个性化(α,l)-多样性 k-匿名隐私保护模型

曹敏姿[1] 张琳琳[1] 毕雪华[2] 赵 楷[1]
(新疆大学信息科学与工程学院 乌鲁木齐 830046)[1]
(新疆医科大学医学工程技术学院 乌鲁木齐 830011)[2]

摘 要 针对传统隐私保护模型对个性化匿名缺乏考虑的问题,对现有的两种个性化匿名机制进行了分析。在 k-匿名和 l-多样性匿名模型的基础上,提出一种个性化(α,l)-多样性 k-匿名模型来解决存在的问题。在该模型中,依据敏感程度的不同,对敏感属性的取值划分类别;设置相应的约束条件,并为特定的个体提供个性化的隐私保护。实验结果表明,所提模型在有效提供个性化服务的同时,具有更强的隐私保护能力。

关键词 隐私保护,k-匿名,l-多样性,个性化匿名,泛化
中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.11.028

Personalized (α,l)-diversity k-anonymity Model for Privacy Preservation

CAO Min-zi[1] ZHANG Lin-lin[1] BI Xue-hua[2] ZHAO Kai[1]
(College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)[1]
(Department of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830011, China)[2]

Abstract Aiming at the problem that traditional privacy preservation model is lack of considering the personalized anonymity, this paper analyzed the existing two personalized anonymity mechanisms. On the basis of k-anonymity and l-diversity model, a personalized (α, l)-diversity k-anonymity model was proposed to solve the existing problems. In the proposed model, the sensitive attribute values are divided into several categories according to their sensitivities, each category is assigned with corresponding constraints, and the personalized privacy preservation is provided for specific individuals. The experimental results show that the proposed model can provide stronger privacy preservation while supplying personalized service efficiently.

Keywords Privacy preservation, k-anonymity, l-diversity, Personalized anonymity, Generalization

1 引言

在当今信息化时代,每天都有大量的个人数据被收集和发布,例如人口普查数据、医疗数据、个人消费数据等。对这些已发布数据进行挖掘和分析,有利于科学研究的发展、商业决策的制定等,给人们的工作和生活带来许多便利;但与此同时,所发布的数据中包含了大量与个人相关的敏感信息,这也会带来相应的隐私问题。"隐私"意指可确认特定当事人的身份或特征,但当事人不愿意暴露的敏感信息[1]。数据发布中的隐私保护是针对公开发布的数据采取一定的隐私保护措施,在保证数据可用性的同时,防止他人通过知识推断、链接攻击、数据挖掘等手段获取目标对象的敏感数据[2]。

在数据发布领域,为了在保证数据可用性的同时保护个人的隐私,研究者们提出了多种隐私保护模型,例如 k-匿名[3]、l-多样性[4]以及 t-closeness[5]等。这些模型均是建立在数据匿名化的基础之上的,其基本思想是在确保所发布的信息公开可用的前提下,隐藏公开数据记录与特定个人之间的对应关系,从而保护个人隐私。然而,这些传统的模型对所有的个体都提供相同程度的隐私保护,并没有考虑个性化的隐私保护需求[6]。

在现实生活中,不同隐私信息的敏感程度不同,相同隐私信息对不同个体的重要性程度也不相同,因此有必要提供个性化的隐私保护。Xiao 等[7]首次提出个性化匿名的概念。此后,许多相似的个性化匿名解决方法也相继被提出[8-13]。一般说来,个性化的匿名原则包含两种机制,分别是面向个人的以及面向敏感值的[14]。其中,文献[9]和文献[10]采用了面向个人的个性化匿名方法,为数据集中的每一条记录设计了不同的隐私保护需求,但这种方法不仅工作量大,而且会造

到稿日期:2017-10-03 返修日期:2018-02-12 本文受国家自然科学基金(61562088),新疆维吾尔自治区科技厅项目(2017D01C232),新疆维吾尔自治区高校科研计划项目创新团队(XJEDU2017T002),新疆维吾尔自治区高校计划项目(XJEDU2017M005),赛尔网络下一代互联网技术创新项目(NGII20170325)资助。

曹敏姿(1992-),女,硕士生,主要研究方向为隐私保护;张琳琳(1974-),女,博士,副教授,主要研究方向为数据可视化、隐私保护,E-mail:zllnadasha@126.com(通信作者);毕雪华(1982-),女,硕士,副教授,主要研究方向为数据挖掘、隐私保护医学信息学;赵 楷(1976-),男,博士,副教授,主要研究方向为大数据与安全、服务计算。

成过多的数据冗余。文献[11]和文献[12]探讨了面向敏感值的个性化匿名。其中，Han 等[11]考虑了不同敏感属性值的敏感程度，为等价组中的每一个敏感属性值设置了不同的频率约束；Shen 等[12]根据个人指定的敏感属性值的隐私保护程度来评估约束条件。一般说来，面向敏感值的匿名机制忽视了个人对隐私保护的需求，然而，如果隐私保护的程度仅仅由个人的喜好决定，便会造成隐私信息的泄露或者过多的信息损失。其中，隐私泄露的主要原因之一就是忽视了敏感属性值的全局分布约束；而过多的信息损失则是由于个人过度的隐私保护要求造成的。

本文主要研究数据发布中的个性化隐私保护问题。在 k-匿名以及 l-多样性模型的基础上，提出一种个性化的(α，l)-多样性 k-匿名模型。该模型有效地集成了个性化匿名的两种机制，不仅对敏感属性值设置了相应的约束条件，而且可以满足个人对隐私保护的需求。

2　相关知识

通常情况下，数据拥有者会将待发布的数据记录以数据表的形式发布出去，表中的每一条记录代表一个个体。数据表的属性根据其与个体的关系可分为 4 类[3]：1)标识符属性(I)，指能够唯一标识到个人的属性，例如身份证号码、姓名等；2)准标识符属性(QI)，指可能同时存在于发布的数据表及外部数据表中，通过链接可以重新标识个人身份的属性，例如属性组{Gender，Age，Zip code}；3)敏感属性(S)，即包含个人敏感信息的属性，例如疾病、薪水等；4)非敏感属性(N)，即其他属性。

表 1 为某原始数据。其中，“Name”为标识符属性；“Gender”“Age”“Zip code”均为准标识属性；“Disease”为敏感属性。一般地，在对表数据集进行匿名化操作时，标识符将被移除，且对非敏感属性不做处理。

表 1　原始数据

Table 1　Original data

Name	Gender	Age	Zip code	Disease
Bob	Male	34	100751	Cancer
David	Male	43	100720	Flu
Lily	Female	25	200386	HIV
Rose	Female	28	200425	HIV
Emma	Female	55	178642	Cancer
Jose	Female	48	178653	Flu
Cherry	Female	59	178634	HIV

2.1　泛化

泛化(generalization)是最常见的数据匿名化方法之一，是将原始数据属性值以一个更一般形式的值替换的过程[3]。例如，将属性“Age”的原始值 30 和 34，使用区间[30，35]替换；将属性“Gender”的原始值男、女，用“个人”替换。泛化之后的值取决于原始属性的特征，一般通过为属性构造泛化层次树来规定属性泛化的规则。

2.2　k-匿名模型

定义 1(等价组)　给定表数据集 D，$D=\{r_1,r_2,\cdots,r_n\}$，其中 r_i 为 D 中的第 i 条记录。若 $A=\{A_1,A_2,\cdots,A_j\}$ 为 D 中准标识符属性集合，j 为 D 中准标识符属性的个数，则每一个在 A 上取值相同的记录集合为一个等价组。

定义 2(k-匿名)　给定表数据集 D 和等价组 M，若 D 中的任意 M 的大小都至少为 k，即 D 中任意一条记录都至少与其他 $k-1$ 条记录在准标识符上不可区分，则称 D 满足 k-匿名。

2-匿名表如表 2 所列，其为表 1 进行 2-匿名化之后的表。将表 1 中的标识符属性“Name”移除，以避免个人身份被泄露；对准标识符属性组{Gender，Age，Zip code}取值进行泛化操作，保留敏感属性“Disease”的原始取值。表 2 中有 3 个等价组，每个等价组中至少包含 2 条记录。

表 2　2-匿名表

Table 2　2-anonymity table

Group-ID	Gender	Age	Zip code	Disease
1	Male	[31，45]	100***	Cancer
	Male	[31，45]	100***	Flu
2	Female	[16，30]	200***	HIV
	Female	[16，30]	200***	HIV
3	Female	[46，60]	1786**	Cancer
	Female	[46，60]	1786**	Flu
	Female	[46，60]	1786**	HIV

k-匿名模型使得攻击者最多只能以 $1/k$ 的概率通过准标识属性关联出目标个体的身份，它可以有效地抵御身份泄露，却不能很好地抵御属性泄露[15]。例如，表 2 中的第 2 个等价组中的两条记录的敏感值都是 HIV，若攻击者事先知道 Lily 的性别、年龄和邮政编码，便可推断出 Lily 在匿名表的第 2 个等价组中，从而可确定其一定患有 HIV。

2.3　l-多样性模型

l-多样性可以解决 k-匿名模型的不足，主要针对属性泄露对数据集进行保护。

定义 3(l-多样性)　给定数据集 D 和等价组 M，若 D 中的任意 M 的不同敏感属性值的个数至少为 l，则称 D 满足 l-多样性。

例如，某 2-多样性匿名表如表 3 所列。表 3 中有 3 个等价组，每个等价组中至少包含 2 个不同的敏感属性值。l-多样性模型使得攻击者推断出目标个体敏感信息的概率至多为 $1/l$。

表 3　2-多样性匿名表

Table 3　2-diversity anonymity table

Group-ID	Gender	Age	Zip code	Disease
1	Male	[31，45]	100***	Cancer
	Male	[31，45]	100***	HIV
2	Female	[61，75]	200***	HIV
	Female	[61，75]	200***	Flu
3	Female	[46，60]	1786**	Cancer
	Female	[46，60]	1786**	Flu
	Female	[46，60]	1786**	HIV

l-多样性模型虽然考虑了等价组中敏感属性值的多样化，但是它会遭受偏斜攻击和相似性攻击[15]，攻击者仍然可以推测出目标个体敏感信息的敏感程度或者取值范围。例如，表 3 中的第 1 个等价组中的敏感属性满足 2-多样性约束，但与表中的“Flu”相比，敏感属性值“Cancer”和“HIV”的敏感程度更高。假设攻击者的目标对象在第 1 个等价组中，其虽然不能确定目标个体的敏感值是“Cancer”还是“HIV”，但可以知道目标用户患有很严重的疾病。

3 个性化(α,l)-多样性 k-匿名模型

为了解决上述问题,并且有效地集成个性化匿名的两种机制,本文在 k-匿名和 l-多样性模型的基础上,提出一种个性化的(α,l)-多样性 k-匿名模型。在数据集满足 k 和 l 约束的条件下,限制敏感级别相同的敏感值在同一个等价组中出现的频率,并为特定的个体提供个性化的隐私保护。

3.1 敏感属性值的敏感级别

本文与文献[16-17]相同,在不失一般性的条件下,将敏感属性 S 的不同取值按照敏感程度划分类别,同一敏感级别的敏感值属于同一个类别,$D(Sid)$为敏感属性 S 的敏感级别的值域。例如,表 4 为敏感属性"Disease"的 8 个取值按照敏感程度划分类别后的表。其中,Cid 表示类别;Sid 表示敏感级别,Sid 的值越大表示敏感级别越高。此时,"Disease"属性所对应的 $D(Sid)=\{1,2,3,4\}$。

表 4 "Disease"属性值的分类[16]

Table 4 Categories of "Disease"[16]

Cid	Sensitive attribute values	*Sid*
1	HIV,Cancer	4 (Top Secret)
2	Phthisis,Hepatitis	3 (Secret)
3	Heart Disease,Asthma	2 (Less Secret)
4	Flu,Indigestion	1 (Non secret)

3.2 (α,l)-多样性 k-匿名模型

定义 4(α-非相关约束) 给定数据集 D、等价组 M、敏感属性 S 以及指定的频率约束 $\alpha(0\leqslant\alpha\leqslant1)$,若对于 D 中的任意 M,和 S 的任意 $Sid\in D(Sid)$,都有 $|(M,Sid)|/|M|\leqslant\alpha$,则称数据集 D 满足 α 约束。其中,$|(M,Sid)|$表示等价组 M 中敏感值的敏感级别为 Sid 的记录条数,$|M|$为等价组的大小。需要强调的是,本文中 α 分布约束的定义与文献[17]中的略有不同,α 可由数据发布者指定,也可以由用户指定。

如表 5 所列,其中任意一个等价组中的敏感属性取值都满足 $|(M,Sid)|/|M|\leqslant0.5$。

定义 5((α,l)-多样性 k-匿名模型) 给定表数据集 D 和等价组 M,如果 M 中至少包含 k 条记录,且这些记录的不同敏感属性值的个数至少为 $l(l\leqslant k)$,同时敏感属性值的分布满足 α 约束,则称 D 为满足(α,l)-多样性 k-匿名模型的表。

例如,表 5 为满足(0.5,2)-多样性 2-匿名的数据表。表 5 中有 3 个等价组,每一个等价组中至少包含 2 条记录,且每个等价组中不同的敏感属性值的个数至少为 2,同时敏感属性值的分布满足 $\alpha=0.5$ 约束。在数据集满足 k-匿名和 l-多样性的基础上,参数 α 的设置可以在一定程度上降低敏感属性值泄露的概率。

表 5 (0.5,2)-多样性 2-匿名表

Table 5 (0.5,2)-diversity 2-anonymity table

Group-ID	Gender	Age	Zip code	Disease	*Sid*
1	Male	[31,45]	100***	Cancer	4
	Male	[31,45]	100***	Flu	1
2	Female	[61,75]	200***	HIV	4
	Female	[61,75]	200***	Asthma	2
3	Female	[46,60]	1786**	Cancer	4
	Female	[46,60]	1786**	Flu	1
	Female	[46,60]	1786**	HIV	3

3.3 个性化(α,l)-多样性 k-匿名模型

依据敏感程度的不同为敏感属性值划分的类别,可能无法充分地满足具体个人对隐私保护的需求。例如,疾病属性值"Flu"可能对大多数人而言并不敏感,然而,对在幼儿园工作的老师 Alice 来说,她不想让学生家长知道自己患有"Flu"。此时,Alice 的敏感属性值"Flu"需要更强的隐私保护。为此,本文通过为敏感属性构造泛化树的方法,允许特定个人为自己的敏感属性值设置隐私保护级别,以在数据集满足(α,l)-多样性 k-匿名模型的基础上进一步提供个性化的隐私保护。

3.3.1 敏感属性泛化树

为敏感属性 S 构建一棵泛化层次树 GHT(Generalization Hierarchy Tree),树的层次级别数与 Sid 的取值相同,其中,属性的原始取值作为树的叶子节点,对应的树层次级别为 1。例如,表 4 中"Disease"属性的 8 个取值构造的泛化层次树如图 1 所示,自底向上,每一层节点所在的层次级别分别为 1,2,3,4。

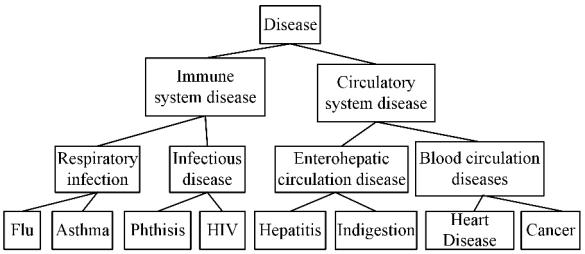

图 1 "Disease"属性的泛化层次树

Fig. 1 Generalization hierarchy tree for "Disease"

3.3.2 隐私保护级别

依据敏感属性的 GHT,允许个人指定自己敏感属性值的隐私保护级别 Ppl(Privacy preservation level),$Ppl=\{1,2,\cdots,n\}$,n 为敏感属性层次树的根节点所在的级别。表 6 为个人指定的隐私保护级别表,其中"—"表示个人没有为自己的敏感属性值设置隐私保护级别。

表 6 个人指定的隐私保护级别

Table 6 Individual-specified privacy preservation levels

No.	Gender	Age	Zip code	Disease	*Sid*	*Ppl*
1	Male	34	100751	Cancer	4	4
2	Male	43	100720	Flu	1	2
3	Female	66	200386	HIV	4	3
4	Female	70	200425	Asthma	2	—
5	Female	55	178642	Cancer	4	—
6	Female	48	178653	Flu	1	1
7	Female	36	178634	Hepatitis	3	2

3.3.3 个性化隐私保护规则

在数据表满足(α,l)-多样性 k-匿名模型的基础上,对敏感值的 $Ppl\leqslant Sid$ 的记录不做处理;对于 $Ppl>Sid$ 的记录,则用 GHT 中与 Ppl 取值相同的级别所在的原始值的父节点代替当前的敏感值。

例如,在对表 6 进行相应的匿名化操作之后,发现第 2 条记录中,敏感属性值"Flu"的 $Ppl=2$ 大于其对应的敏感级别 $Sid=1$,则用 GHT 中对应级别为 2 的"Flu"的父节点"respiratory infection"代替"Flu",作为匿名化之后的敏感属性取值。

定义 6(个性化的(α,l)-多样性 k-匿名模型) 给定表数

据集 D，若 D 满足(α,l)-多样性 k-匿名模型，且 D 中的各元组的敏感属性值满足个性化隐私保护的规则，则称 D 为满足个性化(α,l)-多样性 k-匿名模型的表。

例如，表 7 为表 6 进行个性化$(0.5,2)$-多样性 2-匿名之后的表。

表 7　个性化$(0.5,2)$-多样性 2-匿名表

Table 7　Personalized $(0.5,2)$-diversity 2-anonymity table

Group-ID	Gender	Age	Zip code	Disease
1	Male	[31,45]	100 ***	Cancer
	Male	[31,45]	100 ***	Respiratory infection
2	Female	[61,75]	200 ***	HIV
	Female	[61,75]	200 ***	Asthma
3	Female	[46,60]	1786 **	Cancer
	Female	[46,60]	1786 **	Flu
	Female	[46,60]	1786 **	Hepatitis

3.4　信息损失度量与敏感属性值的识别率

通常采用数据可用性以及隐私保护程度来评价匿名数据集的质量。其中，匿名数据集的可用性可以通过信息损失来衡量，而数据集中敏感值的平均识别率可以用来衡量隐私保护的效果。

3.4.1　信息损失度量

无论是准标识符属性值还是敏感属性值的泛化，都会造成相应的信息损失。本文使用文献[18]中的方法，采用标准化确定性惩罚(NCP)来衡量数据的可用性。

定义 7(数值型属性信息损失度量)　设 $D=(A_1,A_2,\cdots,A_n)$是待匿名化的表数据集，其中$(A_1,A_2,\cdots,A_n)$是表 D 中的属性，且 A_i 为数值型属性。假设 D 中记录 $t=(x_1,x_2,\cdots,x_i,\cdots,x_n)$在 A_i 被泛化为$[y_i,z_i]$，其中 $y_i\leqslant x_i\leqslant z_i$，那么记录 t 在数值属性 A_i 上的标准化确定性惩罚定义为：

$$NCP_{A_i}(t)=\frac{z_i-y_i}{|A_i|} \tag{1}$$

其中，$|A_i|$是属性 A_i 的值域范围。

例如，属性年龄的值域范围为$[0,90]$，若某条记录的年龄属性取值为 35，其泛化到区间$[30,45]$，则该记录的年龄属性值泛化后的信息损失为：$(45-30)/90=1/6$。

定义 8(分类型属性信息损失度量)　设元组 t 在分类属性 A_C 上的属性值 v 和其他一些属性值 $v_1,v_2,\cdots,v_i$ 一起被泛化为 u，那么元组 t 在属性 A_C 上规范化后的确定性惩罚定义为：

$$NCP_{A_C}(t)=\frac{Size(u)}{|A_c|} \tag{2}$$

其中，$|A_C|$表示分类属性 A_C 所有属性值的个数，$Size(u)$为 u 的子孙叶节点的个数。

例如，图 1 中的树有 8 个叶子节点，则$|A_C|=8$，若将分类属性值“Flu”泛化为“respiratory infection”，而“respiratory infection”的子孙叶子节点的个数为 2，则该泛化造成的信息损失为 1/4。

因此，表 T 中所有元组在各个属性上的标准化确定性惩罚为：

$$NCP(T)=\sum_{t\in T}\sum_{i=1}^{n}NCP_{A_i}(t) \tag{3}$$

其中，$NCP_{A_i}(t)$的计算需要区分 A_i 是数值型属性还是分类属性。

3.4.2　敏感属性值的识别率

对于匿名化的数据集而言，其敏感属性值被攻击者识别的概率越低，则该数据集的隐私保护效果越好。本文采用文献[19]中的方法，通过敏感属性值的平均识别率来衡量数据集的隐私保护程度。

定义 9(一条记录的敏感属性值的识别率)　给定一个数据集 D 和等价组 M，s 是等价组 M 中的某条记录 t 的敏感属性的取值，则 M 中 t 的敏感属性值 s 的识别率可以通过下式进行计算：

$$RR_t(s,M)=\frac{|(s,M)|}{|M|\times|f(s)|} \tag{4}$$

其中，$|(s,M)|$为等价组 M 中敏感属性值 s 的个数，$|M|$为等价组的大小，$|f(s)|$的值等于敏感属性值依据 GHT 泛化之后的父节点所在的子树的叶子节点的个数，若敏感值没有进行泛化，则$|f(s)|=1$。例如，表 7 中的等价组 1 中两条记录的敏感值的识别率分别为 $1/(2\times1)=1/2$ 和 $1/(2\times2)=1/4$。

定义 10(一个等价组中敏感值的平均识别率)　给定一个等价组 M，s 是等价组 M 中的某条记录 t 的敏感属性的取值，用 ARR_M 表示 M 中敏感属性值的平均识别率，则 ARR_M 可以通过下式计算：

$$ARR_M=\frac{\sum_{t\in M}RR_t(s,M)}{|M|} \tag{5}$$

定义 11(一个数据集中敏感值的平均识别率)　给定一个数据集 D 和等价组 M，D 中 M 的个数为 n，即数据集中共有 n 个等价组，用 ARR_D 表示 D 中敏感属性值的平均识别率，则 ARR_D 可以通过式(6)进行计算：

$$ARR_D=\frac{\sum_{M\subset D}ARR_M}{n} \tag{6}$$

数据集中敏感值的平均识别率越低，则隐私保护的程度越高。

4　个性化(α,l)-多样性 k-匿名模型的算法实现

个性化(α,l)-多样性 k-匿名模型算法的基本思想是：首先采用文献[18]中提出的基于聚类的自顶向下的分治策略，在数据集各聚簇满足 k 约束的条件下，对各簇中对应的敏感属性值进行参数 l 和 α 的判定，并对不能同时满足 k,l,α 约束的聚类进行相应处理；通过泛化操作，使得数据集满足(α,l)-多样性 k-匿名模型；最后依据个性化隐私保护的规则，生成满足个性化(α,l)-多样性 k-匿名模型的表 D^*。个性化(α,l)-多样性 k-匿名模型的算法如算法 1 所示。

算法 1　个性化(α,l)-多样性 k-匿名模型算法

输入：待发布的数据集 D；匿名参数 k，l，α；敏感属性 S 的敏感级别分类表 C；GHT；部分用户指定的 Ppl

输出：满足约束条件的匿名数据集 D*

1. Initialization：create M for D；
2. If $|M|<2k$ then return；
3. Else $\{\forall r_i\in D$，find r_{maxi}，according to r_i and r_{maxi}，partition M into two exclusive sub-groups M_1 and M_2；such that M_1 and M_2 are more local than C and either M_1 and M_2 has at least k tuples；/* 从 D 中任取一条记录 r_i，找到与其距离最远的记录 r_{maxi}，分别作为两个簇的初始元组并划分聚簇 */
4. If $|M_1|>2k-1$ then recursively partition M_1；

5. If $|M_2|>2k-1$ then recursively partition M_2;}

6. Merge those M in D that $|D(s_i)|<l$ in M or $|(M,Sid)|/|M|>\alpha$ into a new Group G;

/*对各簇中参数 l 以及 α 进行判定,合并不满足条件的簇*/

7. If $G\neq\emptyset$ and $|G|\geqslant 2k$ then D←G and execute lines 1－7; until D can't generate any M that satisfy k,l,α at the same time;

8. For $\forall r_i\in G$, insert r_i into nearest M in D that new M satisfy k,l,α at the same time;

9. Generalize records within the same M; get D′;

/*D′为满足(α,l)-多样性 k-匿名模型的表*/

10. For each record t in D′ that Ppl>Sid, find $note_t$ in GHT, where $l(note_t)=Ppl$ && $note_t$ is the parent node of t;

11. t←$note_t$;

12. Return D^*.

步骤 1－步骤 5 主要依据 NCP 最小化的原则,将数据集划分成大小满足 k 约束的各聚簇;步骤 6－步骤 7 对初始划分好的各聚簇中的敏感值是否满足 l 以及 α 约束进行判定,并将不满足条件的聚簇合并,重复迭代执行步骤 1－步骤 7;步骤 8 将剩余不满足条件的记录插入距离最近的聚簇中,且新生成的聚簇满足 k,l,α 约束;步骤 9 对各聚簇中记录的准标识符执行泛化,得到 D';最后对整个数据集进一步执行个性化的匿名操作,得到满足个性化(α,l)-多样性 k-匿名模型的表 D^*。

5 实验与结果分析

本文中的实验数据来源于隐私保护研究领域中被广泛使用的美国 UCI 机器学习仓库的 Adult 数据集[20]。该数据集中包含了 48842 条记录,去除数据集中具有空值和不确定信息的记录,得到一个包含了 30169 条记录的数据表。本文与文献[16]一样,选取 Age, Work-class, Education, Native-Country, Marital Status, Race, Gender 这 7 个属性作为准标识符属性,并将表 4 中的 8 个"Disease"属性值随机添加到数据表中的每条记录作为敏感属性的取值,其中敏感属性"Disease"的各敏感值的 Sid 与表 4 中的保持一致。从数据集中随机抽取 1/4 的记录,为它们随机设置隐私保护级别 Ppl,$Ppl=\{1,2,3,4\}$。此外,为数据集中的每一个属性都构造一棵泛化树,数据集结构设置如表 8 所列。实验环境为:Intel Core i5 2.5GHz CPU,4GB RAM,Windows 10 专业版 64 位操作系统;编程语言为 Java。

表 8 数据集的结构特征

Table 8 Structure features of dataset

属性	类型	取值个数	泛化层次树高度
Age	numeric	74	6
Work-class	categorical	8	3
Education	categorical	16	4
Native Country	categorical	41	3
Marital Status	categorical	7	3
Race	categorical	5	3
Gender	categorical	2	2
Disease	categorical	8	4

为了验证个性化(α,l)-多样性 k-匿名模型的性能,本文在相同的实验条件下将其与 k-匿名模型以及(α,l)-多样性 k-匿名模型进行比较。所有实验中,l 和 α 的取值均固定($l=4$,$\alpha=0.8$)。主要分析 k 值以及数据集大小的变化对实验结果的影响,并分别从信息损失、敏感属性值识别率以及运行时间 3 个方面来比较 3 种模型的性能。每组实验重复进行 10 次,取相应结果的平均值作为最后的对比数据。

5.1 信息损失量的比较

信息损失量采用式(3)计算。图 2 给出了在数据集大小为 30169、k 值变化的情况下,3 种隐私模型下的信息损失量情况。由图可知,随着 k 值的增加,3 种模型下数据集的信息损失量均在不断增加。这是因为 k 的大小直接影响着数据集中等价组的大小,k 值越大,同一等价组中需要泛化的准标识符属性值的数量越多,从而信息损失量越大。当 k 值增加到一定程度时,(α,l)-多样性 k-匿名模型与 k 匿名模型的信息损失量的差异不大,这是因为固定的 l 和 α 的取值在 k 值增加到一定程度时对数据集的约束不明显。总体上,个性化的(α,l)-多样性 k-匿名模型的信息损失量较其他两种模型更大,这是因为其增加了对敏感值的个性化保护。

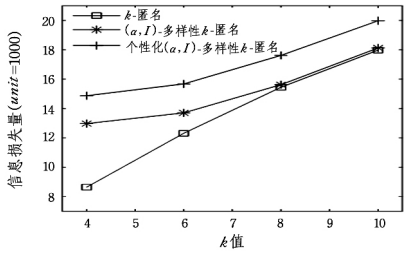

图 2 不同 k 值下的信息损失量

Fig. 2 Information loss with different k values

图 3 给出了在 $k=5$、数据集大小变化的条件下,3 种模型的信息损失量比较。可以看到,随着数据集大小的增加,3 种模型的信息损失量都在增长,这是因为随着需要进行匿名化操作的记录条数的增多,需要处理的属性值的个数也会相应增加,从而导致信息损失量越大。k-匿名模型下数据集的信息损失量总是小于另外两者,这是因为该模型只是对等价组的大小进行约束;而个性化(α,l)-多样性 k-匿名模型比(α,l)-多样 k-匿名模型带来的信息损失量更大,则是因为其增加了对部分敏感属性值的匿名化操作。

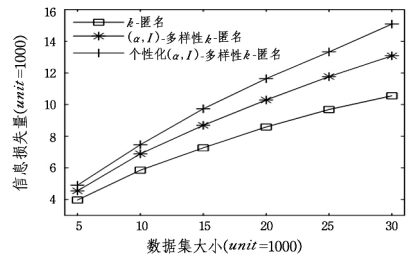

图 3 不同数据集规模下的信息损失量

Fig. 3 Information loss with different dataset sizes

5.2 敏感值识别率的比较

数据集的敏感值识别率采用式(6)计算。图 4 给出了在图 2 中相同的数据背景下,当 k 值变化时,3 种隐私模型的敏感值识别率。可以看到,随着 k 值的增加,3 种模型的数据集

敏感值的平均识别率均在降低。其中，个性化(α，l)-多样性 k-匿名模型与(α，l)-多样性 k-匿名模型的敏感值识别率的变化趋势基本相同，但由于个性化(α，l)-多样性 k-匿名模型增加了对敏感值的个性化保护，因此较其他两种模型而言，其敏感值的识别率总是最低的，隐私保护效果最好。

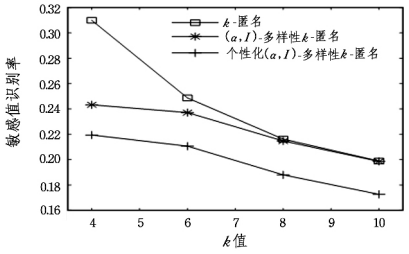

图 4　不同 k 值下的敏感值识别率

Fig. 4　Recognition rate of sensitive value with different k values

图 5 给出了当 k=5 时不同数据集大小下的敏感值识别率的比较。可以看到，总体上数据集大小的改变对 3 种模型的敏感值识别率的影响不大。相同条件下，3 种模型中个性化(α，l)-多样性 k-匿名模型下数据集的敏感值识别率是最小的。

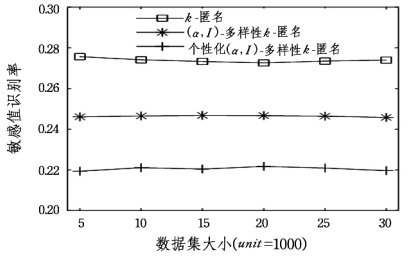

图 5　不同数据集规模下的敏感值识别率

Fig. 5　Recognition rate of sensitive value with different dataset sizes

5.3　执行时间的比较

图 6 和图 7 分别给出了 k 值以及数据集大小变化的情况下 3 种模型的运行时间比较。图 6 中，k 值变化的情况下，3 种模型的运行时间有一定的起伏，当 k=4 时，(α，l)-多样性 k-匿名模型与个性化(α，l)-多样性 k-匿名模型中隐私参数 l 和 α 的设置增加了划分等价组时判定的时间，因此运行时间较长。随着 k 值的增加，3 种模型的运行时间差别不大。图 7 中，k=5 时，随着数据集大小的增加，3 种模型的运行时间都在增长，这是由于所需处理的记录数量增加造成的。可以很明显地看到，由于划分等价组时初始元组的随机选择会影响各模型的实验结果，个性化(α，l)-多样性 k-匿名模型虽然增加了对敏感属性值的匿名化操作，但在某些情况下能使运行时间短于(α，l)-多样性 k-匿名模型。

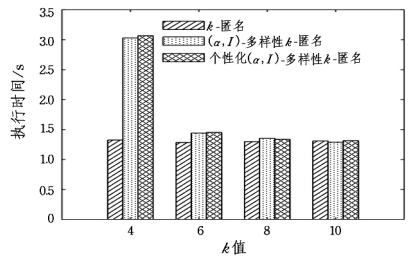

图 6　不同 k 值下的执行时间

Fig. 6　Execution time with different k values

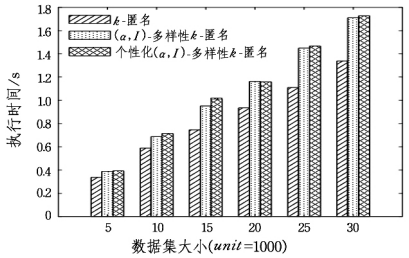

图 7　不同数据集规模下的执行时间

Fig. 7　Execution time with different dataset sizes

结束语　为了满足隐私保护数据发布中个性化的隐私保护需求，文中提出一种新的个性化(α，l)-多样性 k-匿名模型。该模型有效地集成了个性化匿名的两种机制，在数据集满足 k-匿名和 l-多样性模型的基础上，依据敏感程度的不同为敏感属性取值划分类别，限制敏感级别相同的敏感值在同一个等价组中出现的频率；通过为敏感属性构造泛化树的方法，允许特定个人为自己的敏感属性值设置隐私保护级别，并依据个性化隐私保护规则提供相应的保护。实验结果表明，个性化(α，l)-多样性 k-匿名模型在特定的信息损失下，能够有效地提供更强的隐私保护。需要指出的是，隐私保护强度的增加会在一定程度上影响数据的可用性。因此，确定本文模型中的数据可用性与隐私保护程度之间的关系，以及如何优化本文算法来进一步降低信息损失，是未来研究中需要解决的问题。此外，本文主要考虑单一敏感属性的个性化匿名，下一步也将针对多敏感属性的个性化隐私保护问题进行研究。

参 考 文 献

[1] MENG X F，ZHANG X J. Big Data Privacy Management[J]. Journal of Computer Research and Development，2015，52(2)：265-281.(in Chinese)
孟小峰，张啸剑. 大数据隐私管理[J]. 计算机研究与发展，2015，52(2)：265-281.

[2] JIANG H W，ZENG G S，MA H Y. Greedy clustering-anonymity method for privacy preservation of table data-publishing[J]. Joural of Software，2017，28(2)：341-351.(in Chinese)
姜火文，曾国荪，马海英. 面向表数据发布隐私保护的贪心聚类匿名方法[J]. 软件学报，2017，28(2)：341-351.

[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] MACHANAVAJJHALA A，KIFER D，GEHRKE J. l-diversity：Privacy beyond k-anonymity[J]. ACM Transactions on Knowledge Discovery from Data，2007，1(1)：3.

[5] LI N，LI T，VENKATASUBRAMANIAN S. Closeness：a new privacy measure for data publishing[J]. IEEE Transactions on Knowledge & Data Engineering，2009，22(7)：943-956.

[6] HAN J M，YU J，YU H Q，et al. Individuation Privacy Preservation Oriented to Sensitive Values[J]. Acta Electronica Sinica，2010，38(7)：1723-1728.(in Chinese)
韩建民，于娟，虞慧群，等. 面向敏感值的个性化隐私保护[J]. 电子学报，2010，38(7)：1723-1728.

[7] XIAO X,TAO Y. Personalized privacy preservation[C]// ACM SIGMOD International Conference on Management of Data. ACM,2006:229-240.

[8] XU Y,QIN X,YANG Z,et al. A personalized k-anonymity privacy preserving method[J]. Journal of Information & Computational Science,2013,10(1):139-155.

[9] WANG P. Personalized Anonymity Algorithm Using Clustering Techniques[J]. Journal of Computational Information Systems,2011,7(3):924-931.

[10] YE X,ZHANG Y,LIU M. A Personalized (a,k)-Anonymity Model[C]// The Ninth International Conference on Web-Age Information Management. IEEE Computer Society,2008:341-348.

[11] HAN J,YU H,YU J,et al. A Complete (α,k)-Anonymity Model for Sensitive Values Individuation Preservation[C]// International Symposium on Electronic Commerce and Security. IEEE,2008:318-323.

[12] SHEN Y,GUO G,WU D,et al. A novel algorithm of personalized-granular k-anonymity[C]// International Conference on Mechatronic Sciences, Electric Engineering and Computer. IEEE,2013:1860-1866.

[13] WANG B,YANG J. A personalized anonymous method based on inverse clustering[J]. Acta Electronica Sinica,2012,40(5):883-890. (in Chinese)
王波,杨静. 一种基于逆聚类的个性化隐私匿名方法[J]. 电子学报,2012,40(5):883-890.

[14] WANG B,YANG J. Research on Anonymity Technique for Personalization Privacy-preserving Data Publishing[J]. Computer Science,2012,39(4):168-171. (in Chinese)
王波,杨静. 数据发布中的个性化隐私匿名技术研究[J]. 计算机科学,2012,39(4):168-171.

[15] PRASSER F,BILD R,EICHER J,et al. Lightning:Utility-Driven Anonymization of High-Dimensional Data[J]. Transactions on Data Privacy,2016,9(2):161-185.

[16] SUN X,WANG H,LI J,et al. Enhanced P-Sensitive K-Anonymity Models for Privacy Preserving Data Publishing[J]. Transactions on Data Privacy,2008,1(2):53-66.

[17] KAN Y Y,CAO T J. Enhanced privacy preserving K-anonymity model:(α,L)-diversity K-anonymity[J]. Computer Engineering and Applications,2010,46(21):148-151. (in Chinese)
阚莹莹,曹天杰. 一种增强的隐私保护 K-匿名模型—(α,L)多样化 K-匿名[J]. 计算机工程与应用,2010,46(21):148-151.

[18] XU J,WANG W,PEI J,et al. Utility-based anonymization for privacy preservation with less information loss[J]. ACM SIGKDD Explorations Newsletter,2006,8(2):21-30.

[19] LIU X,XIE Q,WANG L. Personalized extended (α,k)-anonymity model for privacy-preserving data publishing[J]. Concurrency & Computation Practice & Experience,2017,29(6):e3886.

[20] BLAKE C. UCI repository of machine learning databases[OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html.

(上接第 163 页)

参 考 文 献

[1] ALKHATHAMI M,ALAZZAWI L,ELKATEEB A. Large Scale Border Security Systems Modeling and Simulation with OPNET[C]// Computing and Communication Workshop and Conference (CCWC). IEEE,2017:1-8.

[2] MEHIC M,MAURHART O,RASS S,et al. Implementation of Quantum Key Distribution Network Simulation Module in the Network Simulator NS-3[J]. Quantum Information Processing,2017,16(10):253.

[3] ORGERIE A C,ASSUNCAO M D,LEFEVRE L. A Survey on Techniques for Improving the Energy Efficiency of Large-scale Distributed Systems[J]. ACM Computing Surveys,2014,46(4):1-31.

[4] BOETTIGER C. An introduction to Docker for reproducible research[J]. ACM SIGOPS Operating Systems Review,2015,49(1):71-79.

[5] LUBKE R,BUSCHEL P,SCHUSTER D,et al. Measuring Accuracy and Performance of Network Emulators[C]// 2014 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom). IEEE,2014:63-65.

[6] MIJUMBI R,SERRAT J,GORRICHO J L,et al. Network Function Virtualization:State-of-the-art and Research Challenges[J]. IEEE Communications Surveys & Tutorials,2016,18(1):236-262.

[7] FANG B X,JIA Y,LI A P,et al. Cyber Ranges:State-of-the-art and Research Challenges[J]. Journal of Cyber Security,2016,1(3):1-9. (in Chinese)
方滨兴,贾焰,李爱平,等. 网络空间靶场技术研究[J]. 信息安全学报,2016,1(3):1-9.

[8] YANG R,LIU Y K. Simulation Fidelity Theory and Measurement:A Literature Review[J]. System Simulation Technology,2014,10(2):85-89. (in Chinese)
杨蓉,刘玉坤. 建模与仿真逼真度理论与方法研究综述[J]. 系统仿真技术,2014,10(2):85-89.

[9] SIATERLIS C,GARCIA A P,GENGE B. On the Use of Emulab Testbeds for Scientifically Rigorous Experiments[J]. IEEE Communications Surveys & Tutorials,2013,15(2):929-942.

[10] WROCLAWSKI J,BENZEL T,BLYTHE J,et al. DETERLab and the DETER Project[M]// The GENI Book. Springer International Publishing,2016:35-62.

[11] ROZA Z C. Simulation Fidelity Theory and Practice[D]. Netherlands:TU Delft,2005.

[12] GARDENGHI L,GOLDWEBER M,DAVOLI R. View-os:A New Unifying Approach against the Global View Assumption[C]// International Conference on Computational Science(ICCS 2008). 2008:287-296.

[13] SHUJA J,GANI A,BILAL K,et al. A Survey of Mobile Device Virtualization:Taxonomy and State of the Art[J]. ACM Computing Surveys,2016,49(1):1.

[14] DETER T. Building Apparatus for Multi-resolution Networking Experiments Using Containers:ISI-TR-683[R]. 2011.