第37卷 第14期 计 算 机 工 程 2011年7月
Vol.37 No.14 Computer Engineering July 2011

·专栏· 文章编号：1000—3428(2011)14—0012—06 文献标识码：A 中图分类号：TP311

基于敏感性分级的(α_i, k)-匿名隐私保护

金 华[1]，张志祥[1,2]，刘善成[1]，鞠时光[1]

(1. 江苏大学计算机科学与通信工程学院，江苏 镇江 212013；2. 紫琅职业技术学院软件工程系，江苏 南通 226002)

摘 要：(α, k)-匿名模型未考虑敏感属性不同取值间的敏感性差异，不能很好地抵御同质性攻击。同时传统基于泛化的实现方法存在效率低、信息损失量大等缺点。为此，提出一种基于敏感性分级的(α_i, k)-匿名模型，考虑敏感值之间的敏感性差异，引入有损连接思想，设计基于贪心策略的(α_i, k)-匿名聚类算法。实验结果表明，该模型能抵御同质性攻击，是一种有效的隐私保护方法。

关键词：隐私保护；(α_i, k)-匿名模型；泛化；有损连接；同质性攻击

(α_i, k)-anonymity Privacy Preservation Based on Sensitivity Grading

JIN Hua[1], ZHANG Zhi-xiang[1,2], LIU Shan-cheng[1], JU Shi-guang[1]

(1. School of Computer Science & Telecommunications Engineering, Jiangsu University, Zhenjiang 212013, China;
2. Department of Software Engineering, Zilang Vocational Technical College, Nantong 226002, China)

【Abstract】(α, k)-anonymity model can not thwart the homogeneity attack well because of the model ignoring the sensitive difference between sensitive attribute. It is achieved traditionally via generalization techniques. It also has some defects on efficiency and data distortion. So this paper proposes an improved (α_i, k)-anonymity model. It considers the sensitive difference between sensitive attribute, and designs a (α_i, k)-anonymity clustering algorithm based on greedy strategy recur to the idea of lossy join. Experimental results show that the proposed model can resist homogeneity attack and is an effective approach.

【Key words】privacy preservation; (α_i, k)-anonymity model; generalization; lossy join; homogeneity attack

DOI: 10.3969/j.issn.1000-3428.2011.14.004

1 概述

随着信息技术的高速发展，产生了大量的与个体相关的数据，这些数据中含有大量的个体隐私信息。如何在共享这些信息的同时，有效地保护好个体的隐私已成为数据库和信息安全领域亟待解决的问题[1]。目前基于匿名化的限制发布技术是隐私保护研究的热点[2-3]。其中比较典型的模型有 k-匿名模型[4]、L-多样性模型[5]、(α, k)-匿名模型[6]和 t-closness 模型[7]。

k-匿名模型要求发布数据中的个体具有不可区分性，但没有对敏感数据做任何约束，无法抵制同质性攻击和背景知识攻击；L-多样性模型对敏感信息作了约束，要求在每个等价类里敏感属性的值至少有 L 种，提高了抗击同质性攻击和背景攻击的能力，但没有控制等价类中敏感值出现的频率，无法抵制概率攻击；(α, k)-匿名模型通过控制每个等价类中敏感信息频率来实现敏感信息的多样性，从而抵制概率攻击实现隐私保护；文献[7]指出 L-多样性的缺陷，提出 t-closeness 模型，要求敏感值在每个等价类中的分布必须接近于其在原始表中的全局分布，但在等价类的记录数较少的情况下，考虑敏感值分布的意义不大，而且原数据表敏感值的分布与敏感值的敏感性并无必然关系，只考虑分布未能够实现对所有高敏感性敏感值的保护。

上述研究工作一般都是从个体的角度去考虑它们的匿名化，或仅考虑了敏感值之间的不同，而没有考虑不同的敏感值之间也存在敏感性的差异。此外，对于匿名化的方法多采用泛化/隐匿的技术实现，存在效率低、信息损失量大的缺陷。为此，本文提出基于敏感性分级的(α_i, k)-匿名模型，并引入有损连接的思想，设计基于贪心策略的(α_i, k)-匿名聚类算法。

2 相关工作

2.1 (α, k)-匿名模型

在数据匿名化的过程中，所处理的原始数据一般为数据表的形式，表中的属性按其作用效果可分为 3 类：(1)显式标识符(Explicit Identifier, EI)，能唯一标识单一个体的属性；(2)准标识符(Quasi Identifier, QI)，同时存在于隐私表与外表中，联合起来能唯一标识一个个体的一组属性；(3)敏感属性(Sensitive Attribute, SA)，包含个体隐私数据的属性。如表 1 所示，其中，Name 为显式标识符；Job、Age、Sex、Zipcode 为准标识符；Disease 为敏感属性。

表 1 原始数据

No	Name	Job	Age	Sex	Zipcode	Disease
1	Alice	clerk	23	F	13010	Fever
2	Betty	accountant	25	F	13050	Flu
3	Bob	barker	30	M	13020	Viral Infection
4	Sophia	editor	31	F	13040	Flu
5	Lucy	housewife	36	F	13220	Fever
6	Steven	manager	39	M	13221	Viral Infection
7	Bill	dustman	42	M	13226	HIV
8	Paul	wleder	45	M	13228	Cancer
9	Ellen	prof.	52	F	14850	Cancer
10	Jack	mechanic	53	M	14862	Fever
11	David	manager	61	M	14802	HIV
12	Martin	prof.	64	M	14800	Heart Disease

基金项目：国家自然科学基金资助项目(60773049)；江苏省自然科学基金资助项目(BK2010192)；江苏省中小企业技术创新基金资助项目(BC2008140)

作者简介：金 华(1977－)，男，讲师、博士研究生，主研方向：数据库安全，数据隐私保护；张志祥、刘善成，硕士研究生；鞠时光，教授、博士、博士生导师

收稿日期：2011-03-10 **E-mail：**jinhua@ujs.edu.cn

定义 1(等价类) 设 QI 为数据表 T 的准标识符，在表 T 上具有相同 QI 的所有记录的组合称为一个等价类。

定义 2(k-匿名) 如果匿名数据表 T 上任意一个等价类的记录数都大于等于 k，则称该数据表 T 满足 k-匿名。

定义 3(α-无关联) 给定一匿名表 T、准标识符 QI、敏感属性 SA(SA∉QI)的某一敏感取值 s，且 E 为该匿名表上的等价类，$|(E, s)|$记作在等价类 E 中，所有含敏感值 s 的记录数，$|E|$记作等价类 E 的所有记录数。如果对于$\forall E$，都有$|(E, s)|/|E| \leqslant \alpha(0<\alpha<1$，$\alpha$为用户设定的阈值)，则称 T 关于准标识符 QI 和敏感值 s 是α-无关联的。

定义 4((α, k)-匿名) 若匿名表 T 既满足 k-匿名，又关于其准标识符 QI 和某一敏感值 s 是α-无关联的，则称该匿名表 T 是关于准标识符 QI 和敏感值 s (α, k)-匿名的，称为特定(α, k)-匿名。如果匿名表 T 中每个等价类的每个敏感值都是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 QI 和敏感属性 SA(α, k)-匿名的，称为一般(α, k)-匿名。

根据上述定义，通过泛化方法得到的表 1 所对应的(α, k)-匿名表，如表 2、表 3 所示。

表 2 特定(0.4, 3)-匿名表

No	Job	Age	Sex	Zipcode	Disease
1	*	[20, 39]	*	130**	Fever
2	*	[20, 39]	*	130**	Flu
3	*	[20, 39]	*	130**	Viral Infection
4	*	[30, 39]	*	13***	Flu
5	*	[30, 39]	*	13***	Fever
6	*	[30, 39]	*	13***	Viral Infection
7	*	[40, 59]	*	1****	HIV
8	*	[40, 59]	*	1****	Cancer
9	*	[40, 59]	*	1****	Cancer
10	*	[50, 69]	M	148**	Fever
11	*	[50, 69]	M	148**	HIV
12	*	[50, 69]	M	148**	Heart Disease

表 3 一般(0.4, 3)-匿名表

No	Job	Age	Sex	Zipcode	Disease
1	*	[20, 39]	*	130**	Fever
2	*	[20, 39]	*	130**	Flu
3	*	[20, 39]	*	130**	Viral Infection
4	*	[30, 39]	*	13***	Flu
5	*	[30, 39]	*	13***	Fever
6	*	[30, 39]	*	13***	Viral Infection
7	*	[40, 69]	*	1****	HIV
8	*	[40, 69]	*	1****	Cancer
9	*	[40, 69]	*	1****	Cancer
10	*	[40, 69]	*	1****	Fever
11	*	[40, 69]	*	1****	HIV
12	*	[40, 69]	*	1****	Heart Disease

2.2 (α, k)-匿名模型的缺点

(α, k)-匿名模型虽能在一定程度上抵御同质性攻击和背景知识攻击，但仍存在如下缺点：

(1)α约束的适应性较差

如表 2、表 3 所示，对特定(α, k)-匿名模型仅考虑了特定敏感值 HIV 的α约束，并没有考虑其他敏感值的约束，因此，对其他敏感值而言，该模型是不安全的。而一般(α, k)-匿名模型将α约束扩展到所有敏感属性值，对所有敏感值设置统一的频率约束，适应性较差。

(2)存在同质性攻击的风险

(α, k)-匿名模型在实现α约束和 k-匿名分组时，仅考虑敏感值的不同，而未考虑敏感值之间的敏感性差异，如表 2 中的第 3 分组虽然满足α约束和 3-匿名的要求，但这 3 条记录的敏感值都属于高风险的敏感属性值，如果攻击者根据背景知识知道某个病人处于该组，虽然他不能确定病人到底得的是何种病，但也可以获知该病人的病情非常严重。因此，仍存在同质性攻击的风险。

(3)传统泛化方法效率低，信息损失量大

1)泛化存在多义性。常用的泛化方法有层次泛化、逐位泛化、分段泛化。不同泛化方法将会产生不同的泛化结果，如何确定属性的合理化泛化域目前还没有可依据的方法，是否合理仅取决于对其语义的理解，因此，对属性值的泛化存在多义性。

2)计算复杂性高，信息损失量大。已经证明采用泛化/隐匿获得最优 k-匿名是 NP-hard 问题[8]，而且在一组准标识符中也很难找到最优的泛化次序，若没有谨慎的组合，匿名化会损失大量的信息。特别当 QI 属性较多，α、k 的约束更为严格时，信息损失更明显。

3 基于敏感性分级的(α_i, k)-匿名模型

针对上述缺陷，本文提出了基于敏感性分级的(α_i, k)-匿名模型。对个人隐私而言，不同的取值往往具有不同的敏感性。例如：HIV 这类比较严重的疾病其敏感性远高于 fever 这类一般的病情；高收入人群对于薪水的隐私保护程度也高于低收入者对于工资的隐私保护度。因此，要实现更好的隐私保护，就需要对敏感属性值进行敏感性分级。下文以 Disease 属性为例来讨论敏感属性的敏感性分级。根据其语义敏感性由高到低分为分为 $L_1, L_2, \cdots, L_\lambda$ 级，Disease 属性的分类如表 4 所示。

表 4 Disease 属性分类

级别	属性值
1	HIV, Cancer, Syphilis, …
2	Neurasthenia, Sclerosis, Zona, …
3	Diabetes, phtisis, Heart Disease, …
4	Flu, Viral Infection, Anemia, …
5	Fever, Cold, Indigestion, …

在实现(α_i, k)-匿名分组时，首先应该避免具有相同级别敏感值的记录分在同一组。其次还应该考虑到组与组之间的差异性也有高低之分，如表 4 中级别 1 和级别 5 之间的差异性要高于级别 1 和级别 2 之间的差异性。为表示这一差异性，通过定义敏感性距离[9]来进行描述。

定义 5(敏感性距离) 设 $D(A)$为属性 A 的级别域，$|D(A)|$表示该属性的级别数，L_i、L_j 为该域中的 2 个级别，则两级别间的敏感距离为：

$$d(L_i, L_j) = |L_i - L_j| \tag{1}$$

显然 $d(L_i, L_j)=d(L_j, L_i)$，因此，可以定义一个特定的等价类矩阵来描述某一分组(等价类)中所有敏感属性值之间的敏感性距离。设等价类 E 有 q 条记录；C_{ij} 表示第 i 条记录和第 j 条记录之间的级别距离，故有 $C_{ij}=d(L_i, L_j)$。则等价类中所有敏感属性值之间的敏感性距离可以用矩阵 ***Matrix***(E)来表示。比如，对于等价类 $E_1=\{1, 2, 3, 3\}$和 $E_2=\{3, 4, 5, 5\}$，它们的距离为：

$$\boldsymbol{Matrix}(E_1) = \boldsymbol{Matrix}(E_2) = \begin{pmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 0 \\ 2 & 1 & 0 & 0 \end{pmatrix}$$

由矩阵定义可知：$C_{ii}=0$，$C_{ij}=C_{ji}$，因此，可以用所有 C_{ij} 的和来度量该等价类 E 的敏感性距离，从而表示等价类中敏感信息的总体敏感性及其差异。

定义 6(等价类敏感性距离度量) 设 E 为等价类，含有 q 条记录，$|D(A)|$为域 $D(A)$的级别数；W_k 表示第 i 条记录和第 j 条记录之间级别距离的权重，则等价类距离的度量定义为：

$$D_s(E)=\frac{\sum_{i=1}^{q}\sum_{j=i+1}^{q}(w_k\times C_{ij})}{(q\times(q-1)/2)\times(|D(A)|-1)} \quad (2)$$

其中，W_k 中的 k=max(L_i, L_j)。对于表 5，可以设 W_k={0.85, 0.9, 0.95, 1}，即表示：级别 5 到级别 4、3、2、1 的权重都为 1；级别 4 到级别 3、2、1 的权重是 0.95；级别 3 到级别 2、1 的权重是 0.9；级别 2 到级别 1 的权重是 0.85。因此，$D_s(E_1)$= D_s(1, 2, 3, 3)=1.042，$D_s(E_2)$=D_s(2, 3, 5, 5)=1.158，较好地反映了 2 个等价类之间的敏感性差异。

定义 7(敏感性分级(α_i, k)-匿名) 给定一匿名表 T'、准标识符 QI、敏感属性 SA，将 SA 的取值按敏感性由高到低分为 $L_1, L_2,\cdots, L_\lambda$ 级，为 SA 中的每一敏感级设置一个频率约束$\alpha_1, \alpha_2, \cdots, \alpha_\lambda$。如果匿名表 T'是 k-匿名的，且对 SA 中的任一敏感级 S_i，关于准标识符 QI 和敏感级 S_i 都是一般α_i-无关联的，则称该匿名表 T'关于准标识符 QI 和敏感属性 SA 满足敏感性分级(α_i, k)-匿名。

敏感性分级(α_i, k)-匿名为敏感属性的每一敏感级别 L_i 设置一个频率约束α_i，要求所有等价类中的敏感值都满足其所属敏感级的频率约束，因此适用性比较强。同时在实现该模型的α_i 约束和 k-匿名分组时，需要充分考虑每个等价类分组中敏感值之间的敏感区分度。传统的泛化技术本身就存在计算复杂度高和信息损失大的缺陷，而且其分组的策略是以 QI 属性为出发点的，如果再要考虑不同敏感值之间的敏感性差异，将会有更高的计算复杂度，因此，本文引入有损连接的思想从敏感属性的角度来实现分组。

4 (α_i, k)-匿名模型的发布方法

4.1 有损连接

基于有损连接的思想[10]是将泛化方法的单表发布改为将准标识符 QI 和敏感属性 SA 分开用 2 个表进行发布，两表之间使用一个等价类 class ID 相关联，从而通过部分连接的损失来达到隐私保护的目的。表 5~表 7 给出了基于有损连接思想对表 1 进行(α, k)-匿名发布的示例。

表 5 一般(0.4, 3)-匿名临时表

No	Job	Age	Sex	Zipcode	Disease	Class ID
1	clerk	23	F	13010	Fever	1
2	accountant	25	F	13050	Flu	1
3	barker	30	M	13020	Viral Infection	1
5	housewife	36	F	13220	Fever	2
7	dustman	42	M	13226	HIV	2
8	wleder	45	M	13228	Cancer	2
4	editor	31	F	13040	Flu	3
6	manager	39	M	13221	Viral Infection	3
9	prof.	52	F	14850	Cancer	3
10	mechanic	53	M	14862	Fever	4
11	manager	61	M	14802	HIV	4
12	prof.	64	M	14800	Heart Disease	4

表 6 一般(0.4, 3)-匿名 QI 表

Job	Age	Sex	Zipcode	Class ID
clerk	23	F	13010	1
accountant	25	F	13050	1
barker	30	M	13020	1
housewife	36	F	13220	2
dustman	42	M	13226	2
wleder	45	M	13228	2
editor	31	F	13040	3
manager	39	M	13221	3
prof.	52	F	14850	3
mechanic	53	M	14862	4
manager	61	M	14802	4
prof.	64	M	14800	4

表 7 一般(0.4, 3)-匿名 SA 表

Class ID	Disease
1	Fever
1	Flu
1	Viral Infection
2	Fever
2	HIV
2	Cancer
3	Flu
3	Viral Infection
3	Cancer
4	Fever
4	HIV
4	Heart Disease

原始表 1 利用泛化的方法可以得到表 2、表 3 的(0.4, 3)-匿名，而基于有损连接的方法是：以满足(α, k)-匿名的原则对原始表 1 进行等价类的划分，并为每个等价类分配一个唯一的 ID 号，属于同一个等价类的记录具有同样的 ID 号，从而得到如表 5 所示的临时表。然后根据 ID，将 QI 和 SA 分表发布得到如表 6、表 7 所示的 2 张表。通过连接的损失来实现(0.4, 3)-匿名隐私保护。但由于(α, k)-匿名模型没有考虑敏感性的差异，在第 2 个等价类中有 2 个高敏感性的敏感值被分在了同一组。这样高敏感性的敏感值容易产生隐私泄露。

4.2 (α_i, k)-匿名模型发布算法

本文设计了一种基于贪心策略的敏感性分级(α_i, k)-匿名聚类算法((α_i, k)-GCA)，算法的具体步骤如下：

算法 (α_i, k)-GCA

输入 原始数据表 T，匿名约束 k，敏感值分级($L_1, L_2,\cdots,L_\lambda$)及其频率约束($\alpha_1, \alpha_2,\cdots, \alpha_\lambda$)

输出 准标识符属性表 T_QI，敏感属性表 T_SA

定义

1. 临时表 T'：包括 QI 属性、SA 属性和 class ID；
2. 等价类 E；
3. 基于敏感性分级的敏感值分组 G(G_{s1},G_{s2},$\cdots$,G_{si},$\cdots$,G_{sn})，n 为敏感属性不同的取值个数，G_{si} 是所有敏感属性值为 S_i 的记录分组，并标记该分组所属的敏感级别 L_i 和频率约束α_i；
4. 分组优先级 P，敏感级别高的优先级高，级别相同者，记录数多的优先级高；

流程

1. 初始化 E = Φ
2. while 可以构造等价类
3. while E 不满足(α_i, k)-匿名时
4. if 存在非空分组，且非空分组数大于等于 k
5. 从优先级最高的分组中提取一条记录 t 添加到 E，从该组中删除 t；
6. Do
7. 从除了 t 所在级别的分组中提取记录，按照级别由低到高的顺序依次从 G 中分别取一条记录添加到 E，同级别中选 P 高的分组，每取出一条记录，即从该分组中删除该记录；
8. if (k>λ-2)
9. 重复第 7 步；
10. end if
11. Until 记录 t 满足(α_i,k)-匿名
12. 将 E 添加到 T'，并标记其 class ID；
13. E=Φ；
14. end if
15. end while
16. end while
17. for each 剩余的记录 rt
18. 如果存在等价类 E，添加 rt 后仍然满足(α_i,k)-匿名，则添加 rt 到该等价类，优先考虑不含相同级别的、与最高级别距离较大的等价类；

19. end for

20. 隐匿所有无法添加到等价类的记录；

21. 将所有等价类以 T_QI,T_SA 2 张表输出；

以表 1 的原始数据为例来说明该算法的执行过程。按表 4 的敏感性分级，并定义敏感级的α_i约束为{0.3, 0.35, 0.4, 0.45, 0.5}，k=3。过程如下：(1)按优先级别由高到低的分组。第 1 级别分组有：HIV={t_7, t_{11}}，Cancer={t_8, t_9}；第 3 级别分组有：Heart Disease={t_{12}}；第 4 级别分组有：Flu={t_2, t_4}，Viral Infection={t_3, t_6}；第 5 级别分组有：Fever={t_1, t_5, t_{10}}；(2)从优先级最高的组中取出一条记录 t_7，再按级别由低到高分别取出 t_1、t_2 构成第一个等价类{t_7, t_1, t_2}；(3)重复上述过程生成临时表 8；(4)进行分表发布。

表 8　基于敏感性分级的(α_i, k)-匿名临时表

No	Job	Age	Sex	Zipcode	Disease	Class ID
7	dustman	42	M	13226	HIV	1
1	clerk	23	F	13010	Fever	1
2	accountant	25	F	13050	Flu	1
8	wleder	45	M	13228	Cancer	2
5	housewife	36	F	13220	Fever	2
3	barker	30	M	13020	Viral Infection	2
9	prof.	52	F	14850	Cancer	3
10	mechanic	53	M	14862	Fever	3
4	editor	31	F	13040	Flu	3
11	manager	61	M	14802	HIV	4
6	manager	39	M	13221	Viral Infection	4
12	prof.	64	M	14800	Heart Disease	4

比较表 5 和表 8 可以看到，后者的等价类中不存在具有相同高敏感性的敏感属性值。从而可以更好地实现隐私保护。

5　实验与分析

采用实际数据集对本文提出的算法进行测试，实验数据来自于 UCI machine learning repository 的人口统计数据集，该数据集下载于 http://kdd.ics.uci.edu 网站，总记录数为 32 562，目前有关隐私保护的多数文献都采用该数据集。由于在下载的原始数据集中，有少数记录的数据不完整，首先过滤掉这部分不完整记录，通过数据格式转换后，随机提取 10 000 条记录进行实验。该数据集共有 14 种属性，本实验选择其中的 9 种作为 QI 属性，1 种作为 SA 属性，具体描述如表 9 所示。

表 9　Adult 实验数据集描述

属性	No.	Attribute	Distinct Values	Type	Generalizations	Height
QI	1	Age	74	Numeric	<20,[20,29],…,>70	7
	2	Workclass	7	Categorical	Hierarchy Tree	3
	3	Education	16	Categorical	Hierarchy Tree	4
	4	Marital-status	7	Categorical	Hierarchy Tree	3
	5	Relationship	6	Categorical	Hierarchy Tree	2
	6	Race	5	Categorical	Hierarchy Tree	2
	7	Sex	2	Categorical	Suppression	1
	8	Native-country	41	Categorical	Hierarchy Tree	3
	9	Salary-class	2	Categorical	Suppression	1
SA	1	Occupation	14	Categorical	–	–

实验运行环境为：Intel Pentium D，3.0 GHz CPU，1 GB 内存；操作系统平台为：Microsoft Windows XP Professional；编程环境为：MyEclipse 7.0+Oracle 9i。

Incognito 全域泛化算法[11]是目前著名的泛化算法，Wong 将该方法应用到(α, k)-匿名给出了适合(α, k)-匿名的 eIncognito 方法，并在此基础上作出了改进，给出了改进后的 Top-Down 方法。笔者分别实现了 Top-Down 算法、本文的(α_i,k)-GCA 算法，同时将贪心策略应用到传统(α, k)-匿名的聚类算法。从信息损失率、执行效率和敏感度距离 3 个方面比较它们的性能。对 Adult 数据集中敏感属性 Occupation 的属性值采用如表 10 所示的分级。

表 10　Occupation 属性分级

级别	α	属性值
1	0.20	Protective-serv, Armed-Forces
2	0.30	Exec-managerial, Prof-specialty, Adm-clerical
3	0.35	Tech-support, Craft-repair, Machine-op-inspct, Transport-moving
4	0.40	Priv-house-serv, Farming-fishing, Handlers-cleaners, Sales
5	0.50	Other-service

5.1　信息损失率比较

一个成熟的发布方法既需要考虑数据的隐私，也要考虑数据之间的相关性。其具体的应用则是反映在查询结果的准确性上。而查询结果是否准确，又取决于数据重构的完整性。任何基于隐私保护的数据发布方法都会有不同程度的信息损失，对于发布后的重构数据不可能，也不应恢复到原始数据，否则也就不可能保证隐私。本文采用文献[10]中关于数据重构距离的方法来比较两者之间的信息损失率。在满足相同的(α, k)-匿名的基础上，重构损失率低的方法更有效。

用一个 d+1 维空间 DS 来表示整个数据集 T，数据集中的每一条记录都可以被映射成该空间的一个点。该空间包括 d 维的 QI 属性和一个 SA，记录 t 可表示为(QI[1], QI[2], …, QI[d], t[d+1])。用概率密度函数 P_t 来描述 t 这条记录可以发生重构的概率。在多维空间 DS 中，记录 t 的概率密度函数可定义为：

$$P_t(x)=\begin{cases}1 & \text{if } x=t\\ 0 & \text{otherwise}\end{cases} \tag{3}$$

其中，x 是 d+1 维空间 DS 中的随机变量；条件 x=t 表示 $x[i]$=$t[i]$。对于所有的 $i\in[1, d+1]$，这里的 $t[i]$表示记录 t 的第 i 个字段取值。对于泛化表的概率密度函数如下：

$$\beta_t(x)=\begin{cases}\dfrac{1}{\prod_{i=1}^{d}L(QI[i])} & \text{if } x[i]\in QI[i],\forall i\in[1,d]\\ 0 & \text{otherwise}\end{cases} \tag{4}$$

其中，$L(QI[i])$表示记录 t 的 $QI[i]$字段的泛化长度。如表 2 所示，记录 t_1 的 Age 字段长度即为 20。

对于基于有损连接的方法：设 v_1,v_2,…,v_n 是临时表中某一等价类的 n 个不同取值(如表 4 所示，classID1 等价类的 n=3)，$c(v_k)(1\leqslant k\leqslant n)$为等价类中 SA=$v_k$ 的记录数；$|\mathrm{QI}|_{\mathrm{classID}}$ 为等价类中的所有记录数。则基于有损连接方法的概率密度函数可推广为：

$$\beta_t(x)=\begin{cases}c(v_1)/|\mathrm{QI}|_{\mathrm{classID}} & \text{if } x=(t[1],t[2],\ldots,t[d],v_1)\\ \quad\mathrm{M}\\ c(v_n)/|\mathrm{QI}|_{\mathrm{classID}} & \text{if } x=(t[1],t[2],\ldots,t[d],v_n)\\ 0 & \text{otherwise}\end{cases} \tag{5}$$

对于基于概率的密度函数，每条记录都可以计算它的重构距离，本文采用表中所有记录的平均重构距离来具体描述信息损失率。数据集 T 的平均重构距离为：

$$Dis(\mathrm{T})=\sum_{i=1}^{N}\sum_{x\in\mathrm{DS}}\left(\beta_{t_i}(x)-P_{t_i}(x)\right)^2\Big/N \tag{6}$$

实验分别从 2 个方面计算了 Top-Down 和 GCA 方法的平均重构距离：

(1)α和|QI|不变，k 变化时重构距离比较如图 1 所示。

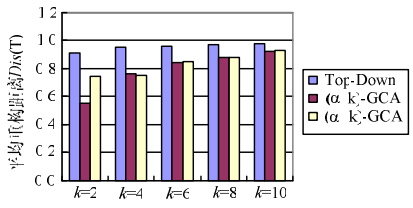


图 1　k 变化时平均重构距离比较

对于 Top-Down 和(α, k)-GCA 方法，本文选取了α=0.5，|QI|=6，分别计算了的平均重构距离，对于(α_i, k)-GCA 方法，选取了同样的 6 种 QI 属性，α的约束如表 10 所示。由图 1 可以看到，当α和|QI|不变时，随着的 k 的增大，3 种方法的重构距离都在增加，说明 k 的要求越高前者的泛化程度越高，后两者的等价类也越大，信息损失都会变大。但总的来讲，Top-Down 方法的信息损失率始终要高于本文的 GCA 方法。而本文的 2 种方法都是采用的有损连接的发布思想，仅在分组策略上根据各自的模型算法有所不同，因此，其分组的重构距离基本一样。只有当 k=2 时，(α_i, k)-GCA 方法的距离要明显大于(α, k)-GCA 的距离，这是因为后者的α=0.5，相对于前者α_i的约束要求要低一些，而此时 k 比较小，因此保护发布信息的约束完全取决于α_i。当 k 较大时，其约束就由α和 k 同时决定。

(2)k 和|QI|不变，α变化时重构距离比较如图 2 所示。

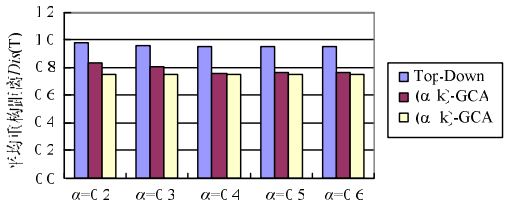


图 2 α变化时平均重构距离比较

本文选择 k=4，|QI|=6，分别计算随α变化时，前 2 种方法的重构距离，(α_i, k)-GCA 方法与此处的α没有关系，在图 2 中其距离是同一个数据，放在图中是为了体现方法之间的比较，其α_i同样由表 10 所定义。根据实验结果可知，当α取 0.2 或 0.3 时，条件比较严格，此时会有较多的记录被隐匿，且 2 种方法的重构距离都比较大，当α大于 0.5 之后，α的变化已经没有意义，此时等价类的构造实际由 k=4 来限定，因此重构距离不再变化。但是不管α如何变化，本文 GCA 方法的平均重构距离始终小于 Top-Down 方法的平均重构距离。

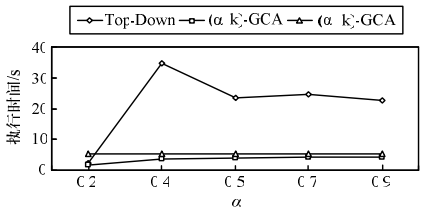
通过上述 2 组实验的比较可以看到，本文提出的基于贪心策略的聚类算法与基于泛化的 Top-Down 方法相比，具有较低的信息损失率。

5.2 执行效率比较

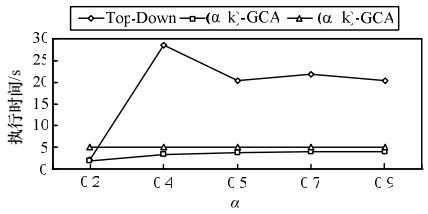
为了反映算法的实际应用性，同样从 2 个方面进行了实验，比较算法的执行效率。

(1)k 和|QI|不变，随α变化的算法执行时间

实验分别选取了 k=4，|QI|=6 和 k=10，|QI|=6 这 2 组数据来比较执行时间。结果如图 3 所示。当α为 0.2 时，执行时间很短，这是因为α取值很小时，很难得到满足，大量的记录被隐匿了，而 Top-Down 方法是首先从最高层“*”开始逐步具体化，当α很小时，等价类中所含记录数较多，泛化程度高，所以执行时间短；同样本文算法此时也需要隐匿大量的记录，所构成的等价类个数较少，因此执行时间也短，而且本文算法只需要对 SA 一个属性进行操作，速度比 Top-Down 方法快。随着α的增大，Top-Down 对 QI 的具体化程度将会增加，而且需要多个 QI 的泛化组合，因此执行时间大幅增大，当α大到一定程度时，α的约束已经失去意义，此时完全受 k 的约束，因此执行时间基本不变。对于(α_i, k)-GCA 方法，由于其α_i约束是一个固定值，因此其执行时间不变。该方法需要考虑敏感性的分级排序，所以，执行时间上略高于(α, k)-GCA 算法。



(a)k=4, |QI|=6

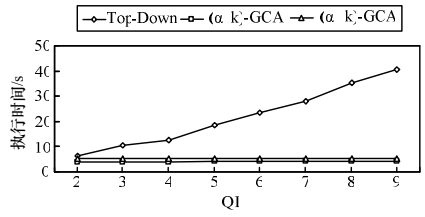


(b)k=10, |QI|=6

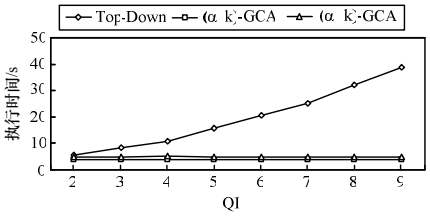
图 3 α变化时执行时间比较

(2)α和 k 不变，随|QI|变化的执行时间

由图 4 可以看到，当α和 k 不变时，随着|QI|的变化，本文的算法执行时间基本没有变化，因为不管|QI|怎么变化，对应的都是一条记录，对本文算法没有影响；而|QI|越多，Top-Down 要处理的泛化组合越复杂，时间复杂度越高。对于(α, k)-GCA 方法，由于α=0.5，实际上两组实验在约束上都取决于 k。而(α_i, k)-GCA 方法，当 k=4 时，α_i和 k 都会起到约束作用，当 k=10 时，其约束仅由 k 决定。因此，图 4(a)中的执行时间略高于图 4(b)。



(a)α=0.5, k=4



(b)α=0.5, k=10

图 4 |QI|变化时执行时间比较

上述 2 个方面的实验结果都表明，本文算法 GCA 在执行效率上要明显优于基于泛化的 Top-Down 方法。

5.3 隐私保护度比较

本文通过高敏感性等价类距离度量来比较算法的隐私保护度。高敏感性等价类距离指在最终发布的敏感属性表 T_SA 中，所有含有较高级别敏感属性的平均等价类敏感性距离(D_s(T_SA))，实验选择含有 L_1 和 L_2 2 个级别的等价类，并比较它们的等价类敏感性距离。等价类距离越大，则等价类中敏感值的差异越大，对于高敏感属性受攻击的难度就越大，则数据发布的安全性就越好。

根据上文关于等价类敏感性距离度量的定义，D_s(T_SA)的计算公式为：

$$D_s(\text{T_SA}) = \sum_{i=L_1,L_2} D(E_i) / |E_{L_1,L_2}| \quad (7)$$

其中，E_{L_1,L_2}为含有敏感级别为 L_1 或 L_2 的等价类；$|E_{L_1,L_2}|$为

满足条件的等价类个数。

对于一般(α, k)-匿名模型，结合(α_i, k)-匿名模型的敏感性分级，选取|QI|=6，k=5，分别计算α=0.2、α=0.3、α=0.35、α=0.4、α=0.5 情况下的高敏感性等价类距离度量，实验结果如图 5 所示。

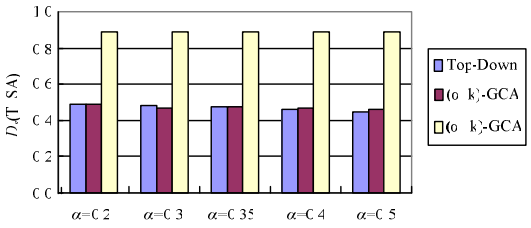


图 5　高敏感性等价类距离度量

实验结果表明，(α_i, k)-匿名模型的平均等价类距离要明显大于一般(α, k)-匿名模型的平均等价类距离。由于后者在构造等价类时，仅从 QI 的角度实现聚类分组，并没有考虑到记录敏感属性值之间的关系，因此在等价类中的敏感属性值虽然不同，但往往属于同样的高敏感级，或相近的高敏感级。而(α_i, k)-匿名模型在实现聚类分组时，充分考虑了敏感属性值之间的敏感性差异，在满足α_i和 k 约束的前提下，总是尽可能地让敏感属性值级别差异较大的记录分在同一组。因此，(α_i, k)-匿名模型的隐私保护度较高，其高敏感值抗攻击的能力要强，该模型具有更高的发布安全性。

6　结束语

本文针对(α, k)-匿名模型的缺陷，提出了基于敏感性分级的(α_i, k)-匿名模型。通过定义敏感性距离度量的方法来描述它们的敏感性差异，引入有损连接的思想通过损失部分准标识符属性和敏感属性间的关联来发布隐私数据，并给出基于贪心策略的(α_i, k)-匿名聚类算法。从信息损失度、执行性能和隐私保护度 3 个方面进行了实验比较和分析，结果表明，本文的模型能充分反映敏感值之间的敏感性差异，对于高敏感性的敏感值具有更好的隐私保护效果，同时所提出的算法具有更好的执行效率和发布安全性。

参考文献

[1] 周水庚, 李　丰, 陶宇飞, 等. 面向数据库应用的隐私保护研究综述[J]. 计算机学报, 2009, 32(5): 847-860.

[2] 黄春梅, 费耀平, 李　敏, 等. 基于多维泛化路径的 k-匿名算法[J]. 计算机工程, 2009, 35(2): 154-156.

[3] Fung B C M, Wang Ke, Chen Rui, et al. Privacy-preserving Data Publishing: A Survey on Recent Developments[J]. ACM Computing Surveys, 2010, 42(4): 1-55.

[4] Meyerson A, Williams R. On the Complexity of Optimal k-anonymity[C]//Proc. of the 23rd ACM Symposium on Principles of Database Systems. Paris, France: ACM Press, 2004: 223-228.

[5] Machanavajjhala A, Gehrke J, Kifer D, et al. L-diversity: Privacy Beyond k-anonymity[C]//Proc. of the 22nd International Conference on Data Engineering. Atlanta, Georgia, USA: IEEE Press, 2006: 24-36.

[6] Wong R C, Li Jiuyong, Fu A W, et al. (α, k)-anonymous Data Publishing[J]. Journal of Intelligent Information Systems, 2009, 33(2): 209-234.

[7] Li Ninghui, Li Tiancheng, Venkatasubramanian S. T-closeness: Privacy Beyond k-anonymity and L-diversity[C]//Proc. of the 23rd International Conference on Data Engineering. Istanbul, Turkey: IEEE Press, 2007: 44-56.

[8] 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.

[9] Han Jianmin, Yu Huiqun, Yu Juan. An Improved L-diversity Model for Numerical Sensitive Attributes[C]//Proc. of the 3rd International Conference on Communications and Networking in China. Hangzhou, China: [s. n.], 2008: 938-943.

[10] Xiao Xiaokui, Tao Yufei. Anatomy: Simple and Effective Privacy Preservation[C]//Proc. of the 32nd International Conference on Very Large Data Bases. Seoul, Korea: [s. n.], 2006: 139-150.

[11] LeFevre K, DeWitt D J, Ramakrishnan R. Incognito: Efficient Full-domain k-anonymity[C]//Proc. of SIGMOD'05. Baltimore, Maryland, USA: ACM Press, 2005: 49-60.

编辑　金胡考

(上接第 11 页)

5　结束语

本文利用宏块间的运动相关性和宏块间距离的关系，提出了一种利用竖直方向上的丢失片边缘正确宏块的运动信息加权预测出错误宏块运动矢量的掩盖算法，算法复杂度远低于 BMA 算法。实验结果说明，使用本文算法后主观质量和客观质量都能得到较大的提升。该方法可广泛使用在各种高清视频应用中，如视频会议、视频监控、数字电视等。

参考文献

[1] 盛　赞, 张有志, 张丽君. 一种改进的H.264自适应空域错误掩盖算法[J]. 计算机工程, 2007, 33(21): 213-215.

[2] 朱小辉. 基于 H.264 的错误掩盖技术研究[D]. 西安: 西安电子科技大学, 2008: 25-27.

[3] Kale A, Hampiholi V. Real Time Error Concealment in H.264 Video Decoder for Embedded Devices[C]//Proc. of the 2nd International Conference on Internet Multimedia Services Architecture and Applications. Bangalore, India: [s. n.], 2008.

[4] Zhao Qingguo, Li Chongrong. Effective Error Concealment for Real-time High Definition Video Delivered over IP Network[C]//Proc. of the 2nd International Conference on Communications and Networking. [S. l.]: IEEE Press, 2007.

[5] Liu Xingang, Sohn K. A Pervasive Temporal Error Concealment Algorithm for H.264/AVC Decoder[C]//Proc. of the 10th IEEE International Conference on Computer and Information Technology. [S. l.]: IEEE Computer Society, 2010.

[6] Texas Instruments. Digital Media System on Chip Data Manual[Z]. 2007.

编辑　顾逸斐