差异化多敏感属性 Lq-Diversity 模型和算法①

左苏楠, 卞艺杰, 吴 慧

(河海大学 商学院, 南京 211100)

摘 要: 针对多维敏感属性数据发布面临的一般泄露、交叉泄露、相似性泄露、多维独立泄露的威胁, 本文提出了敏感属性敏感等级和敏感属性值敏感等级的概念, 基于单维 l-diversity 模型, 对各维敏感属性进行单独分组, 提出了差异化多维敏感属性模型, 验证了该模型在面向多敏感属性数据发布的安全性, 并根据此模型提出了相应的 DMSA 算法, 通过实验验证, 该算法正确可行, 且隐匿率和附加信息损失度的值都很低, 数据可用性高, 具有良好的隐私保护效果.

关键词: 多敏感属性; 敏感属性敏感等级; 敏感属性值敏感等级; lq-diversity 模型; DMSA 算法

Lq-Diversity Model and Algorithm of Differentiated Multisensitive Property

ZUO Su-Nan, BIAN Yi-Jie, WU Hui

(Business School, Hohai university, Nanjing 211100, China)

Abstract: According to the threats of general leakage, cross leakage, similar leakage and multidimensional independent leakage in redistribution of data of multi sensitive attributes, this paper puts forward the concept of sensitive attribute sensitivity level and sensitivity level of sensitive attribute values. Then, it separates each dimension of sensitive attributes based on l-diversity model. It also puts forward the lq-diversity model of differentiated multi sensitive property. Experiments prove that it is safe for the distribution of data of multiple sensitive attributes. Finally, according to lq-diversity model, the paper puts forward a corresponding DMSA algorithm, which is proved to be correct and feasible, and has low occult rate and loss degree of additional information. The result indicates data has high availability and good privacy after released with this method.

Key words: multisensitive attributes; sensitive level of sensitive attribute; sensitive level of sensitive attribute values; lq-diversity model; DMSA algorithm

数据发布隐私保护的关键是提防攻击者选择某一具体个体的信息而导致该个体隐私泄露[1]. 关于数据单次发布的隐私保护技术, 按照敏感属性进行划分, 可以分为两大类: 单一敏感属性数据发布隐私保护技术和多敏感属性数据发布隐私保护技术[2]. 目前国内外对单一敏感属性数据发布隐私保护理论研究相对成熟且受到了广泛应用, 而对于多敏感属性数据发布隐私保护研究正在发展之中. 文献[3]沿用 Anatomy 算法的思维, 提出了多维桶分组技术 MSB, 从而确保数据具有保密性. 但是 MSB 技术也存在不足之处: 其无法抵挡背景知识攻击和相似性攻击. 文献[4]提出了(p,α)-Sensitive k-anonymity 模型,该模型具有很好的隐私保护效果, 但是数据有效性较差. 文献[5]分析了 MSB 有可能会遭受推理攻击, 提出了(dou-l)-anonymity 模型, 使得攻击者即使获取了部分敏感属性值, 但其他敏感属性仍然具有较高的安全性, 有效防止了交叉泄露. 文献[6]引入聚类思想, 提出了(K,P)-Sensitive 分组算法, 从而使数据达到高保密性. 文献[7]沿用了 MSB 分组的思想, 提出了引入聚类概念的 l-cover 模型, 并提出了相对应的 lccg 算法. 文献[8,9]也对多个敏感属性数据发布隐私保护展开了相关的研讨, 并给出了相应的算法.

但是总体来说, 目前国内学者对于多维敏感属性

① 收稿时间:2015-05-30;收到修改稿时间:2015-08-17

数据发布研究较少. 对于多敏感属性数据发布面临的一般泄露、交叉泄露、相似性泄露、多维独立泄露的威胁, 我们需要从增强数据发布安全性和提高数据可用性出发, 进行深入的研究.

1 L-diversity模型

2006 年, Machanavajhala 等学者[10]针对同一个 E 中 SA 值相近甚至相等的情况, 提出了基于 K-Anonymity 模型的 l-diversity 模型, 同一个 E 中 SA 值至少有 1 个不同值:

1)一般的 l-diversity 模型: 同一等价类 E 中不同 SA 值的数量 N_{SA} 即($N_{SA} \geq 1$), 保证同一个 E 中 SA 值被窃取的概率 $\Pr obability \leq \frac{1}{l}$, 一般的 l-diversity 模型可以很好的抵制同质攻击, 但是有可能会遭到概率推理攻击; 当某一个 SA 值在 E 中出现的频率大于其他 SA 值在该 E 中出现的频率, 恶意分子就可以以高概率推出该个体的敏感属性;

2)基于熵的 l-diversity 模型: 同一等价类 E 中敏感属性值 SA 的信息熵 $Entropy(E) \geq \log l$, 其中熵定义为 $Entropy(E) = -\Sigma_{SA \in set_{SA}} V(E, SA) \log p(E, SA)$, 式中 Set_{SA} 为 SA 值的集合, V(E,SA)为 SA 值在 E 中出现的频率; E 的熵越大, 则 E 中的 SA 值的分布就越均匀, 攻击者就越难窃取 SA 值, 从而使数据具有较高保密性, 因此普通的 l-diversity 模型在数据安全性上往往不如基于熵的 l-diversity 模型;

3)递归(c, l)-diversity 模型: 如果每个等价类 E 都满足 $r_i < c(r_1 + r_2 + \cdots + r_m)$, 我们称匿名发布表符合递归(c, l)–diversity, E 中 SA 值 i 出现的次数用 r_i 表示, 其中 $r_i (1 < i < m)$, E 中不同 SA 值的个数记作 l. c 记作 E 中 SA 出现次数出现的最大值, 对 E 中 SA 值出现的最大次数进行了限定, 所以递归(c, l)-diversity 模型在数据安全性上要优于基于熵的 l-diversity 模型.

但是 l-diversity 模型面对 SA 敏感等级相近时则导致无法抵御同质性攻击 HA 和相似性攻击 SIA, 这是因为未对 SA 值在同一 E 中出现的次数给予限制, 所以也无法抵御概率攻击.

2 差异化多敏感属性数据发布模型

根据前面的介绍, 我们知道了 l-diversity 这一单敏感属性数据发布模型, 如果将这一模型直接应用于多个敏感属性数据的发布将会造成严重的隐私泄露. 但是对于找到合适的多敏感属性数据发布模型, 我们可以借鉴单一敏感属性数据发布模型. 从单一敏属性数据发布的隐私保护模型可知, 单一敏感属性数据发布为了满足 1-diversity 模型, 尽量选取了互不相等的敏感值放在同一组, 且每组选取 l 条数据, 怎样选取互不相等的 1 条敏感值是解决此问题的重点. 由于单一敏感属性表只有一维敏感数据, 那么问题就成了对一维数据的操作, 所以单一敏感属性的数据发布问题处理起来就很方便, 但对于多敏感属性数据发布问题则没那么简单. 因此, 本文将引入以下概念:

1)敏感属性的敏感等级

当基础数据表中含有多个 SA, 将人们对 SA 的保护程度的大小称之为 SA 的敏感等级. 同一敏感属性在不同基础数据表中的敏感等级是不一样的, 所以, 每个敏感属性的敏感等级必须根据实际待发布的数据源来确定. 但是如何确定 SA 的敏感等级成了问题的关键, 这时需引入信息熵的概念. 在信息论中, 信息熵可以用来衡量系统的有序化程度, 信息熵的高低决定了系统的稳定程度的大小, 且两者成正比关系. 因此可以根据 SA 的信息熵值大小来确定敏感属性敏感等级, 即敏感属性信息熵值越大, 该类 SA 的敏感等级就越高, 那么敏感属性受保护的程度也越高. 如果敏感属性的信息熵值越小, 则认为该类 SA 敏感等级较低, 那么这类敏感属性受保护的程度也就较低.

2)敏感属性值的敏感等级

多敏感属性数据表中每一维敏感属性都具有多种敏感等级不同的属性值, 对于同一 SA 敏感等级不同的属性值, 需要受到的保护程度也有所不同. 如何计算 SA 值的敏感等级是解决问题的关键, 本文引入逆文档频率 IDF(Inverse Document Frequency)[11]来对 SA 值的敏感等级进行计算. IDF 指出, 如果一个词或者一个字经常出现, 那么人们对于这个词或者字就都非常了解, 对该词或者字的获取的需求就越低. 根据逆文档频率 IDF 可知, 在多敏感属性数据表 T 中, 同一 SA 中某一 SA 的敏感程度要明显低于出现次数比较低的 SA 值. 敏感属性值出现的频率是衡量敏感属性值敏感等级的标准, 敏感属性值出现的次数与敏感属性值的敏感等级成反比, 具体来说, 敏感属性值在多敏感属性表 T 中出现的次数越低, 敏感属性值的敏感等级就越高; 反之, 敏感属性值出现的次数越高, 敏感属性

值的敏感等级就越低, 用公式表示如下:

$$SA_{ij}\_degree = \frac{c(SA_{ij})}{m} \quad (1)$$

式中 $SA_{ij-}\deg ree$ 表示敏感属性 SA_i 中敏感属性值 SA_{ij} 的敏感度, 其中 m 为多敏感属性数据表 T 中的总记录数, $c(SA_{ij})$ 表示敏感属性值 SA_{ij} 在敏感属性 SA_i 中出现的次数.

2.1 差异化多敏感属性 lq-diversity 模型

设表 $T\{QI_1, QI_2, \cdots, QI_p, SA_1, SA_2, \cdots, SA_q\}$ 是待发布的基础数据, 表中含有多个敏感属性. 其中 $QT_i(1 \le i \le p)$ 为准标识符属性, $SA_j(1 \le j \le q)$ 为敏感属性. 待发布表 T 中有 m 行数据, n 个属性, 即 $|T| = m$, 其中第 d 条数据记为 $t_d(1 \le d \le m)$, t(x)表示表中第 t 记录中属性 x 的值.

定义 1. 多敏感属性集: 是指表 T 中所有敏感属性的集合, 用符号 SA-Set 表示. 其中, $SA_{Set_b}(1 \le b \le q)$ 表示敏感属性 b 所有值的集合, 也是 SA-Set 中的第 b 维, 其中 $R(SA - Set_b)$ 表示 $SA - Set_b$ 的值域, $|SA - Set_b|$ 表示 R_{SA-Set_b} 的基数.

定义 2. 多敏感属性向量: 每一维敏感属性值可以构成一个值向量, 所有的敏感属性值向量构成一个多敏感属性向量集, 记为 $(\overrightarrow{SA_1}, \overrightarrow{SA_2}, \cdots, \overrightarrow{SA_q})$.

定义 3. 分组: 将表 T 的记录按照一定的规则把准标识符属性和多敏感属性分开用分组号链接起来, 每个分组号中存在多条记录. 任何一条数据只能拥有一个分组号, 所有的分组数据构成表 T. 表 T 的分组记为 $GT(G_1, G_2, \cdots, G_h)$, 并且每一个分组都没有交集, 用符号可以表示为 $U_{a=1}^{h} G_i = T$且$G_x \cap G_y = \phi(1 \le x \le h, 1 \le y \le h, x \ne y)$; h 表示 T 表分组总数.

定义 4. 敏感数据记录分组: 在 SAT 表中, 敏感属性集合记作 $(SA_{i1}, SA_{i2}, \cdots, SA_{in}), U_{j=1}^{n} SA_{ij} = SA_i$.

定义 5. 单敏感属性 l-diversity 模型[12]: 其中 SA 为唯一的敏感属性, SA 中出现次数最多的值是 v, c(v)为其出现的次数, 如果 $\frac{c(v)}{|SA|} \le \frac{1}{l}$, (其中 $|SA|$ 表示属性 SA 中记录总数), 那么 SA 就满足 l-diversity 模型.

定理 1. 具有多敏感属性数据表 T 中的分组 G, 若每一维 SA 都符合 l-diversity 模型性质, 则发布表中的每一维 SA 都符合 l-diversity 模型.

定义 6. 多敏感属性 l-diversity 模型: 基础数据表 T 中含有多个敏感属性, 对于发布后的表中的任意分组 G, 如果 G 中所有 SA 的每一维 SA 值都满足 l-diversity 性质, 对于分组 G 中的多敏感属性就符合 l-diversity 性质.

定义 7. 如果具有多敏感属性数据表 T 上的分组 $GT(G_1, G_2, \cdots, G_h)$ 中的任意分组都满足模型, 那么 GT 就符合模型.

定理 2. 具有多个敏感属性的数据表 T 上的分组集 $GT(G_1, G_2, \cdots, G_h)$ 满足模型性质, 发布后的数据是具有安全性的, 可以达到隐私保护的效果.

定理 3. 多敏感属性数据表 T 可以被处理成符合模型分组的充要条件是每个敏感属性 SA_i, 其中 $1 \le i \le q$, 有 $c(v_j) \le \frac{m}{l}$, 其中属性值 v_i 是在 SA_i 属性中出现次数最频繁的属性值, v_j 出现次数记为 $c(v_j)$.

2.2 DMSA 算法

定义 8. 如果多敏感属性数据表 T 上的分组 $GT(G_1, G_2, \cdots, G_h)$ 满足模型性质且每个分组 G_i 中每一维敏感属性的值都至少包含 1 个不同敏感等级的敏感属性值, 则称多敏感属性数据表 T 上的分组满足差异化多敏感属性 I_q-diversity 模型. 在某一分组 G_f 中, 如果 G_f 满足差异化多敏感属性 I_q-diversity 模型, 那么攻击者获得多敏感属性表 T 中某一记录的敏感属性值 SA_{ij} 的几率小于等于 $\frac{1}{l}$, 其中 $1 \le i \le q, 1 \le j \le m$. 每一维 SA 的多样性参数 l, 可以根据实际情况选取不一样的值.

定义 9. 对多敏感属性数据表 T 进行分组 $GT(G_1, G_2, \cdots, G_h)$, 使得每个分组 G 中每一维敏感属性分别满足 $l_1, l_2, \cdots, l_q$, 我们称 $GT(G_1, G_2, \cdots, G_h)$ 满足 $(l_1, l_2, \cdots, l_q) - diversity$. 由定理 2 可知, 如果多敏感属性数据表 T 满足 $(l_1, l_2, \cdots, l_q) - diversity$ 模型, 那么发布的数据具有良好的隐私保护能力, 发布数据是安全的.

根据上述对多敏感属性数据发布问题的分析, 我们可以发现找到多敏感属性数据表 T 上满足差异化多敏感属性 lq-diversity 模型的分组 GT 是解决多敏感属性数据发布隐私保护问题的关键点. 当多敏感属性数据表 T 中的某些分组 G 无法满足差异化多敏感属性 lq-diversity 模型时, 可以选择分组 G 中每一维敏感属性向量 $\overrightarrow{SA_j}$ 中敏感属性值的敏感等级数量最多的

记录, 在保证其满足差异化多敏感属性 lq-diversity 模型的同时, 尽可能少的选择重复记录, 隐匿这些记录的准标识属性 QI 或敏感等级较低的敏感属性值数据, 实现发布数据的支持度和置信度的平衡, 通过少量的数据损失来保证发布数据的安全性. 接下来本文将根据差异化多敏感属性 lq-diversity 模型提出相应的差异化多敏感属性数据发布算法(Differentiation Multiple Sensitive Attributes, 简称 DMSA).

2.2 DMSA 算法

输入: 待发布的多敏感属性表 $T\{QI_1,QI_2,\cdots,QI_p,SA_1,SA_2\cdots,SA_q\}$, 多样性约束 $(l_1,l_2,\cdots,l_q)$, $QT_i(1\le i\le p)$ 为准标识符属性, $SA_j(1\le j\le q)$ 为敏感属性, 多敏感属性数据表 T 中有 m 条数据, n 个属性, q 个敏感属性.

输出: 准标识符属性表 QIT 和敏感属性表 SAT, 两张表通过分组号联接起来.

Step1: 根据待发表的多敏感属性表 T 中的敏感数据(数据量要多)计算每个敏感属性 $\{SA_1,SA_2,\cdots,SA_q\}$ 的信息熵值, 根据每个敏感属性的信息熵值来确定敏感属性的敏感等级. 假设多敏感属性数据表 T 中对敏感属性进行重新排序 $\{SA_1',SA_2',\cdots,SA_q'\}$, 其中 SA_1' 的敏感属性敏感最高, SA_q' 的敏感属性敏感等级最低; 然后根据 IDF 技术确定各个 SA 的敏感值敏感等级.

Step2: 计算多敏感属性表 T 中各个敏感属性的敏感属性值的多样性总数, 每个敏感属性值的多样性个数为 $(l_1',l_2',\cdots,l_q')$, 根据 $(l_1',l_2',\cdots,l_q')$ 的值来定敏感属性多样性约束 $(l_1,l_2,\cdots,l_q)$, 其中 $(l_1\le l_1',l_2\le l_2',\cdots,l_q\le l_q')$.

Step3: 引入 Anatomy 技术的思想对多敏感属性表进行分组. 分组如下:

把每一维不同的敏感属性值看成是一个向量, 按照敏感属性值的敏感级别进行排列, 并将表中敏感属性对应的敏感值所在的第 d 条数据(用符号 t_d 表示)一一映射到每一维敏感属性对应的敏感属性值向量中. 在敏感属性等级最高的敏感属性 SA_i' 中的敏感属性值等级最高的向量中任选一条记录 t_d, 然后排除其他敏感属性和 t_d 在同一向量中的所有数据. 由于同一分组中每一维敏感属性值要多样化, 并且相近的敏感值不能单独分在一组, 所以在敏感属性等级最高的敏感属性 SA_i' 中的敏感属性值等级最高的向量中选出与 t_d 不同的数据记录, 然后排除这些数据记录在其他敏感属性的同一向量中的其他数据. 接下来对下一维敏感属性进行操作, 从剩下的数据记录中挑选一条记录. 然后以此规律循环, 找到符合 $(l_2,l_2,\cdots,l_2)-diversity$ 的分组.

Step4: 处理未被顺利分组的记录. 对于未被顺利分组的记录, 在不损害现有分组满足模型的假设下, 可以把其分配到已分好的组中. 最后, 在数据发布时, 用抑制或泛化技术隐匿未被顺利分组的数据, 本文是把不能分组的 SA 值用*来代替. 经过这四步对原始数据进行匿名化, 最后将每一分组 G 中的准标识符属性和多敏感属性分别发布成两张表: QIT 表和 SAT 表, 两张表通过分组号进行有损链接, 从而达到多敏感属性数据发布隐私保护的目的.

时间复杂度: Step1, 需要对整个多敏感属性表 T 中的敏感数据进行一次遍历, 其时间复杂度为 O(m); Step2, 运算过程中, 需要确定 q 个敏感属性的多样性约束 $(l_1',l_2',\cdots,l_q')$, 且 $(l_1\le l_1',l_2\le l_2',\cdots,l_q\le l_q')$, 因此, 其时间复杂度不会超过 $O(l_1'+l_2'+\ldots+l_q')$; Step3, 需要对每一维不同的敏感属性进行不少于 ml_q/l_q' 次的循环排除操作, 由于 $l_q\le l_q'$, 故 $ml_q/l_q'\le m$, 因此, 其时间复杂度不会超过 O(qm); Step4, 只需处理少量未被顺利分组的记录, 因此, 其时间复杂度会远小于 O(m). 所以, 整个算法耗费的时间不会超过 O(m)+ $O(l_1'+l_2'+\ldots+l_q')$ + O(qm)+ O(m)= O(qm). 由此可知, DMSA 算法的时间复杂度只与多敏感属性数据表 T 中数据条数 m 和敏感属性个数 q 有关, 在空间方面需要在内存中保存每一不同的敏感属性分组的数据记录, 而当前的内存技术足以满足该算法的需要, 因此, DMSA 算法在实时数据发布隐私保护中的应用是可行的.

3 实证分析

3.1 数据集的筛选

本文实验数据使用著名的数据挖掘测试数据集 Adult, 由 UCI Machine Learning Repository 维护提供[12]. Adult 数据集中共含有 14 个属性, 48842 个样本, 实验之前, 我们采用与文献[13,14]类似的方式先对 Adult 数据集进行预处理, 剔除掉不完整的数据, 还剩有 30162 条数据. 其中选取属性集{age, gender, native country, salary}作为准标识属性, {education, marital status, occupation, work class, Race}作为敏感属性, 如表 1, 表 2 所示. Adult 数据集提供的数据是 TXT 文本

格式, 实验时, 要将 TXT 文本转换为 EXCEL 后通过 JAVA 实现导入 Mysql 数据库中.

本次实验选择数据集中的 80%作为训练集, 20%作为测试集. 实验的硬件平台配置为 AMD Athlon(tm) II Dual-Core M320 2.09GHz 处理器, 2G 内存, 操作系统为 Windows xp SP3; 测试在 Java 环境下编程实现, 数据库使用 mysql5.0.

表 1 各个敏感属性基数

敏感属性	education	maritalstatus	occupation	work class	Race
基数	10	10	10	16	8

表 2 敏感属性信息

敏感属性个数	敏感属性向量
2	marital status, education
3	occupation , education, marital status
4	work class , education, occupation, marital status
5	education, Race, marital status, work class, occupation

3.2 评价指标

由于多敏感属性数据表 T 在分组大小不能为 1 的情况下, 为了保证发布数据安全性, 对于未能顺利被处理的数据, 就要把它分到已有的分组中, 这样就会造成信息损失, 本文对附加信息损失度[3] 进行了改进.

定义 10 (附加信息损失度). 文献[15] 中定义, 对于表 T 满足多敏感属性 l-diversity 的分组 $GT(G_1,G_2,\cdots,G_h)$, $|G_i|\geq l(1\leq i\leq h)$, h 表示分组数, l 为设定的多样性参数值, 附加信息损失度 (additional information loss) $aio=\Sigma_{1\leq i\leq h}(|G_i|-l)/hl$. 附加信息损失度的高低与数据有效性成反比, 即 aio 越低, 数据有效性越好, 数据可用性越高.

本文根据每个敏感属性值都设定了不同多样性值, 因此文献中的附加信息损失度对于多敏感属性数据发布表 T 上满足 $(l_1,l_2,\cdots,l_q)-diversity$ 的敏感属性分组 $SAT\{SA_1G_1,SA_2G_1,\cdots,SA_qG_1,SA_2G_2,\cdots,SA_qG_2,\cdots,SA_1G_h,\cdots,SA_qG_h\}$, 其中 SA_1G_1 表示敏感属性 SA_1 在分组 G_1 中的数据记录向量. 附加信息损失度用公式表示如下:

$$AIL=\frac{1}{q}\Sigma_{1\leq i\leq q}\Sigma_{1\leq j\leq h}\frac{|SA_iG_j|-l_i}{hl_i} \quad (3)$$

其中 q 表示敏感属性的总数, $|SA_iG_j|$ 表示 SA_i 在分组 G_j 中的个数, I_i 表示之前设定的敏感属性 SA_i 多样性值, h 表示分组总数.

对于未被顺利分组的记录, 在不损害现有分组满足差异化多敏感属性多样性模型的假设下, 可以把其分配到已分好的组中. 最后, 在数据发布时, 用抑制或泛化技术隐匿未被顺利分组的数据, 本文是把不能分组的SA值用*来代替. 因此还需要用隐匿率来衡量 DMSA 算法发布数据的信息损失. 隐匿率[16] 指隐匿的数据总量与表 T 中记录总数的比. 用公式表示隐匿率如下:

$$SuppRatio=\frac{\Sigma_{i=1}^{q}n_{SA_i}}{q|T|} \quad (3)$$

其中 q 表示敏感属性的个数, |T|表示表中记录数, n_{SA_i} 表示敏感属性 SA_i 中隐匿的敏感属性值的个数, 其中隐匿率越小, 信息的损失就越小, 最理想的情况是隐匿率为 0.

3.3 实验过程及结果分析

3.3.1 多样性参数的变化对 AIO 和 SuppRatio 的影响

选取敏感属性{occupation, education, marital status}, 敏感数 q=3, 数据量 Num=7000, Sweeney 等人[17] 研究指出 k 的取值一般不能大于 5 或者 6, 每维敏感属性多样性参数取以下几组数据: (2,2,2), (2,3,4), (3,3,3), (3,4,5), (4,4,4), (3,5,6), (5,5,5), 通过多样性参数的变化观察 AIO 和的变化, 结果由图 1、图 2 所示.

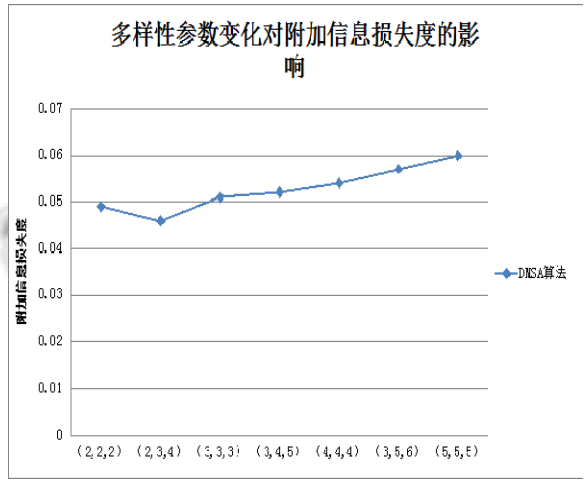

图 1 附加信息损失度多样性参数的变化

从图 1、图 2 中观察得知, 对于这组数据, 当 l 的取值越来越大时, 从待发布的表中挑选符合 $(l_1,l_2,\cdots,l_q)-diversity$ 模型的数据就会越来越少, 最后会造成不能被顺利分组的数据记录增加, 然而为了达到隐私保护的效果就必须隐藏剩下的数据, 这样就会造成巨大的信息损失. 所以我们的 1 取值不能太大.

当多样性参数$(I_1,I_2,I_3)=(2,3,4)$时，附件信息损失度和匿名率达到最低，数据有效性高．当多样性取值为(3,3,3)、(4,4,4)、(5,5,5)时，我们发现附加信息损失度和匿名率都比$(I_1, I_{,2}, I_{,3})=(2,3,4)$时高，这说明多维敏感属性数据表可以对每一维敏感属性设定不同的多样性参数 l，而且多维敏感属性设置的不同 l 值会比每一维敏感属性设定相同的 l 值的数据有效性高，隐私保护效果更好．所以对多敏感属性采取差异化策略是正确可行的．

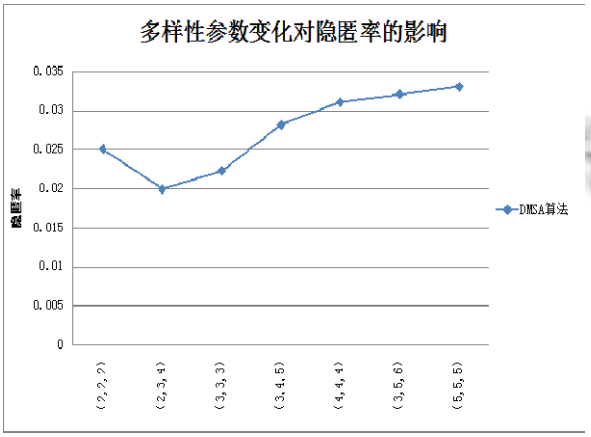

图 2　隐匿率随着多样性参数的变化

3.3.2 数据量的变化对附加信息损失度和隐匿率的影响

选取敏感属性{ occupation , education, marital status }，敏感数 q=3，设定各个敏感属性的多样性参数为 3，即$l_1=2,l_2=3,l_3=4$．数据量 Num 从 1k、逐渐递增到 10k，间隔为 1k，观察附加信息损失度和隐匿率的变化，实验结果如图 3、图 4 所示:

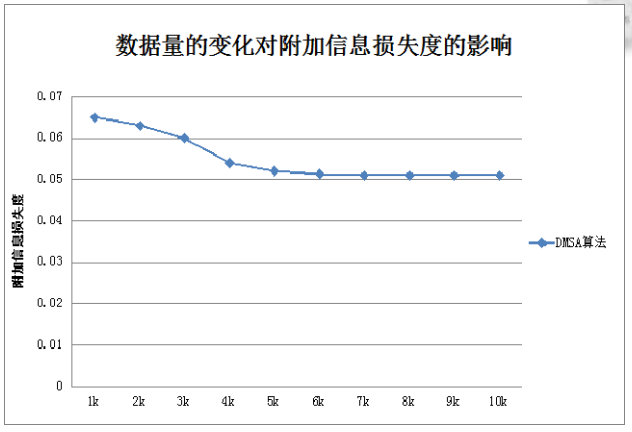

图 3　附加信息损失度随着数据量的变化

从图 3、图 4 中我们发现，不管数据量怎么变化，附加信息损失度一直低于 0.07，隐匿率一直低于 0.035，说明该算法具有良好的隐私保护效果且数据有效性高．随着数据量的增大，附加信息损失度在减小，但是并不是一直在减小，而是当数据量增大到一定的时候，附加信息损失度值减小到趋于稳定的状态．而隐匿率是随着数据的增大一直在减小，甚至隐匿率为 0，这是因为随着多敏感属性表中数据记录数的增大，数据挑选的范围就会更广，数据之间的组合也会越来越多，选到符合的分组就更加容易，从而数据基本上都可以被顺利分组，这样 aio 和就会更低，发布后的数据有效性较好，数据可用性高．

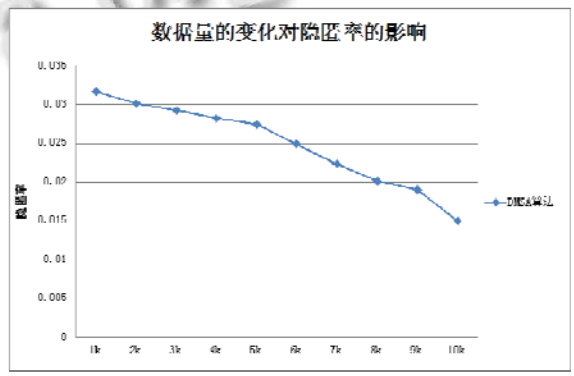

图 4　隐匿率随着数据量的变化

4　结语

本文为了使相同分组中的敏感属性具有不同的敏感值等级来实现抵制相似性攻击的目标，采取了在 l-diversity 模型的基础上，对各维敏感属性进行单独分组，提出了面向多敏感属性数据发布的模型，并验证了该模型在面向多敏感属性数据发布的安全性；然后根据模型提出了相应的 DMSA 算法，该算法基于有损链接技术，对多敏感属性数据表 T 进行匿名分组处理，保证同一分组中每一维敏感属性都满足 l-diversity 模型，确保在同一分组中敏感属性值属于不同的敏感等级，且保证没有敏感属性等级相近的敏感属性值分在同一分组中，即满足差异化多敏感属性模型，从而到达隐私保护的效果．最后从多样性参数、数据量这两个变量对 DMSA 算法的各个指标进行比较分析，最终实验验证该算法正确可行，附加信息损失度和隐匿率都很低，说明数据有效性较高，且达到了隐私保护的效果．

参考文献

1 魏琼.数据发布中的隐私保护方法的研究[博士学位论文].

武汉:华中科技大学,2008.
2 周水庚,李丰,陶宇飞,肖小奎.面向数据库应用的隐私保护研究综述.计算机学报,2009,32(5):847–860.
3 杨晓春,王雅哲,王斌,于戈.数据发布中面向多敏感属性的隐私保护方法.计算机学报,2008,31(4):574–587.
4 王茜,曾子平.(p,α)-Sensitive K-匿名隐私保护模型.计算机应用研究,2009,6:2177–2179.
5 王胜和,王佳俊,刘腾腾,倪巍伟.多维敏感属性隐私保护数据发布方法.计算机工程与应用,2011,38(4):45–50.
6 李立,袁方,郗亚辉.面向相关多敏感属性的隐私保护方法.山东大学学报,2011,46(5):82–85,90.
7 金华,刘善成,鞠时光.面向多敏感属性医疗数据发布的隐私保护技术.计算机科学,2011,38(12):171–177.
8 刘腾腾,倪巍伟,崇志宏,张勇.多维数值敏感属性隐私保护数据发布方法.东南大学学报,2010,40(4):699–703.
9 徐龙琴,刘双印.语义相似和多维加权的联合敏感属性隐私保护.计算机应用,2011,31(4):999–1002.
10 Machanavajhala A, Gehrke J, Kifer D. L-diversity: Privacy beyond k-anonymity. Proc. of the 22nd International Conference on Data Engineering(ICDE), 2006.
11 http://baike.baidu.com/view/6219237.htm.
12 http://archive.ics.uci.edu/ml/datasets/Adult.
13 倪巍伟,徐立臻,崇志宏,等.基于领域属性熵的隐私保护数据干扰方法.计算机研究与发展,2009,46(3):498–504.
14 童云海,陶有东,唐世渭,杨冬青.隐私保护数据发布中身份保持的匿名方法.软件学报,2010,21(4):771–781.
15 张兴兰,刘乐伟.面对多敏感属性的隐私保护方法.计算机与现代化,2013,38(8):168–174.
16 魏志强,康密军,贾东宁,殷波,周炜.普适计算隐私保护策略研究.计算机学报,2010,33(1): 128–138.
17 Sweeny L. K-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557–570.