

## 一种增强的个性化匿名隐私保护模型

刘英华<sup>1,2</sup>, 刘永彬<sup>1</sup>, 李广原<sup>1</sup>, 郭建威<sup>1</sup>

(1 北京科技大学 信息工程学院, 北京 100083; 2 中国青年政治学院, 北京 100089)

**摘 要:** 匿名模型是近年来隐私保护研究的热点技术之一, 主要研究如何在数据发布中避免敏感数据的泄露, 又能保证数据发布的高效性。提出了一种个性化( $\alpha[s]$ ,  $\ell$ )-多样  $k$ -匿名模型, 该方法将敏感属性泛化成泛化树, 根据数据发布中隐私保护的具体要求, 给各结点设置不同的  $\alpha$  约束, 发布符合个性化匿名模型的数据。该方法在保护隐私的同时进一步提高信息的个性化要求。实验结果表明, 该方法提高了信息的有效性, 具有很高的实用性。

**关键词:** 数据发布; 个性化隐私保护; 匿名;  $k$ -匿名

**中图分类号:** TP311

**文献标识码:** A

**文章编号:** 1000-7180(2011)08-0004-05

## An Enhanced Personalized Privacy Preserving $k$ -anonymity Model

LIU Ying-hua<sup>1,2</sup>, LIU Yong-bin<sup>1</sup>, LI Guang-yuan<sup>1</sup>, Guo Jian-wei<sup>1</sup>

(1 School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China;

2 China Youth University for Political Science, Beijing 100089, China)

**Abstract:** Recently, anonymization model is one of the hot topic techniques in privacy preserving research. The mainly research is how to avoid leakage of sensitive data in data publishing, but also ensures the efficient use of data. This paper proposed a personalized ( $\alpha[s]$ ,  $\ell$ )-diversity  $k$ -anonymity model. This method publish the personalized data though generalization technology and  $\alpha$  restriction for different code of the generalization tree. This method reserves more information while maintaining the individual privacy. The model are evaluated in an experimental scenario, reserving more information and demonstrating practical applicability of the approaches.

**Key words:** data dissemination; personalized privacy preserving; anonymization;  $k$ -anonymity

### 1 引言

随着信息化趋势的快速发展, 高效化的数据收集和数据发布为政府部门、商业机构、研究团体间的合作和研究提供了便利的条件。然而丰富的数据资源中难免会涉及到个人敏感信息, 这些数据的直接发布, 会造成个人敏感信息的泄露, 对个人隐私带来威胁。因此在数据发布时既要保证发布的数据满足数据分析和数据挖掘的需求, 又要保护个人隐私不被泄露。

1998 年 Samarati 等人<sup>[1]</sup>首次提出匿名的概念, 2002 年 Sweeney 等人<sup>[2-3]</sup>提出  $k$ -匿名 ( $k$ -anonymity) 隐私保护模型, 它是数据发布领域隐私保护的核

心模型。 $k$ -匿名使攻击者不能区分出隐私信息所属的个体, 从而保护了个体隐私<sup>[4-6]</sup>。为了达到隐匿效果,  $k$ -匿名多采取泛化或抑制的方法, 这样也不可避免地造成了原始数据的缺失。随着  $k$  值的增大, 攻击者挖掘出隐私的难度越大, 但同时信息的缺失也越多了<sup>[7-8]</sup>。2006 年 XIAO 等人<sup>[9]</sup>提出个性化匿名 (personalized anonymity) 的概念, 个性化匿名为满足个人隐私要求的最小量的概括, 最大程度地保留了原始数据中信息。

### 2 基本概念

#### 2.1 属性分类

数据匿名一般将要处理的原始数据的属性分成

收稿日期: 2011-05-16; 修回日期: 2011-06-27

基金项目: 国家自然科学基金项目(60875029)

四类:(1)个体标识属性(Individually Identifier Attribute)用来唯一地标识个体,如姓名、身份证号等;(2)准标识属性(Quasi Identifier Attribute)是一组属性,这组属性可以通过背景知识唯一地标识个体。例如出生日期和地址属性的集合可能唯一标识个体;(3)敏感属性(Sensitive Attribute)是包含敏感信息的属性,例如薪酬、宗教信仰、疾病等;(4)非敏感属性(Not Sensitive Attribute)是可以公开的属性<sup>[10-11]</sup>。

## 2.2 抑制和泛化

匿名技术主要采用泛化和抑制两种操作实现。抑制技术就是属性被隐藏,即该属性不发布,攻击者不能看到被抑制的属性。泛化技术是对原始数据的概括,例如用范围代替具体数据值,泛化虽然保护了隐私,但同时出现了数据缺损的问题。

## 3 相关研究和局限性

$k$ -匿名隐私保护模型很好地解决了链接攻击问题,但由于没有约束敏感数据,所以攻击者可以利用背景知识攻击  $k$ -匿名数据,这就是背景知识攻击(background knowledge attack)。攻击者也可以利用  $k$ -匿名数据中存在某等价类的敏感信息基本相同,结合准标识属性确定等价类的所属个体,对个体隐私造成威胁,这就是同质性攻击(Homogeneity Attack)。为了进一步提高  $k$ -匿名隐私保护模型抵御背景知识攻击和同质性攻击的能力,提出了简单  $(\alpha, k)$ -匿名隐私保护模型,这种模型要求数据满足  $k$ -匿名且某个敏感属性出现的频率小于  $\alpha$  ( $\alpha \in [0, 1]$ ),随后提出了一般  $(\alpha, k)$ -匿名隐私保护模型,这种模型要求数据满足  $k$ -匿名且每个敏感属性的出现频率均小于  $\alpha$  ( $\alpha \in [0, 1]$ )。完全  $(\alpha, k)$ -匿名模型要求数据满足  $k$ -匿名且每个敏感属性的约束值各不相同。 $p$ -sensitive  $k$ -anonymity 模型,要求数据满足  $k$ -匿名且要求每个等价类中至少有  $p$  个不同的敏感属性。这种方法解决了敏感属性泄露的问题,但数据的丢失率很高,而且当  $k$  远远大于  $p$  值时,无法抵御概率攻击。上述的匿名化模型均实现了一定程度上的数据隐私保护,但保护数据的同时由于泛化和隐匿必然带来原始数据的缺失。

$\ell$ -多样 ( $\ell$ -diversity) 模型要求每个等价类中的敏感属性至少有  $\ell$  个好表现(well-represented),好表现既可以是一个等价类中至少包含不同的  $\ell$  个敏感属性值,也可以是熵的  $\ell$ -多样 (Entropy  $\ell$ -Diversity)。Positive Disclosure-Recursive ( $c, \ell$ )-Diversity

方法和 Negative/Positive Disclosure-Recursive ( $c_1, c_2, \ell$ )-Diversity 方法在  $\ell$ -多样方法的基础上进行了扩展,提出积极泄露和消极泄露的概念,确定数据的泄露在一定范围之内。 $\ell$ -多样模型在一定程度上解决了背景知识攻击和同质性攻击,但是当原始数据过大,而  $\ell$  值较小(例如  $\ell=2$ )时会导致等价类数量庞大;而且对单个敏感属性而言,如果两个敏感属性值差异过大,确定敏感属性值的敏感度是很困难的。

针对  $\ell$ -多样模型存在的问题,提出了  $t$ -Closeness 模型,这种模型在  $k$ -匿名和  $\ell$ -多样模型的基础上综合了敏感属性的分布,要求在所有等价类内的敏感属性分布与在总体分布的差异小于  $t$ 。

个性化匿名(Personalized Anonymity)是个人可以通过分类树中的节点指定自己的隐私保护程度,实现了满足个人隐私要求的最小量的概括,最大程度地保留了原始数据中信息。

Anatomy 方法将所有的准标识属性和敏感值分别放入两个单独的表,利用有损链接,结合分组机制来保护隐私。

## 4 个性化 $(\alpha[s], \ell)$ -多样 $k$ -匿名模型

### 4.1 $k$ -匿名模型

表 1 是隐藏了身份证、姓名等个体标识属性信息的原始数据表, {性别, 年龄, 邮编} 是准标识属性, {疾病} 是敏感属性。若有某人和 Mike 是邻居, 知道 Mike 的年龄和邮编信息, 则通过表 1 和表 2 很容易知道 Mike 患了肺癌。 $k$ -匿名模型要求发布的数据表中每个准标识属性分组中至少包含  $k$  个元组。即每个 {性别, 年龄, 邮编} 准标识属性分组中至少有  $k$  个元组, 通过泛化可以将精确取值转换为概要取值, 例如邮编“10075”可以泛化成“1007\*”、“100\*\*”等, 年龄 24 可以泛化成 [20, 25]。表 3 是符合 2-匿名模型的发布数据, 即使某人知道 Mike 的年龄和邮编也无法通过表 2 和表 3 确定 Mike 是否患病? 患的是什么病?

表 1 原始数据表

| 编号 | 性别 | 年龄 | 邮编    | 疾病   |
|----|----|----|-------|------|
| 1  | 男  | 20 | 10075 | 流感   |
| 2  | 女  | 24 | 10075 | 流感   |
| 3  | 男  | 45 | 10088 | 消化不良 |
| 4  | 男  | 45 | 10088 | 胃溃疡  |
| 5  | 女  | 35 | 10087 | 肺炎   |
| 6  | 男  | 31 | 10099 | 肺癌   |

表 2 选民登记表

| 姓名   | 性别 | 年龄 | 邮编    |
|------|----|----|-------|
| Anne | 女  | 21 | 10088 |
| Sam  | 男  | 35 | 10088 |
| Mike | 男  | 31 | 10099 |

表 3 2-匿名

| 组号 | 性别 | 年龄      | 邮编      | 疾病   |
|----|----|---------|---------|------|
| 1  | *  | [20,25] | 10075   | 流感   |
| 1  | *  | [20,25] | 10075   | 流感   |
| 2  | 男  | 45      | 10088   | 消化不良 |
| 2  | 男  | 45      | 10088   | 胃溃疡  |
| 3  | *  | [31,35] | 100 * * | 肺炎   |
| 3  | *  | [31,35] | 100 * * | 肺癌   |

#### 4.2 $\ell$ -多样模型

$k$ -匿名模型虽然在一定程度上保护了患者的隐私泄露问题,但在组 1 中,两个元组无法区分,若邻居确定 Linda 家的两个孩子(一个男孩,一个女孩)患病了,且年龄在 20~25 之间,邮编是 10075,则邻居可以判定 Linda 家至少一个孩子患流感.若在  $k$ -匿名模型的基础上保证数据发布中的各组中至少包含  $\ell$  个敏感属性,则发布的数据符合  $\ell$ -多样模型.表 4 既是 2-匿名模型也是 2-多样模型.则这个邻居无法确定 Linda 家的孩子所患的疾病.

表 4 2-匿名,2-多样

| 组号 | 性别 | 年龄      | 邮编      | 疾病   |
|----|----|---------|---------|------|
| 1  | *  | [20,35] | 100 * * | 流感   |
| 1  | *  | [20,35] | 100 * * | 流感   |
| 2  | 男  | 45      | 10088   | 消化不良 |
| 2  | 男  | 45      | 10088   | 胃溃疡  |
| 3  | *  | [20,35] | 100 * * | 肺炎   |
| 3  | *  | [20,35] | 100 * * | 肺癌   |

#### 4.3 个性化( $\alpha[s], \ell$ )-多样 $k$ -匿名模型

表 1 中的敏感属性{疾病}可以根据泛化技术设计成一个分类树,各种疾病是叶子结点,中间节点是其子树各节点疾病的泛化,越接近根节点,泛化程度越高.表 5 就是表 1 中部分疾病的一个泛化树.个性化匿名模型就是根据数据发布中隐私保护度的具体要求,发布符合个性化匿名模型的数据.

针对不同的人群,敏感属性的强弱是不同的,例

如家长更关注幼儿园教师是否患有某种传染性疾病,例如肺炎,而股民更关注公司高层是否有某种致命疾病,例如癌症.在个性化匿名模型中可以根据敏感属性的强弱设置  $\alpha$  值.

表 5 分类树

|      |        |      |      |
|------|--------|------|------|
| 疾病   | 呼吸系统疾病 | 呼吸传染 | 流感   |
|      |        |      | 肺炎   |
|      |        |      | 支气管炎 |
|      |        |      | ...  |
|      | 消化系统疾病 | 胃病   | 胃溃疡  |
|      |        |      | 胃炎   |
|      |        |      | 消化不良 |
|      |        |      | ...  |
|      | 肿瘤     | 恶性肿瘤 | 肺癌   |
|      |        |      | ...  |
| 良性肿瘤 |        |      |      |
| ...  |        |      |      |
| ...  |        |      |      |

**定义 1**  $\alpha$  约束. 原始数据集  $D$  可以划分成  $n$  个等价类  $E_i (i \in [1, n])$ , 等价类  $E_i$  包含  $N_i$  个元组,  $N_i$  个元组中包含敏感属性  $s$  的记录数为  $n_i$ , 则  $\text{freq}(E_i, s) = n_i / N_i$ .  $\alpha$  是数据发布用户设定的参数,  $\alpha \in (0, 1)$ , 若等价类  $E_i$  符合  $\alpha$  约束, 则任意等价类  $E_i$  的  $\text{freq}(E_i, s) < \alpha$ .

**定义 2** 所有敏感属性值  $s$  组成敏感属性集合  $S$ , 按照树形结构分类, 构造分类树. 敏感属性值  $s$  是分类树中的一个结点,  $\text{subs}(s)$  是结点  $s$  的子树,  $\text{parent}(s)$  是结点  $s$  的父结点.

$\text{subs}(\text{呼吸传染})$  表示“呼吸传染”结点的子树, 包含“流感”、“肺炎”、“支气管炎”等结点.  $\text{parent}(\text{呼吸传染})$  表示“呼吸系统疾病”.

**定义 3** 从根结点( $\text{root}$ )到结点  $s$  的路径均是结点  $s$  的保护结点.

结点“流感”的保护结点是“疾病”、“呼吸系统疾病”、“呼吸传染”.

**定义 4** 个性化  $\alpha[s]$  约束. 原始数据集  $D$  可以划分成  $n$  个等价类  $E_i (i \in [1, n])$ , 等价类  $E_i$  包含  $N_i$  个元组,  $N_i$  个元组中包含敏感属性  $s$  的记录数为  $n_i$ , 则  $\text{freq}(E_i, s) = n_i / N_i$ .  $\alpha$  是数据发布用户设定的关于敏感属性  $s$  的参数,  $\alpha[s] \in (0, 1)$ , 若等价类  $E_i$

符合  $\alpha[s]$  约束,则等价类  $E_i$  的  $\text{freq}(E_i, s) < \alpha[s]$ .

**定义 5** 个性化  $(\alpha[s], \ell)$ -多样  $k$ -匿名模型. 原始数据集  $D$  匿名后得到数据集  $D'$ ,  $D'$  满足  $k$ -匿名, 且每个等价类  $E_i$  至少包含  $\ell$  个不同的敏感属性, 不同的等价类  $E_i$  满足个性化  $\alpha[s]$  约束.

表 6 是符合  $(\alpha[\text{流感}] = 0.3, 2)$ -多样 2-匿名的匿名数据表, 其中每个匿名后的等价类中至少包含 2 个元组(等价类 1 包含 4 个元组, 等价类 2 包含 2 个元组), 且每个等价类  $E_i$  至少包含 2 个不同的敏感属性(等价类 1 包含 4 个不同的敏感属性, 等价类 2 包含 2 个不同的敏感属性),  $\text{freq}(E_1, \text{“流感”}) = 1/4 = 0.25$ , 即  $\text{freq}(E_1, \text{“流感”}) < \alpha$ , 满足个性化  $\alpha[\text{流感}]$  约束. 所以表 6 是满足  $(\alpha[\text{流感}] = 0.3, 2)$ -多样 2-匿名的匿名模型.

表 6  $(\alpha[\text{流感}] = 0.3, 2)$ -多样 2-匿名

| 组号 | 性别 | 年龄      | 邮编      | 疾病     |
|----|----|---------|---------|--------|
| 1  | *  | [20,35] | 100 * * | 流感     |
| 1  | *  | [20,35] | 100 * * | 呼吸系统疾病 |
| 1  | *  | [20,35] | 100 * * | 肺炎     |
| 1  | *  | [20,35] | 100 * * | 肿瘤     |
| 2  | 男  | 45      | 10088   | 消化不良   |
| 2  | 男  | 45      | 10088   | 胃溃疡    |

生成个性化  $(\alpha[s], \ell)$ -多样  $k$ -匿名模型的算法:

输入: 原始数据集  $D, \alpha[s], k, \ell$ .

输出: 符合个性化  $(\alpha[s], \ell)$ -多样  $k$ -匿名模型的匿名数据表.

过程:

产生符合  $k$ -匿名,  $\ell$ -多样的数据集  $D'$

while ( $i < n$ )

while( $\text{freq}(E_i, s) > \alpha[s]$ )

任取等价类  $E_i$  中的一个元组, 用  $\text{parent}(s)$

替代  $s$

end while

$i = i + 1$

end while

返回最终生成的数据集  $D^*$

## 5 实验结果与分析

本文基于 Weka 平台封装了个性化  $(\alpha[s], \ell)$ -多样  $k$ -匿名模型算法, 对 UCI 机器学习数据集中的 adult 数据集进行了个性化  $(\alpha[s], \ell)$ -多样  $k$ -匿名模型算法和  $k$ -匿名模型算法的对比实验.

选择 {age, work class, gender, country} 作为

准标识属性, 为 adult 数据集中添加一个敏感属性字段 illness. 字段 illness 的取值是 {flu, pneumonia, bronchitis, gastriculcer, dyspepsia, gastritis}.

个性化  $(\alpha[s], \ell)$ -多样  $k$ -匿名模型算法中随着  $\alpha[s]$  值的逐步增大, 运行时间逐步下降. 因为步骤“产生符合  $k$ -匿名,  $\ell$ -多样的数据集  $D'$ ”中采用的是 TopDown 算法, 此算法中若父结点满足符合  $k$ -匿名, 则在子分支中无需细化, 所以随  $\alpha[s]$  值的逐步增大, 运行时间逐步下降. 随着  $k, \ell$  值的逐步增大, 运行时间逐步降低. 随着  $k, \ell$  值的逐步增大, 生成的等价类减少, 需要泛化的数据就越少, 因此运行时间逐步降低.  $\alpha[s]$  取值与运行时间关系见图 1. 两种匿名模型运行时间的比较见图 2.

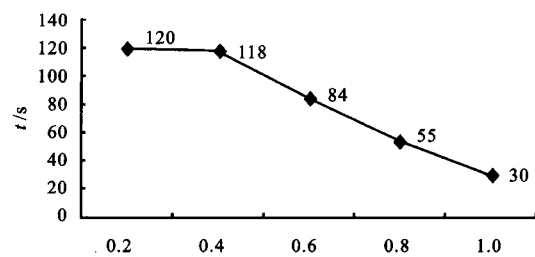


图 1  $\alpha[s]$  取值与运行时间的关系

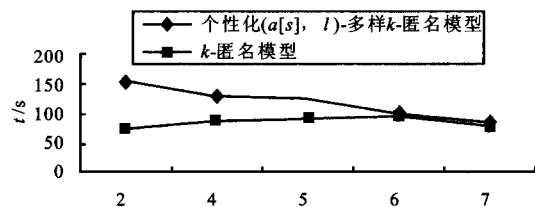


图 2 两种匿名模型运行时间的比较

个性化  $(\alpha[s], \ell)$ -多样  $k$ -匿名模型算法中随着  $\alpha[s]$  值的逐步增大, 信息损失度逐步减少. 因为随着  $\alpha[s]$  值的逐步增大, 约束越少, 需要泛化的数据就越少, 因此信息损失度逐步减少. 随着  $k, \ell$  值的逐步增大, 信息损失度逐步增大. 随着  $k, \ell$  值的逐步增大, 要求每个等价类中的元组数增加, 需要泛化的数据就越多, 因此信息损失度逐步增大.

## 6 结束语

随着数据发布中个性化要求的提高, 本文分析了现有  $k$ -匿名模型算法, 总结各种  $k$ -匿名模型算法的问题和缺陷, 针对  $k$ -匿名模型中敏感属性的个性化设置, 提出个性化  $(\alpha[s], \ell)$ -多样  $k$ -匿名模型算法, 该算法针对不同的等价类  $E$  的个性化  $\alpha[s]$  的约束, 实现数据发布中敏感属性的个性化保护.

## 参考文献:

- [1] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information[C]// Proc of the PODS '98. New York, 1998: 188.
- [2] Sweeney L. Achieving K-anonymity privacy protection using generalization and suppression[J]. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 571-588.
- [3] Sweeney L. K-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 557-570.
- [4] Wong R, L J. ( $\alpha$ , k)-anonymity: an enhanced k-anonymity model for privacy-preserving data publishing [C] // Proc of the 12th ACM SIGMOD Int'l Conf. New York, 2006: 754-759.
- [5] Truta T M, Vinay B. Privacy protection: p-sensitive k-anonymity property [C]//Proc of 22nd IEEE Int'l Conf. on Data Engineering Workshops. Washington DC: IEEE computer Society, 2006: 94-103.
- [6] Nergiz M E, Clifton C. MultiRelational k-Anonymity [C]//Proc of the IEEE 23rd Int'l Conf. Turkey: Istanbul, 2007: 1417-1421.
- [7] Machanavajjhala A, Kifer D, Gehrke J, et al.  $\ell$ -diversity: privacy beyond k-anonymity [C]//Proc of the 22nd Int'l Conf. on Data Engineering. New York: ACM, 2006: 24-35.
- [8] Li Ninghui, Li Tiancheng. T-closeness: privacy beyond k-anonymity and  $\ell$ -diversity [C]//Proc of 23rd Int'l Conf. on Data Engineering. Turkey, 2007: 106-115.
- [9] Xiao Xiaokui, Tao Yufei. Personalized privacy preservation [C]//Proc of ACM SIGMOD Conf. on Management of Data. USA, Chicago, 2006: 229-240.
- [10] Bugra Gedik, Liu Ling. Protecting location privacy with personalized k-anonymity: architecture and algorithms [J]. Mobile Computing, 2008(7): 1-18.
- [11] Xiao Xiaokui, Tao Yufei. M-invariance: towards privacy preserving republication of dynamic datasets [C]//Proc of the ACM SIGMOD Int'l Conf. on Management of Data. China: Beijing, 2007: 689-700.

## 作者简介:

刘英华 女, (1975-), 博士研究生, 讲师. 研究方向为数据挖掘、知识发现、隐私保护.

## (上接第 3 页)

- [6] Shi, Y, Casey, K, Ertl, M and Gregg, D. Virtual machine showdown: stack versus registers [J]. ACM Transactions on Architecture and Code Optimization (TACO), 2008, V4(4): 1-36.
- [7] 于森, 孙强. 对超粒度混杂技术的改进: 基于瘦虚拟机的指令集交替技术 [J]. 计算机应用, 2005, 25(12): 2808-2810.

## 作者简介:

汤战勇 男, (1979-), 博士生. 研究方向为网络与信息安全和软件安全与保护.

王怀军 男, (1981-), 博士生. 研究方向为软件安全与保护.

房鼎益 男, (1959-), 教授, 博士生导师. 研究方向为网络与信息安全、软件安全与保护、无线传感器网络.

陈晓江 男, (1973-), 博士, 副教授. 研究方向为传感器网络、信息安全.