面向敏感值的个性化隐私保护

韩建民[1],于 娟[1],虞慧群[2],贾 ▮[1]
(1.浙江师范大学数理与信息工程学院,浙江金华 321004; 2.华东理工大学计算机科学与工程系,上海 200237)

摘 要: 现有隐私保护匿名模型不能实现敏感值的个性化保护,为此,论文提出完全(α,k)-匿名模型,该模型通过设置等价类中敏感值的出现频率来实现敏感值的个性化保护.论文还提出(α,k)-聚类算法来实现各种(α,k)-匿名模型.实验表明:完全(α,k)-匿名模型能够以与其它(α,k)-匿名模型近似的信息损失量和时间代价,获得更好的隐私保护.

关键词: (α,k)-匿名模型; k-匿名; l-多样性; 同质性攻击; 背景知识攻击
中图分类号: TP309.2 **文献标识码:** A **文章编号:** 0372-2112 (2010) 07-1723-06

Individuation Privacy Preservation Oriented to Sensitive Values

HAN Jian-min[1], YU Juan[1], YU Hui-qun[2], JIA Jiong[1]
(1. *Mathematics, Physics and Information Engineering College of Zhejiang Normal University, Jinhua, Zhejiang* 321004, *China*;
2. *Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai* 200237, *China*)

Abstract: Existing anonymity models for privacy preservation cannot implement individuation preservation oriented to sensitive values. To solve the problem, the paper proposes a complete (α,k)-anonymity model which can implement individuation privacy preservation for sensitive values by setting the frequency constraints on each sensitive value in every equivalence class. The paper also proposes a (α,k)-clustering algorithm to implement all kinds of (α,k)-anonymity models . Experimental results show that the complete (α,k)-anonymity model provides better privacy preservation than other (α,k)-anonymity models with the similar information loss and execution time.

Key words: (α,k)-anonymity model; k-anonymity; l-diversity; homogeneity attack; background knowledge attack

1 引言

k-匿名[1]作为一种有效的个体隐私保护模型,近年来受到了广泛的关注[2~4],它要求发布的数据中存在一定数量(至少为 k 个)不可区分的个体,使攻击者不能标识出隐私信息所属的个体,从而保护了个体的隐私.但 k-匿名模型不能抵制同质性攻击和背景知识攻击[5],为此,该领域提出了许多敏感属性多样性模型,主要有:l-多样性模型[5],它要求每个等价类的敏感值要满足一定程度的多样性约束,以提高敏感值与其所属个体的链接难度;p-敏感性 k-匿名的模型[6],它要求每个等价类中元组个数不少于 k,敏感值种类不少于 p 个;(α,k)-匿名模型[7],它是通过控制等价类中敏感值的出现频率来实现敏感值的多样性.t-closeness 框架[8],它要求每个等价类的敏感值的分布要接近于原始数据表中敏感属性的分布;个性化匿名模型[9],该模型考虑了隐私保护的个性化需求,通过为每个用户指定不同的敏感属性泛化约束来实现个性化匿名;(k,l)-模型[10],该模型事先设定每个元组的匿名度和敏感信息的多样度,并用这两个参数作为泛化约束,实现个性化匿名度和个性化多样性.以上模型都没有考虑敏感值的个性化需求,个性化匿名模型和(k,l)-模型的个性化约束是面向个体的,而不是面向敏感值的.当数据表非常大时,为每个元组设置个性化约束将是一项庞大的工作.因此面向个体的个性化隐私保护模型具有一定的局限性.

很多情况下,不同敏感值的敏感性有很大的差异,需要保护的程度也不一样.比如患者的疾病信息,"AIDS"的敏感性比"Flu"的大,需要保护的强度要强,因此等价类中出现的频率应该相对"Flu"低,而像"Flu、Fever"等敏感值,保护强度可以弱些,等价类中出现的频率可以高些.比如,等价类中 90% 元组的敏感值是"Flu",10% 是"AIDS",那么这个等价类是安全的,但反过来,90% 元组的敏感值是"AIDS",则这个等价类是不安全的,因为攻击者更关注的是能否从匿名表中推导出

收稿日期:2008-12-01;修回日期:2010-01-18
基金项目:国家自然科学基金(No.60773094, No.60473055);上海市曙光计划(No.07SG32);上海市浦江人才计划(No.05PJ14030)

某个体是否患有“AIDS”.针对该问题,本文提出一个实现敏感值个性化约束的完全(α,k)-匿名模型,该模型为不同的敏感值设置不同频率约束α,用以控制不同的敏感值在等价类中出现的频率.本文还给出了频率α取值策略和基于聚类的匿名化算法.

2 (α,k)-匿名模型

2.1 k-匿名模型

数据表的属性按其与个体的关系可分为三类:(1)显式标识符,指能唯一标识个体身份的属性,如用户身份证号码,这些属性在数据发布前应被删除或加密;(2)准标识符 QI(Quasi Identifier),指同时存在于匿名表和外部公共数据表中,并能通过这些属性的链接来标识个体身份的一组属性,如属性组{Race,Birth,Gender};(3)敏感属性,指包含个体隐私信息的属性,如薪水、身体状况等.

定义 1 k-匿名,给定数据表 $T(A_1,A_2,\cdots,A_n)$,QI 是 T 的准标识符,$T[QI]$为 T 在 QI 上的投影(元组可重复),当且仅当在 $T[QI]$上出现的每组值至少要在 $T[QI]$上出现 k 次,则 T 满足 k-匿名.

定义 2 匿名表中的等价类,设 T'为一个 k-匿名表,把 T'在准标识符上具有相同值的元组的集合称为匿名表的等价类.

易知,k-匿名表中的每个等价类的大小至少为 k.例如,表 1 为 2 个等价类的 k-匿名表($k=4$),该表中准标识符 $QI=\{Zip\ Code, Age, Nationality\}$,敏感属性为 $Condition$.该匿名表中,每个记录至少与另外 3 个记录在准标识符上具有相同的值,这样攻击者在对该数据表在准标识符上进行链接攻击时,至少链接到该数据表的 4 个记录,不能确定具体个体的敏感值,隐私信息得到了保护.

表 1 k-匿名患者数据表($k=4$)

	Quasi Identifier			Sensitive attribute
	Zip Code	Age	Nationality	Condition
1	130**	<30	*	Heart Disease
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	3*	*	Cancer
6	130**	3*	*	Cancer
7	130**	3*	*	Cancer
8	130**	3*	*	Cancer

定义 3 同质性攻击,同质性攻击是由于 k-匿名表中某等价类的敏感信息基本相同,以至攻击者获得匿名表后,通过外表与准标识符链接确定某个体所属的等价类,即可获得相应个体的隐私信息.

例如,若攻击者通过外部公共数据表知道 Alice 的 $Zip\ Code$ 为:13010,Age 为:35,且她的疾病信息存储在表 1 中,则该攻击者就可以确定 Alice 的记录在第 2 个等价类中,因此确定其患有 Cancer.

定义 4 背景知识攻击,背景知识攻击指攻击者利用自己的背景知识来从 k-匿名表中获取个体的隐私信息.

例如,若攻击者通过外部公共数据表知道 Bob 的 $Zip\ Code$ 为:13010,Age 为:25,且他的疾病信息存储在表 1 中,则该攻击者就可以推断出 Bob 的记录在第 1 个等价类中,可能患有 Heart Disease 或者 Viral Infection,如果攻击者具有 Bob 的一些背景知识,比如知道 Bob 是运动员,心脏很健康.则攻击者可以推断出 Bob 患有 Viral Infection.

为加强 k-匿名表抵制同质性攻击和背景知识攻击的能力,Wong 等[7]提出了简单(α,k)匿名模型和一般(α,k)匿名模型.

2.2 简单(α,k)-匿名模型

定义 5 α-无关联,给定一匿名表 T',准标识符 Q,敏感属性 $S(S\notin Q)$的一个敏感值 s,设(E,s)为等价类 E 中包含敏感值 s 的元组的集合,α 为频率阈值,$0<\alpha<1$.如果 s 在每个等价类中的频率都不大于 α,即 $\forall E$,都有$|(E,s)|/|E|\leqslant\alpha$,则数据表 T' 关于准标识符 Q 和敏感值 s 是 α-无关联的.

定义 6 简单(α,k)-匿名,给定一匿名表 T',准标识符 Q,一个敏感属性值 s,如果匿名表 T'既是关于准标识符 Q 和敏感值 s α-无关联的,又是 k-匿名的,则称匿名表 T'是关于准标识符 Q 和敏感值 s 简单(α,k)-匿名.

简单(α,k)-匿名约束是面向一个特定的敏感值的,比如,表 2 是关于准标识符(Job,$Birth$,$Postcode$)和敏感值“HIV”的简单(0.4,3)-匿名的.因为该表两个等价类中“HIV”出现的频率分别为 0.33 和 0.25,不大于 0.4.表 2 虽然限制了敏感值“HIV”的频率,但没有限制其他敏感值(比如“Cancer”)的频率,攻击者可能会以较高的概率推导出患有其他疾病的个体的敏感信息.因此,该模型是不安全的.

表 2 简单(0.4,3)-匿名表

Job	Birth	Postcode	Illness
*	1975.*.*	1541	HIV
*	1975.*.*	1541	Flu
*	1975.*.*	1541	Fever
*	1975.1.*	1542	Cancer
*	1975.1.*	1542	Cancer
*	1975.1.*	1542	Flu
*	1975.1.*	1542	HIV

2.3 一般(α,k)-匿名模型

定义 7 α-稀少,设 E 为数据表的任意等价类,敏感属性为 S,$D(S)$为 S 的值域,(E,s)为 E 中包含敏感值 s 的元组的集合,α 为频率阈值,$0\leqslant\alpha\leqslant1$,如果敏感属性 S 的每个敏感值 s 在等价类中的频率都不大于 α,即:$\forall s\in D(S)$,都有$|(E,s)|/|E|\leqslant\alpha$,则等价类 E 关于属性 S 是 α-稀少的.

定义 8 一般 α-无关联,给定一匿名表 T',准标识符 Q,敏感属性 S,α 为频率阈值,$0\leqslant\alpha\leqslant1$,如果对任一等价类 E,E 是关于 S α-稀少的,则匿名表 T' 是关于准标识符 Q 和敏感属性 S 一般 α-无关联的.

定义 9 一般(α,k)-匿名,给定一匿名表 T',准标识符 Q,敏感属性 S,如果匿名表 T' 既是关于准标识符 Q 和敏感属性 S 一般 α-无关联的,又是 k-匿名的,则称该匿名表 T' 是关于准标识符 Q 和敏感属性 S 一般(α,k)-匿名的.

表 3　一般$(0.4,3)$-匿名表

Job	Birth	Postcode	Illness
*	1975.*.*	154*	HIV
*	1975.*.*	154*	Flu
*	1975.*.*	154*	Fever
*	1975.*.*	154*	Cancer
*	1975.1.*	1542	Cancer
*	1975.1.*	1542	Flu
*	1975.1.*	1542	HIV

一般(α,k)-匿名模型将简单(α,k)-匿名约束扩展到敏感属性的所有值.比如,表 3 是一般$(0.4,3)$-匿名表.一般(α,k)-匿名模型为所有的敏感值设置统一的频率约束,适应性差,因为数据表中不同的敏感值可能需要不同的频率约束.比如:患者信息中,"HIV"和"Cancer"出现的频率比较小,设为较小的频率是合理的,但"Flu"、"Fever"出现的频率比较大,如果设置与"HIV"和"Cancer"相同的频率约束,将难以生成合适的匿名表.而若把"HIV"和"Cancer"设置为与"Flu"和"Fever"相同的频率约束,则"HIV"和"Cancer"的隐私就不能得到有效的保护.

2.4 完全(α,k)-匿名模型

定义 10 完全(α,k)-匿名,给定一匿名表 T',准标识符 Q,敏感属性 S,为 S 中的每一敏感值 s 设置一个频率约束 α_s.如果匿名表 T' 是 k-匿名的,且对于 S 的任一敏感值 s,关于准标识符 Q 和敏感值 s 都是一般 α_s-无关联的,则称该匿名表 T' 为关于准标识符 Q 和敏感属性 S 完全(α,k)-匿名的.

完全(α,k)-匿名模型为每个敏感值 s 设置一个频率约束 α_s,要求等价类中各敏感值 s 均满足简单(α_s,k)-匿名约束.敏感值 s 敏感性越强,则 α_s 值应越小.比如:"HIV""Cancer"的频率约束设为 0.4,而"Flu""Fever"约束设为 0.9,则表 3 是满足这些参数的完全(α,k)-匿名约束.

完全(α,k)-匿名模型可以看作是简单(α,k)-匿名模型和一般(α,k)-匿名模型的推广,当仅为一个敏感值设置频率约束时,完全(α,k)-匿名模型退化为简单(α,k)-匿名模型.当为所有的敏感值设置相同的频率约束时,完全(α,k)-匿名模型退化为一般(α,k)-匿名模型.当所有的敏感值的频率约束都为 1 时,完全(α,k)-匿名模型就退化为 k-匿名模型.

3 (α,k)-匿名模型的聚类算法

3.1 频率约束 α_s 的设置原则

敏感值的频率约束 α 的设置应遵守 2 个原则:(1)敏感性高的敏感值,其频率约束 α_s 应相对低些,敏感性低的敏感值,其频率约束 α_s 应相对高些;(2)α_s 应该不小于该敏感值在原始数据表中的频率,否则,难以生成满足完全(α,k)-匿名约束的匿名表.设 T 为匿名表,$|T|$为表中元组个数,E 为一等价类,S 为敏感属性,v_s 为一敏感值,α_s 为 v_s 的频率约束,则 α_s 应满足式(1).

$$\alpha_s\geqslant\frac{|\{t\mid t[S]=v_s\}|}{|T|}\tag{1}$$

3.2 距离度量

本文参考了文献[11]的泛化层次树的构造方法和基于权重的层次距离,下面给出相关的定义.

定义 11 加权层次距离,设 h 为域的泛化高度,层 $1,2,\cdots,h-1,h$ 分别为从域的最泛化层到域的最具体层的层数,层 j 与层 $j-1$ 之间的权重定义为 $w_{j,j-1}$,$2\leqslant j\leqslant h$.当一个属性值由层 p 泛化到层 q,$p>q$,这个泛化的加权层次距离定义为式(2).

$$\mathrm{WHD}(p,q)=\frac{\sum_{i=q+1}^{p}w_{j,j-1}}{\sum_{j=2}^{h}w_{j,j-1}}\tag{2}$$

其中 $w_{j,j-1}$为泛化层间的泛化权重,文献[14]给出了权重的 2 种定义方法:

(1) $w_{j,j-1}=1$,$2\leqslant j\leqslant h$;(2) $w_{j,j-1}=1/(j-1)^{\beta}$

其中 $2\leqslant j\leqslant h$,β 由用户指定.

定义 12 元组泛化失真度(Distortion),设 $t=\{v_1,v_2,\cdots,v_m\}$为一元组,$t'=\{v_1',v_2',\cdots,v_m'\}$为 t 的泛化元组,Level(v_j)为 v_j 在第 j 个属性的泛化树中的层次,从 t 到 t' 泛化的失真度定义为式(3).

$$\mathrm{Distortion}(t,t')=\sum_{j=1}^{m}\mathrm{WHD}(\mathrm{level}(v_j),\mathrm{level}(v_j'))\tag{3}$$

定义 13 数据表的泛化失真度,设 D' 为数据表 D 的泛化表,t_i 为 D 中第 i 个元组,t_i' 为 t_i 的泛化,$t_i'\in$

D',将数据表 D 泛化为 D' 的失真度定义为式(4).

$$\text{Distortion}(D, D') = \sum_{i=1}^{|D|} \text{Distortion}(t_i, t_i') \tag{4}$$

定义 14 最近公共泛化,任一属性的所有可能泛化形成一个层次树,树中的每个节点对应一个值,节点的子节点对应更具体的值.元组 t_1 和 t_2 的最近公共泛化 t_{12} 的各属性的值 v_{12}^i 定义为式(5).

$$v_{12}^i = \begin{cases} v_1^i, & \text{if } v_1^i = v_2^i \\ \text{最近公共祖先的值}, & \text{否则} \end{cases} \tag{5}$$

其中 v_1^i, v_2^i 表示 t_1 和 t_2 的第 i 个属性值.

定义 15 元组间距离,设 t_1 和 t_2 为两个元组,t_{12} 为 t_1 和 t_2 的最近公共泛化,则 t_1 和 t_2 之间的距离定义为式(6).

$$\text{Dist}(t_1, t_2) = \text{Distortion}(t_1, t_{12}) + \text{Distortion}(t_2, t_{12}) \tag{6}$$

定义 16 等价类间距离,设等价类 C_1 包含 n_1 个元组,$t_1 \in C_1$,等价类 C_2 包含 n_2 个元组,$t_2 \in C_2$,t_{12} 为 t_1 和 t_2 的最近公共泛化,则 C_1 和 C_2 的距离定义为式(7).

$$\begin{aligned} \text{Dist}(C_1, C_2) = {} & n_1 \times \text{Distortion}(t_1, t_{12}) \\ & + n_2 \times \text{Distortion}(t_2, t_{12}) \end{aligned} \tag{7}$$

3.3 算法描述

两个类 C_1, C_2 能合并为一个类 C_{12} 的条件是类 C_{12} 也要满足完全(α, k)-匿名约束,即约束 1.

约束 1 设 $n = \max\{k, |C_1 + C_2|\}$,$x$ 为 $C_1 + C_2$ 中的敏感值,$(C_1 + C_2, x)$ 为 $C_1 + C_2$ 中含有敏感值 x 的元组的集合,α_x 为 x 的频率约束,那么 $\forall x$,$|(C_1 + C_2, x)|/n \leqslant \alpha_x$.

定义 17 相容等价类,将满足约束 1 的两个等价类称为可合并等价类,其中一个等价类称为另一个等价类的相容等价类.

(α, k)-聚类算法基本思想是:循环选择小于 k 的等价类 C_1,寻找 C_1 的最近相容等价类 C_2,将其合并为 C_{12}.直到所有的等价类的大小大于 k 或大小小于 k 的等价类不能再合并为止.算法步骤 2 循环结束后,若还存在大小小于 k 的等价类,那么这些等价类不存在相容等价类,因此这些等价类中的元组应该被隐匿.算法描述见图 1.

算法第 1 步,时间代价为 $O(n)$.对于算法第 2 步,假设每个等价类都不小于 k,则类的个数至少为 n/k,计算一个等价类间的距离为 $O(1)$,因等价类均比较小,因此判断相容等价类的时间代价可忽略.考虑平均情况,第一次聚类需要的计算距离 n 次,第 2 次聚类需要的计算距离为 $n-1$ 次,以此类推,直到类的个数为 n/k,所以聚类的平均时间花销为:$O(O(n) + O(n-1) + \cdots + O(n/k)) = O(n^2)$.算法第 3 步为 $O(m)$,m 为需要隐匿的元组,一般比较小.所以,总的时间花销为:$O(n) + O(n^2) + O(m) = O(n^2)$.

算法:(α, k)-聚类算法
输入:数据表 D,匿名约束 k,各个敏感值的频率约束 α_i
输出:满足完全(α, k)-匿名约束的匿名表
步骤:
1. 数据表 D 的每个元组构成一个等价类;
2. 循环,直到不存在大小小于 k 的等价类或大小小于 k 的等价类不能再合并到其他类
   (1)随机选大小小于 k 的等价类 C_1;
   (2)寻找距离 C_1 最近的相容等价类 C_2,若有多个,则选第一个选中的;
   若找到 C_2,则合并并泛化等价类 C_1 和 C_2,生成新类 C_{12};
3. 若还存在大小小于 k 的等价类,隐匿这些元组.

图 1 完全(α, k)-聚类算法

4 实验数据及结果分析

4.1 实验数据及参数

实验环境:3.0 GHz Pentium CPU,512M 内存,Linux 操作系统.实验采用 Adult 数据库,该数据库可从 http://archive.ics.uci.edu/ml/datasets/Adult 获得.目前很多泛化算法[7,8,10,12]都是以该数据库为实验数据.本文测试数据元组个数 45222,准标识符设为 6 个,敏感属性 1 个,数据表结构见表 4.

表 4 Adult 数据库描述

No.	Attribute	Type	Distinct values	Height
1	Age	Numeric	74	4
2	Workclass	Categorical	8	3
3	Education	Categorical	16	4
4	Marital Status	Categorical	7	3
5	Race	Categorical	5	2
6	Gender	Categorical	2	2
7	Occupation	Sensitive	14	

实验参数:(1)简单(α, k)-匿名模型:敏感属性值 Prof.specialty 频率约束为 0.4,其它敏感值不加约束;(2)一般(α, k)-匿名模型:所有的敏感值的频率约束为 0.4;(3) 完全(α, k)-匿名模型:不同的敏感值设置不同的频率约束,频率约束参数见表 5.

实验从信息损失量和运行时间两个角度,来分析(α, k)-聚类算法实现的 k-匿名模型以及 3 种(α, k)-匿名模型的性能.

4.2 信息损失量比较

信息损失量采用式(4)度量,图 2 为准标识符属性个数为 6,元组数为 45222,k 值变化时,4 种匿名模型的信息损失量的比较.由图 2 知:4 种模型的信息损失量都会随 k 值的增加而增加,因为 k 的增加要求每个等价类中的元组数变多,要对元组进行更高层次的泛化,

所以会产生更多的变形.图3为 k 取5,数据集元组数为45222,准标识符属性个数取3到6时,4种模型的信息损失量比较.由图3知:随着准标识符大小 $|QI|$ 增大,信息损失量逐渐增大,因为随着准标识符的属性个数增多,计算元组的匿名化的信息损失量要考虑的因素就会变多,信息损失量就会变大.图4为准标识符大小 $|QI|$ 取6,k 取5,数据集大小变化时,4种模型的信息损失量的比较.由图4知:4种模型的信息损失量都会随数据集的增大而增加,因为数据集变大会使匿名的元组数变多,需要更多的变形,所以信息损失量增加.

表5　敏感值的频率约束参数 α

Sensitive value	Freq	α	Sensitive value	Freq	α	Sensitive value	Freq	α
Tech. support	0.044	0.4	Prof. specialty	0.124	0.4	Transport. moving	0.068	0.5
Craft. repair	0.126	0.7	Handlers. Cleaners	0.043	0.5	Priv.house. serv	0.004	0.4
Other. service	0.169	0.7	Machine. op. inspct	0.061	0.5	Protective. serv	0.0	0.4
Sales	0.112	0.7	Adm. Clerical	0.094	0.7	Armed. Forces	0.0	0.4
Exec. manager	0.124	0.7	Farming. Fishing	0.031	0.5			

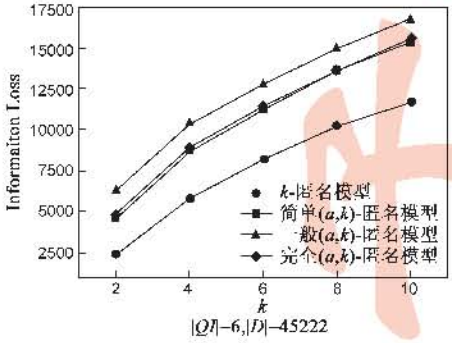

图2　不同k值下信息损量的比较

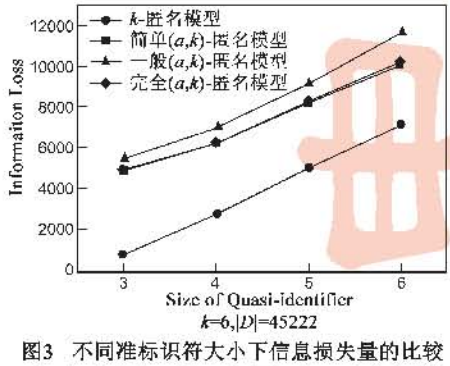

图3　不同准标识符大小下信息损失量的比较

另一方面,由图2、图3和图4知:在相同情况下,k-匿名模型的信息损失量最小,简单(α,k)-匿名模型的信息损失量其次,完全(α,k)-匿名模型与简单(α,k)-匿名模型类似,一般(α,k)-匿名模型最大.因为 k-匿名不要求敏感属性的频率约束,而简单(α,k)-匿名模型、完全(α,k)-匿名模型、一般(α,k)-匿名模型对敏感值的频率约束依次增强,所以信息损失量会依次增加.完全(α,k)-匿名模型信息损失量与简单(α,k)-匿名模型差别很小,但完全(α,k)-匿名模型保护能力最强.可见,完全(α,k)-匿名模型以与简单(α,k)-匿名模型类似的信息损失量获得更好的隐私信息保护.

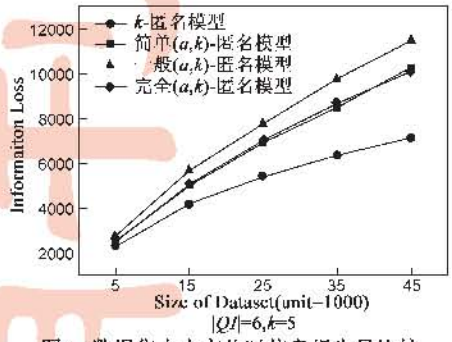

图4　数据集大小变化时信息损失量比较

4.3　执行时间比较

图5为准标识符属性个数为6,数据集大小为45222,k 分别取2,4,6,8,10时,算法实现4种匿名模型的执行时间的比较.由图5知:算法实现4种模型的执行时间都会随 k 的变大而增加,因为(α,k)-聚类算法采用自底向上的聚类策略,k 变大,聚类的次数变多,所以时间花销就会变大.图6为 k 取5,数据集大小为45222,准标识符中属性个数分别为3,4,5,6时,算法实现4种匿名模型的执行时间的比较.由图6知:算法时间花销会随着准标识符属性个数变多而变大,因为准标识符属性个数增多,每次聚类所要考虑的因素会增多,相应的计算量会变大,所以时间开销会变大.图7为准标识符属性个数为6,k 取5,数据集大小变化时,算法实现4种匿名模型的执行时间的比较.由图7知:算法实现4种模型的执行时间都会随数据集的增大而增加,因为数据集变大会使匿名的元组数变多,所以需要更多的时间开销.

另一方面,从图5、图6和图7可以看出,在相同情况下,(α,k)-聚类算法实现这4种匿名模型的时间花

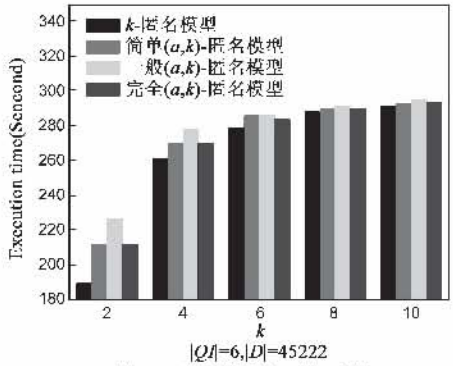

图5　不同k的执行时间比较

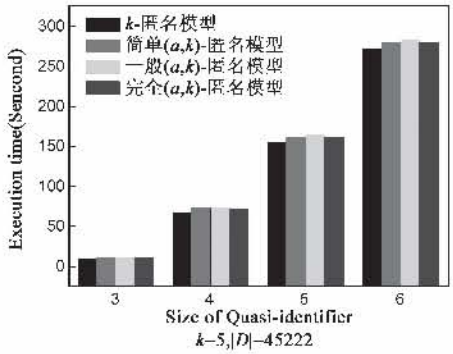

图6 不同准标识符大小下的执行时间的比较

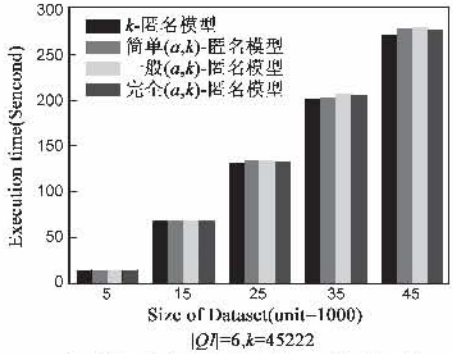

图7 数据集大小变化时执行时间的比较

销相差不大.所以,完全(α,k)-匿名模型以与其它3个模型近似的时间代价获得更好的隐私信息保护.

5 结束语

本文针对敏感值个性化保护的需求,提出完全(α,k)-匿名模型.该模型通过设置敏感值个性化约束,实现了敏感值的个性化保护.本文还提出了(α,k)-聚类算法,通过对参数α的设置,该算法可以实现k-匿名模型、简单(α,k)-匿名模型、一般(α,k)-匿名模型和完全(α,k)-匿名模型.实验结果表明,完全(α,k)-匿名模型能够以与其他(α,k)-匿名模型近似的信息损失量和时间代价,获得更好的隐私信息保护.

下一步工作:本文的模型和算法主要考虑单个敏感属性,下一步将针对多个敏感属性的敏感值个性化隐私保护问题进行研究.另外,本文主要考虑的是分类型敏感属性的隐私保护,数值型敏感属性个性化隐私保护还需要进一步的研究.

参考文献:

[1] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information (abstract)[A]. Proceedings of the 17th ACM-SIGMOD-SIGACT- SIGART Symposium on the Principles of Database Systems[C]. Seattle, WA, USA: IEEE press, 1998.188.

[2] Samarati P. Protecting respondents' identities in microdata release[J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(6): 1010 - 1027.

[3] Tiancheng Li, Ninghui Li. Towards optimal k-anonymization [J]. Data and Knowledge Engineering, 2008, 65(1): 22 - 39.

[4] 韩建民,岑婷婷,虞慧群.数据表k-匿名化的微聚集算法研究[J].电子学报,2008,36(11):2021 - 2029.
Han Jian-min, Cen Ting-ting, Yu hun-qun. Research in Micro aggregation Algorithms for K-anonymization[J]. Acta Electronica Sinica, 2008, 36(11): 2021 - 2029. (in Chinese)

[5] Machanavajjhala A, Gehrke J, Kifer D. L-diversity: privacy beyond k-anonymity[A]. Proceedings of the 22nd International Conference on Data Engineering[C]. Atlanta, GA, USA: IEEE Press, 2006.24 - 36.

[6] Truta T M, Vinay B. Privacy protection: p-sensitive k-anonymity property[A]. Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW)[C]. Washington, DC, USA: IEEE Computer Society, 2006.94.

[7] Wong C R, Li J, Fu A, et al. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing[A]. Proceedings of the 12th ACM SIGKDD Conference [C]. Philadelphia, PA: ACM Press, 2006.754 - 759.

[8] Ninghui Li, Tiancheng Li, Venkatasubramanian S. t-Closeness: privacy beyond k-anonymity and l-diversity[A]. Proceedings of the 23rd International Conference on Data Engineering (ICDE)[C]. Istanbul, Turkey: IEEE Press, 2007.106 - 115.

[9] Xiaokui Xiao, Yufen Tao. Personalized privacy preservation [A]. Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data[C]. Chicago, Illinois, USA: ACM Press, 2006.229 - 240.

[10] Zude Li, Guoqiang Zhan, Xiaojun Ye. Towards an anti-inference (k, l)-anonymity model with value association rules [A]. Database and Expert Systems Applications (DEXA)[C]. Krakow, Poland: Springer-Verlag, Berlin Heidelberg, 2006.883 - 893.

[11] Li J, Wong R, Fu A, Pei J. Achieving k-anonymity by clustering in attribute hierarchical structure[A]. Proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery (DaWak)[C]. LNCS 4081, Springer-verlag, Berlin, Heidelberg, 2006.405 - 416.

作者简介:

韩建民 男,1969年生于辽宁大连,博士,副教授,中国计算机学会会员.研究方向为信息安全.
E-mail: hanjm@zjnu.cn