 来表示数据发布的程度. 给定关系 R, 其隐匿率如 (6) 式所示:

$$\text{removal_rate} = \frac{\#_\text{of_hidden_entries_in_}R}{\#_\text{of_entries_in_}R}. \qquad (3)$$

对所提出的算法与经典的关联规则隐藏算法 CR 算法 [9] 进行了比较. 算法用 C++ 实现, 运行环境为: 奔腾 4 处理器 2.4 GHz, 512 MB 内存. 对于每一种实验设置, 运行算法 10 次并取平均值得到实验结果.

6.1 隐匿非敏感记录的敏感属性分量

在每个数据集中选取 30 000 条记录进行实验. 其中每条记录有 7 个分类属性. 在每个数据集中随机选取 2∼35 条记录作敏感记录. 设 minconf=0.6, minsupp=60, 占表大小的 0.2%.

图 2(a) 显示了敏感记录数对需要隐匿的数据项的数量的影响. y 轴是隐匿率, 表示被隐匿的数据项在表中所占的比例. 图中显示了随着敏感记录数增加, 需要隐匿的数据项的数量也随之增加. 这是因为记录数增加可能导致从发布表中导出更多的不安全关联规则. 对于数据集 1, 当有 4 个敏感记录时, 算法隐匿了 0.006% 的数据项, 当有 35 个敏感记录 时,

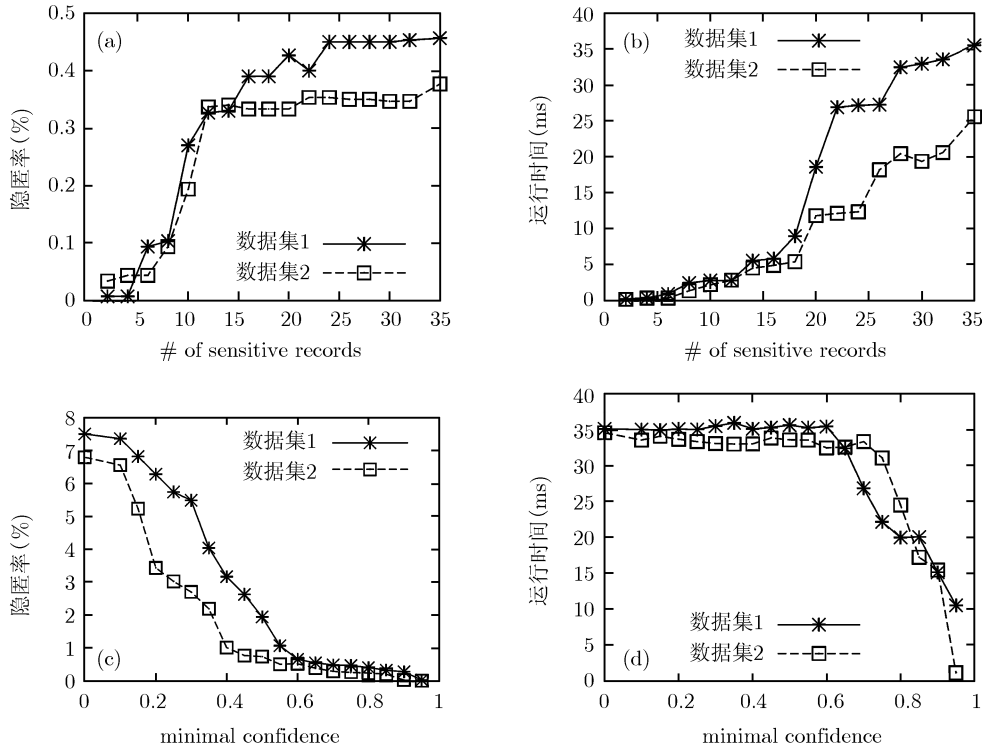


图 2　在隐匿非敏感记录的敏感属性分量前提下算法 SMC 的测试结果

(a) minimal confidence=65%; (b) minimal confidence=65%; (c) # of sensitive records=25; (d) # of sensitive records=25

算法隐匿了 0.45% 的数据项. 图 2(b) 显示了 SMC 算法的运行时间, 证明算法可以有效地 (少于 40 ms) 找到一个安全部分表. 图 2(c),(d) 对最小置信度的影响进行了测试. 任意 20 条不同的敏感记录, 且最小置信度 minconf 在 0~1 之间变化. 如图 2(c) 所示, 当最小置信度增加, 不安全的关联规则会相应减少, 从而导致少量的隐匿数据项就可以保证发布表的安全性. 同理, 如图 2(d) 显示算法的运行时间随最小置信度的增加而减少.

6.2　最小支持度和敏感规则条数的影响

利用图 3 的实验研究了最小支持度对于算法的影响. 选择 40 个具有相同值的敏感数据项. 设最小置信度为 70%, 且最小支持度的值可变. 图 3(a) 显示随着最小支持度的增加, 不安全的关联规则数在不断减少. 这导致需要隐匿的额外数据项减少. 为了更好地研究这种影响, 通过改变敏感记录数来调整不安全关联规则数, 测试与分析不安全规则数和隐匿率之间的关系, 如图 3(b) 所示. 正如期望的那样, 随着不安全关联规则数增加, 隐匿率也随之增加.

6.3　隐匿非敏感记录的条件属性分量

用数据集 1 进行实验. 设最小置信度 minconf=0.6, 最小支持度 minsupp=60(相当于表中记录数的 0.2%). 实验随机选择 10 个数据项 (有相同数据值) 作为敏感数据.

图 4(a) 显示了记录数从 500 增加到 30 000 时隐匿率的变化情况. 随着关系表规模的增大, 对敏感记录的关联规则的支持也增加. 因此要得到安全的部分表就需要隐匿更多的数据项. 4.2 小节的算法比 CR 算法隐匿少得多的数据项. 如图 4(b) 所示, 所提出的算法的运行时间也小于 CR 算法. 下面分析条件属性的个数对隐匿率和运行时间的影响. 在数据集 1 中随机选择 30 000 条数据记录进行实验. 设最小支持度 minsupp=60(占整个表中记录数的 0.2%). 条件属性数在 3~6 之间变化. 如图 4(c) 所示, 当条件属性数增加时, 有更多的条件值可被利用导出关联规则, 因此需要隐匿更多的数据项. 图 4(d) 说明了随着条

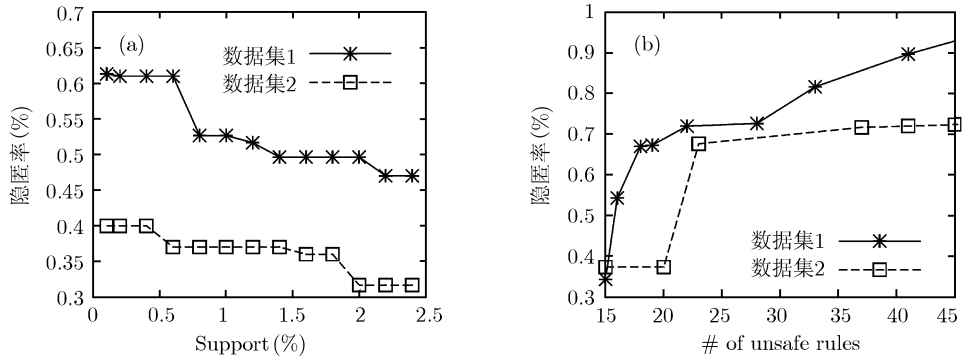


图 3　最小支持度和敏感规则条数的影响

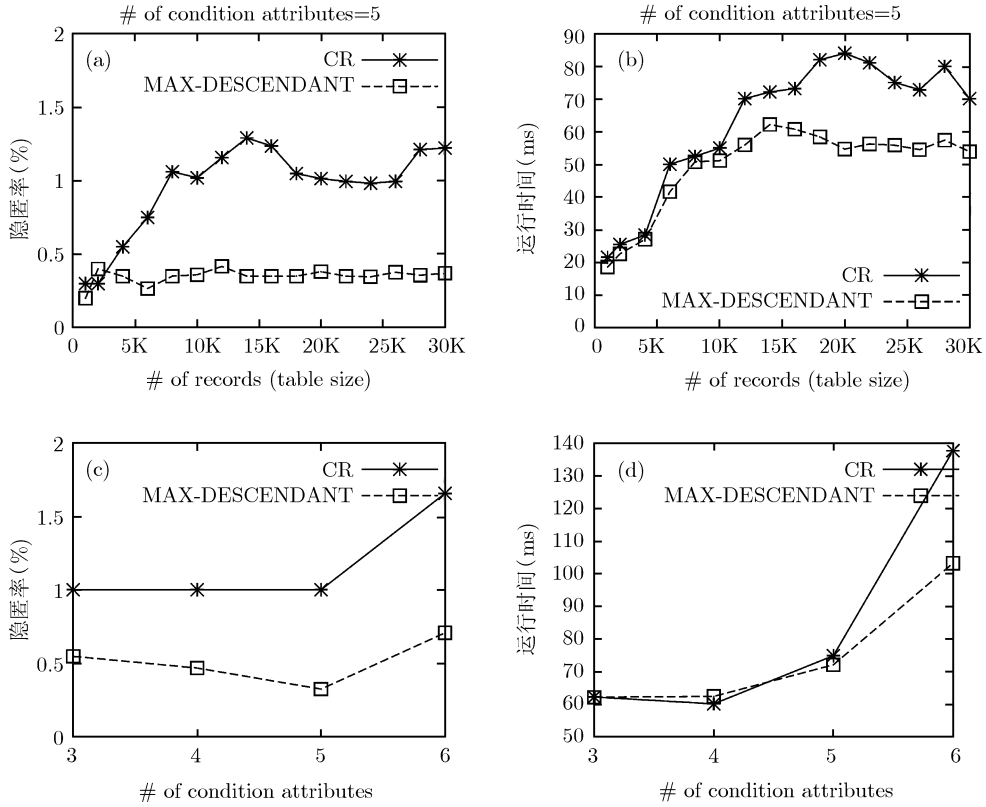


图 4　4.2 小节的算法与 CR 算法的比较 (数据集 1, 唯一敏感数据值)

件属性数增加, 两种算法都需要花费更多的运行时间, 而 4.2 小节的算法比 CR 算法性能更好. 其原因有二. 1) 4.2 小节的算法选择能影响更多不安全规则的记录模式 (如顶层结点), 而 CR 算法选择覆其原因有二: i) 4.2 小节的算法选择能影响更多不安全规则的记录模式 (如顶层结点), 而 CR 算法选择覆盖较小规则的记录, 以减小使安全规则变为不安全规则的可能性. ii) 4.2 小节的算法在选数据项进行隐匿时, 选择 $R-$图中在不安全子孙模式中出现频率较高的数据项 (它能影响更多的不安全规则). 然而, CR 算法选择隐匿表中具有最小支持度的数据项.

6.4　隐匿多敏感数据

图 5 显示了多个敏感数据值对于算法的影响. 对数据集 1, 设 minsupp=60, minconf=0.7. 随即选择 20 条敏感记录. 将不同的敏感数据值的数量控制在 1∼9 之间. 实验结果显示了当敏感数据数量增加时,

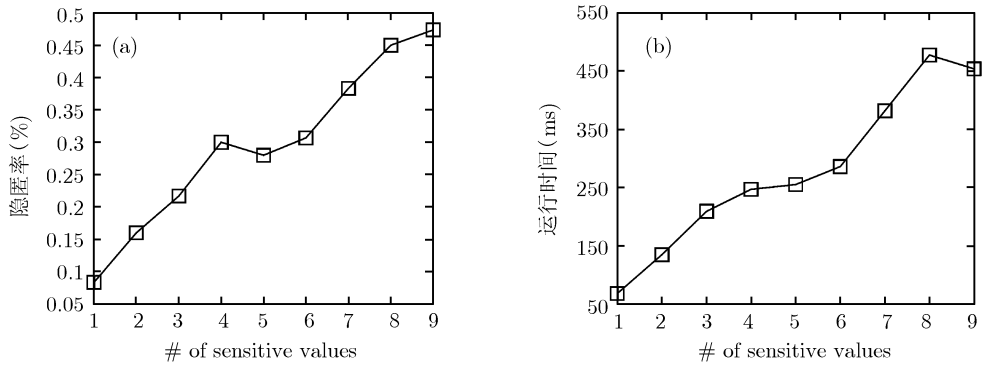


图 5 隐匿多敏感数据

为得到安全部分表, 需要隐匿更多的数据项. 其原因是隐匿一个不安全规则的条件可能增加其它不安全规则的置信度. 这就需要隐匿更多数据来保证发布表的安全性.

7 结论

对数据发布或数据共享环境中涉及到的重要的安全问题进行了研究. 涉及到的问题有: 如何保护个人的敏感信息不受基于关联规则的推演攻击. 为了确定需要发布哪些数据, 考虑了特殊的应用需求, 即有些应用中要求必须共享某些数据. 首先对问题进行了形式化描述, 并研究了问题的计算复杂性. 将问题划分为不同的子情况, 提出了计算安全部分表的有效算法. 基于实际数据集测试并分析了不同的因素对安全数据发布的影响. 还有待研究的问题包括: 可以引入其他定义信息损失的衡量方法, 例如表中不同的数据项可能对于信息损失有不同的权重等.

参考文献

1 Byun J W, Bertino E. Micro-views, or on how to protect privacy while enhancing data usability - concepts and challenges. Sigmod Rec, 2006, 35(1): 9–13

2 Sweeney L. Datafly: A system for providing anonymity in medical data. In: Lin T Y, Qian S, eds. Database Securty XI: Status and Prospects, IFIP TC11 WG11.3 Eleventh International Conference on Database Security (DBSec), IFIP Conference Proceedings, Vol 113. California: Chapman & Hall, 1997. 356–381

3 Yang X, Li C. Secure xml publishing without information leakage in the presence of data inference. In: Nascimento M A, Özsu M T, Kossmann D, et al., eds. Proceedings of the Thirtieth International Conference on Very Large Data Bases (VLDB). Toronto: Morgan Kaufmann, 2004. 96–107

4 Aggarwal G, Feder T, Kenthapadi K, et al. Anonymizing tables. In: Eiter T, Libkin L, eds. Database Theory (ICDT). Lect Notes in Computer Science, Vol 3363. Berlin: Springer-Verlag, 2005. 246–258

5 Meyerson A, Williams R. On the complexity of optimal k-anonymity. In: Deutsch A, ed. Proceedings of the Twenty-third ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS). Paris: ACM, 2004. 223–228

6 Sweeney L. k-Anonymity: a model for protecting privacy. Int J Uncertain Fuzz, 2002, 10(5): 557–570

7 Agrawal R, Srikant R. Privacy-preserving data mining. In: Chen W, Naughton J F, Bernstein P A, eds. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Dallas: ACM, 2000. 439–450

8 Atallah M, Bertino E, Elmagarmid A, et al. Disclosure limitation of sensitive rules. In: Proceedings of 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99). Chicago: IEEE Computer Society, 1999. 45–52

9 Saygin Y, Verykios V S, Clifton C. Using unknowns to prevent discovery of association rules. Sigmod Rec, 2001, 30(4): 45–54

10 Verykios V S, Elmagarmid A K, Bertino E, et al. Association rule hiding. IEEE Trans Knowl Data En, 2004, 16(4): 434–447

11 Damiani E, Vimercati S D C D, Paraboschi S, et al. Design and implementation of an access control processor for xml documents. Comput Netw, 2000, 33(1-6): 59–75

12 Damiani E, Vimercati S D C D, Paraboschi S, et al. A fine-grained access control system for xml documents. ACM Trans Inf Syst Secur, 2001, 5(2): 169–202

13 Miklau G, Suciu D. Controlling access to published data using cryptography. In: Freytag J C, Lockemann P C, Abiteboul S, et al., eds. Proceedings of 29th International Conference on Very Large Data Bases (VLDB). Berlin: Morgan Kaufmann, 2003. 898–909
14 Brodskyand A, Farkas C, Jajodia S. Secure databases: Constraints, inference channels, and monitoring disclosures. IEEE Trans Knowl Data En, 2000, 12(6): 900–919
15 Dawson S, Vimercati S D C D, Lincoln P, et al. Minimal data upgrading to prevent inference and association attacks. In: Proceedings of the Eighteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS). Philadelphia: ACM Press, 1999. 114–125
16 Jajodia S, Meadows C. Inference problems in multilevel secure database management systems. In: Inofrmation Security: An Integrated Collection of Essays. Los Alamitos: IEEE Computer Society Press, 1995. 570–584
17 Qian X, Lunt T. A semantic framework of the multilevel secure relational model. IEEE Trans Knowl Data En, 1997, 9(2): 292–301
18 Bertino E, Carminati B, Ferrari E. A secure publishing service for digital libraries of xml documents. In: Davida G I, Frankel Y, eds. Information Security, Proceedings of 4th International Conference (ISC) Malaga, Spain. Lect Notes in Computer Science, Vol 2200. Berlin: Springer-Verlag, 2001. 347–362
19 Gabillon A, Bruno E. Regulating access to xml documents. In: Olivier M S, Spooner D L, eds. Database and Application Security XV, IFIP TC11/WG11.3 Fifteenth Annual Working Conference on Database and Application Security (DBSec). Ontario: Kluwer, 2001. 299–314
20 Zhang H, He Y P, Shi Z G. A formal model for access control with supporting spatial context. Sci China Ser F-Inf Sci, 2007, 50(3): 419-439
21 Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. In: Bocca J B, Jarke M, Zaniolo C, eds. Proceedings of 20th International Conference on Very Large Data Bases (VLDB). Santiago de Chile: Morgan Kaufmann, 1994. 487–499
22 Rajagopalan S, Vazirani V V. Primal-dual RNC approximation algorithms for set cover and covering integer programs. SIAM J Comput, 1999, 28(2): 525–540
23 Hochbaum D S. Approximation Algorithms for NP-Hard Problems. Boston: PWS Publishing Co, 1997. 54–55
24 Garey M R, Johnson D S. Computer and intractability: a guide to the theory of NP-completeness. Freeman W H 1979.