软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(8):2138−2148 [doi: 10.3724/SP.J.1001.2012.04157]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于取整划分函数的 *k* 匿名算法*

吴英杰[1,2,3+], 唐庆明[1], 倪巍伟[2], 孙志挥[2]

[1](福州大学 数学与计算机科学学院,福建 福州 350108)

[2](东南大学 计算机科学与工程学院,江苏 南京 210096)

[3](网络系统信息安全福建省高校重点实验室(福州大学),福建 福州 350108)

Algorithm for *k*-Anonymity Based on Rounded Partition Function

WU Ying-Jie[1,2,3+], TANG Qing-Ming[1], NI Wei-Wei[2], SUN Zhi-Hui[2]

[1](College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)

[2](College of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

[3](Key Laboratory of Network System Information Security (Fuzhou University), Department of Education of Fujian Province, Fuzhou 350108, China)

+ Corresponding author: E-mail: yjwu@fzu.edu.cn

Wu YJ, Tang QM, Ni WW, Sun ZH. Algorithm for *k*-anonymity based on rounded partition function. *Journal of Software*, 2012,23(8):2138−2148 (in Chinese). http://www.jos.org.cn/1000-9825/4157.htm

Abstract: This paper proposes an algorithm based on rounded partition function for *k*-anonymity. By rigorous theoretical proof, the study will show that a better upper bound on size of the anonymization groups can be obtained in non-trivial data sets. In particular, when the size of the original dataset is greater than $2k^2$, the upper bound will be reduced to k+1. Further, the average size of all anonymization groups of the anonymous data will be close enough to k when the size of the original dataset is large enough. Experimental results on real datasets show that this algorithm is effective and feasible.

Key words: privacy preservation; data publishing; algorithm for *k*-anonymity; rounded partition function; upper bound on size of anonymization group

摘 要: 提出一种基于取整划分函数的 k 匿名算法,并从理论上证明该算法在非平凡的数据集中可以取得更低的上界.特别地,当数据集大于 $2k^2$ 时,该算法产生的匿名化数据的匿名组规模的上界为 k+1;而当待发布数据表足够大时,算法所生成的所有匿名组的平均规模将足够趋近于 k.仿真实验结果表明,该算法是有效而可行的.

关键词: 隐私保护;数据发布;k 匿名算法;取整划分函数;匿名组规模上界

中图法分类号: TP309 **文献标识码**: A

在现实生活中,由于数据统计和科学研究的需要,许多研究机构或组织都会对外发布数据.如何保证所发布的数据既是可用的,又不会泄漏数据中所包含的个体的隐私信息,成了当前非常热门的研究课题[1−4].为了保护个体的隐私,显标识符(例如姓名、身份证号或信用卡号)将在数据正式发布之前被删除.然而,已有的研究指出,

* 基金项目: 国家自然科学基金(61003057); 福建省自然科学基金(2010J01330)
收稿时间: 2011-01-15; 定稿时间: 2011-11-02

仅仅删除显标识符不足以保障个人隐私信息的安全.因此,相关学者提出了多种针对数据发布的隐私保护机制.k匿名就是其中最早被提出的一种隐私保护机制[2−4].经过多年研究,该机制日趋成熟.由于k匿名机制简单且实用,它已经被引入移动数据库、无线传感器网络和一些管理系统的应用中[5,6].

满足 k 匿名的安全要求的数据往往含有许多的匿名组,且每个匿名组内部的记录是无法区分的.在过去的研究中,研究者都考虑了如何提高满足k匿名安全要求的匿名化数据集的数据质量[1,7−12].一般地,一个最终发布的匿名化数据集中包含的匿名组越多,这个数据集的信息就越丰富;同样地,若数据集的平均匿名组规模越小,这个数据集的可用性也就越高.据我们所知,现有算法所产生的匿名化数据所包含的匿名组的规模在最坏情况下的上界为 $2k-1$[1−4,7−11,13,14].

本文针对关系型数据库设计了一种 k 匿名算法.该算法在非平凡的数据集中可以取得更低的上界;当数据集规模大于$2k^2$时,新算法产生的匿名化数据中所有匿名组规模的上界为k+1;而当待发布数据表足够大时,新算法所生成的所有匿名组的平均规模将足够趋近于k.此外,新算法在时间复杂性方面也具有较好的性能.

1 基础知识和相关算法

1.1 k匿名

本文主要针对关系型数据表进行讨论.在关系型数据中,表是列属性和行元组的一系列数据元素的集合.基于隐私考虑,属性可以分成以下几类:

- 显标识符(ID).显标识符能够唯一确定一个元组(一条用户记录).它(们)在数据发布前必须被删除.不失一般性,可以认为表中每一个元组唯一对应一个用户记录.
- 准标识符(QI).一般来说,准标识符是能够结合其他外部信息,以较高概率识别出目标所对应记录的最小属性集合.事实上,不同的攻击者,根据其背景知识会拥有不同的准标识符.本文采取通常的做法,假设每张表的准标识符是确定的.
- 敏感属性(SA).敏感属性是需要保护的信息,但它一般无法预先获知,也一般无法唯一确定一个用户记录.

我们用 $T(Q_1,Q_2,\ldots,Q_d,S_1,S_2,\ldots,S_m)$来描述一张表,简称为 $T(d)$,其中,d 是准标识符的个数,m 是敏感属性的个数.k 匿名机制要求表中的每一条记录至少与其他 $k-1$ 条记录在准标识符上相一致.令$\Pi_{QI}(T)$为表 $T(d)$在属性集合 QI 上的投影,表 $T(d)$在属性集合 QI 下满足 k 匿名,当且仅当 $\prod_{QI}(T)$ 中的任意一条记录都至少重复出现 k 次.在Π运算符下,有相同值的所有记录组成一个匿名组.对于给定的 k,每一个这样的匿名组可以称为一个 k 匿名组.

许多匿名化技术可使发布数据达到k匿名的安全要求.在这些技术中,概化[1−4,7,9−11,13−17]与剖分[12]是其中最常使用的两种技术.前者一般可以产生更加安全的数据,即便攻击者已经知道攻击目标肯定在某个已发布的数据表中,概化技术所产生的数据依然有安全性上的优势[18].它将每条原始记录的某些属性值替换为一个更一般但语义上相关的值,从而达到隐藏信息的目的.例如,“画家”可被替换为“艺术家”,一个确定的数字可被替换为一个区间段.

例如,表 1 是一张原始的医疗数据表,表 2 是利用概化技术对表 1 进行匿名处理后得到的一张 2 匿名表.

Table 1 Medical data table

表 1 医疗数据表

Name	Age	Zipcode	Disease
Linda	20	101	H1N1
Bill	20	103	HIV
Sam	30	102	FLU
Sarah	40	102	Pneumonia
Mary	50	101	HBV
Jacky	50	103	HIV

Table 2 A 2-anonymous form of Table 1

表 2 表 1 的一种 2 匿名形式

Age	Zipcode	Disease
20	[101~103]	H1N1
20	[101~103]	HIV
[30~40]	102	FLU
[30~40]	102	Pneumonia
50	[101~103]	HBV
50	[101~103]	HIV

1.2 数据质量的度量

对于满足 k 匿名安全要求的匿名化数据而言,可以通过其最大匿名组的规模、匿名组的平均规模以及所含匿名组的总量来衡量其数据质量.匿名组越多、匿名组平均规模越小,匿名化数据就越接近原来的真实数据,信息损失就越少,其可用性也就越高.显然,未经任何匿名化操作的原始数据包含的匿名组最多,且匿名组的规模为 1(或者是其他比较小的数值).因此,原始未匿名化数据的可用性最高、信息最丰富.这显然是合理的.

试想一下,在使用(例如查询)某个对外发布的数据表时,用户总是希望得到比较精确的信息.而匿名组的规模影响了信息返回的精确程度.在原始数据集中,返回的往往是一条单一的记录或若干条相近的记录;而在匿名数据中,返回的将是 1 个或若干个匿名组.减小每个匿名组的粒度,显然将使发布的数据表能够更好地满足用户的需求.

事实上,许多之前的研究都从不同角度探讨了如何避免一个巨大的匿名组的形成[1−4,7−10,13,14].然而迄今为止,经过理论严格证明的最好的匿名组规模上界仍然是 $2k-1$[10].

不少研究者采用某些精心设计的度量函数来衡量匿名化数据的质量.一个度量函数往往从某个角度来考察匿名化数据的质量.根据某个度量函数而设计的算法,一般能够在该度量函数下达到最优或理想的效果.根据以往的文献,最常见的度量函数包括可辨别度量(discernibility metric,简称 DM)函数和分类度量(classification metric,简称 CM)函数.当表中没有元组被删除时,DM 和 CM 可以定义如下:

$$DM(T')=\sum_{\forall E\in T'}|E|^2 \tag{1}$$

$$CM(T')=\sum_{\forall E\in T'}|minority(E)| \tag{2}$$

这里,T'为给定的表 $T(d)$经匿名处理后的发布表,E 是任意的匿名组.CM 需要一个类标识符来将元组分为几个类,$minority(E)$表示 E 中最小的类集合.由于目前 DM 和 CM 已普遍用来度量发布数据质量[1,7,13,14,18],本文的实验也将使用这两个度量函数.

1.3 现有算法的匿名组规模上界

现有的 k 匿名算法可以划分为 3 类:单维概化(single-dimensional generalization,简称 SG)、多维概化(multi-dimensional generalization,简称 MG)和局部编码(local recoding,简称 LR)算法[8,10,13].

给定表 $T(d)$,SG 将域$\bigcup_{1\leqslant i\leqslant d}Q_i$中的每一个元素映射成一个值,MG 将笛卡尔乘积$\prod_{1\leqslant i\leqslant d}Q_i$中的每一个元素映射成一个值.SG 和 MG 均属于全局编码(global recoding,简称 GR).与 GR 相比,LR 的限制条件较少,它允许将$\prod_{1\leqslant i\leqslant d}Q_i$中的每一个元素映射到多个值.

一般认为,LR 比 GR 更有效,也更为灵活[8,10,13].如果我们为表 $T(d)$的每一个属性域定义一个顺序,那么,$\prod_{1\leqslant i\leqslant d}Q_i$可以映射到一个多维空间中,而 $T(d)$的每一条记录都可以看成是该多维空间中的一个点.此时,寻找一张 k 匿名表,等价于寻找与其对应的多维空间中某个多维矩形区域的一个划分.在二维的情况下,这个矩形区域是一个欧几里德意义下的平面矩形;而在三维的情况下,它是一个长方体;在更高维的情况下,这样一个矩形区域在各个平面的投影都应该是一个平面矩形.不失一般性,该矩形区域可以取为在该多维空间中能够覆盖所有记录的最小的矩形区域.这样,每一个 k 匿名组就等价于这个矩形区域中的某个划分子区域,而每个匿名组的规模就是其所对应的矩形区域内所包含的记录的总数量.文献[10]证明了如下事实:SG 和 MG 所产生的匿名化

数据的最大匿名组的规模在最坏情况下是 $O(|T(d)|)$的,匿名组的数量在最坏情况是 1,而匿名组的平均规模在最坏情况下同样也是 $O(|T(d)|)$的.这是一个非常不理想的上界结论.许多之前的工作采取删除记录的方法来避免 SG 和 MG 产生过大的匿名组[2−4,8].而对于 LR 所产生的匿名化数据,它的最大匿名组规模和匿名组平均规模在最坏情况下的上界是不超过 $2k-1$ 的.这是因为对于规模超过 $2k$ 的匿名组,LR 技术总能将它们划分成为两个更小的 k 匿名组.然而,现有的研究所提出的匿名算法的最坏上界要么不优于 $2k-1$,要么缺乏对匿名算法最坏上界的有效分析.本文的工作就是在前人工作的基础上设计一种新算法,改进最坏情况下最大匿名组的上界以及匿名组平均规模的上界,从而达到提高发布数据质量的目的.

2 基于取整划分函数的 k 匿名算法

现有的许多 k 匿名算法都采用基于分治策略的概化技术[1,7,10,14,19].对于一张表 $T(d)$,不妨假设Ω是其对应的多维空间中能覆盖所有记录的最小的多维矩形区域.一种有效的基于分治策略和概化技术的 k 匿名算法框架是:先将Ω划分为两个多维矩形区域;然后再递归地将每个小区域划分为更小的子多维矩形区域,直到所得到的区域不能再被划分为更小的满足 k 匿名安全要求的区域为止;最后,对每个子区域中的记录进行概化,使得它们具有相同的 QI 值,从而形成一个匿名组.以上划分过程被称为二划分,图 1 给出了二划分的基本过程.事实上,这种划分是将一个大的问题不断分解为基于一系列较小的多维矩形区域上的问题.显然,在基于 LR 技术的划分过程中,每一个所含记录总量不少于 $2k$ 的多维矩形区域总是可以继续划分.下文称那些在二划分过程中产生的匿名组(多维矩形区域)为临时匿名组.

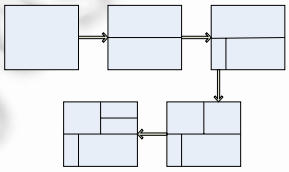


Fig.1 Binary partitioning

图 1 二划分过程

二划分过程的关键是采取什么样的策略将一个较大的临时匿名组划分为两个相对较小的匿名组.Mondrian 算法[10]是现有文献中使用二划分的先例,该算法采用均衡策略,每次将一个大的临时匿名组划分为两个容量尽可能相等的较小的匿名组.Mondrian 比许多基于 MG 和 SG 的算法要有效得多,它产生的匿名化数据的匿名组大小在最坏情况下不会超过 $2k-1$.Mondrian 已被作为一种算法框架成功地应用到其他隐私保护机制上,例如 l-多样性模型[20].

2.1 均衡二划分存在的问题

假设待匿名的数据表记录数为 3×2^k,并且假设目标是发布 2 匿名的数据表.那么,均衡二划分会将整个数据集表分成 2 个子数据表,记录数均为 $3\times2^{k-1}$.而后,继续对这 2 个子数据表分别进行均衡二划分,形成 4 个数据表,每个数据表的记录数都是 $3\times2^{k-2}$.依此策略,可以用数学归纳法证明,最后将得到 2^k 个 3 匿名组.而事实上,应该可以产生 $3\times2^{k-1}$ 个 2 匿名组.

例如,表 1 经过映射后可被看作是一个含有 6 个点的平面矩形.如果数据发布者希望表 1 发布后满足 2 匿名要求,则采用均衡二划分策略只能将其划分成两个匿名组,且每个匿名组的大小为 3,如图 2 所示.然而,我们显然可以将这个表划分成含有 3 个匿名组,每个匿名组有且仅有两条记录,如图 3 所示.

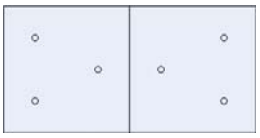


Fig.2 Balance binary partitioning
图 2 均衡二划分

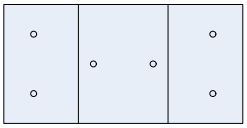


Fig.3 A better partitioning schema
图 3 一个更好的划分结果

不难看出,Mondrian 所采用的“均衡划分”策略本质上是一种局部贪心的策略,这种策略体现了局部公平性,并不具备“全局眼光”.依此策略的划分过程,将可能减少潜在的匿名组数量.

一般地,假设表 $T(d)$有 $n=d\times k+r$ 条记录.其中,r 是一个比 k 小的非负数.那么表格 $T(d)$在理论上应该可以划分为 d 个匿名组.设 $T(d)$经若干次划分后,其中某个临时匿名组 X 的容量为 $n'=d'\times k+r'$.这里,r'同样是一个比 k 小的非负数.如果 d'是奇数,则均衡二划分所产生的两个更小的临时匿名组的容量将分别是 $\frac{d'-1}{2}\times k+\left\lceil\frac{r'+k}{2}\right\rceil$ 和 $\frac{d'-1}{2}\times k+\left\lfloor\frac{r'+k}{2}\right\rfloor$.而这两个子匿名组最多能够被继续划分为共 $d'-1$ 个匿名组.因此,当 d'为奇数时,使用均衡二划分实际上减少了潜在的匿名组数量.

2.2 基于取整划分函数的划分策略

不妨假设上一节中的临时匿名组 X 被划分成两个子匿名组 X_1 和 X_2,且其规模分别为 $\alpha_1k+\beta_1$ 和 $\alpha_2k+\beta_2$.显然,$\alpha_1+\alpha_2\leqslant d'$.若希望最大化可能产生的匿名组数量,就必须让 $\alpha_1+\alpha_2$ 尽量大.对于上述不等式而言,等号成立的充分必要条件是 $\beta_1+\beta_2=r'$.基于此,我们设计如下的划分函数,其划分后两个匿名组的容量规模分别为

$$\begin{cases}X_1:|X_1|=\left\lfloor\frac{d'}{2}\right\rfloor k+\left\lfloor\frac{r'}{2}\right\rfloor\\ X_2:|X_2|=\left\lceil\frac{d'}{2}\right\rceil k+\left\lceil\frac{r'}{2}\right\rceil\end{cases}.$$

上式即为本文算法中最为重要的划分函数,其中,开口向上的符号是下取整函数,开口向下的符号是上取整函数.下取整函数求得一个不超过给定数的最大整数,而上取整函数求得一个不小于给定数的最小整数.由此划分式能够引导产生上界更优的 k 匿名方案.在本节中,我们先给出算法的详细描述.

给定表格 $T(d)$,我们首先为 QI 的每个属性在其对应属性域上的取值定义一个顺序,使得每个属性 Q_i 的域成为有序域;接着,按照属性域的序将属性域上的所有元素一一映射到实数域中.具体而言,对于每个 Q_i,都存在一个和它对应的实域序列$\{q(i,1),q(i,2),\ldots,q(i,t_i)\}$.这里,$q(i,j)$对应着 Q_i 的域中的第 j 个元素,且 $1\leqslant i\leqslant d,1\leqslant j\leqslant t_i=|Q_i|$.由此所有的记录都可被看作是一个 d 维正交空间中的一个点.我们用 P 表示所有点所形成的集合,用 Ω 表示这样一个 d 维空间中能够覆盖 P 的最小的多维矩形区域.同时,用 $\prod_i(p)$ 表示一个点 p 在这个 d 维空间中的第 i 维上的投影.

算法. 基于取整划分函数的 k 匿名算法(划分部分).

算法输入:表 $T(d),\Omega,P,k$.

步骤 1:令 $S=\Omega,TMP=P,|P|=\alpha k+\beta$,其中,$\beta$是比 k 小的非负数.

步骤 2:选择Ω的任意一维 i,并找到一个合适的正整数 j,使得 $\left|\bigcup_{p\in TMP\wedge\prod_i p\leqslant q(i,j)}\right|\geqslant\left\lceil\frac{\alpha}{2}\right\rceil k+\left\lceil\frac{\beta}{2}\right\rceil$, 并且

$$\left|\bigcup_{p\in TMP\wedge\prod_i p\geqslant q(i,j)}\right|\geqslant\left\lfloor\frac{\alpha}{2}\right\rfloor k+\left\lfloor\frac{\beta}{2}\right\rfloor.$$

步骤 3:从 j 处将 S 划分成为两个多维矩形区域 S_1 和 S_2.

步骤 4:将 TMP 划分成为两个点集 P_1 和 P_2:$|P_1|=\left\lceil\frac{\alpha}{2}\right\rceil k+\left\lceil\frac{\beta}{2}\right\rceil$,$|P_2|=\left\lfloor\frac{\alpha}{2}\right\rfloor k+\left\lfloor\frac{\beta}{2}\right\rfloor$,并且要求对于 P_1 中的任意一个元素 p,都有 $\prod_i(p)\leqslant q(i,j)$;而对于 P_2 中的任意一个元素 p,都有 $\prod_i(p)\geqslant q(i,j)$.$P_1$ 的点属于 S_1,P_2 的点属于 S_2.

步骤 5:如果$|P_1|\geqslant 2k$,则利用参数 S_1,P_1 和 k 继续执行.

步骤 6:如果$|P_2|\geqslant 2k$,则利用参数 S_2,P_2 和 k 继续执行.

不妨设临时匿名组 X 在第 i 维上的投影是线段$(q(i,x),q(i,y))$,$1\leqslant x\leqslant y\leqslant|Q_i|$,那么 X 所包含的点集 P_X 中的任意一个点 p 在第 i 维上的投影都在线段$(q(i,x),q(i,y))$中,且为 Q_i 对应的实域序列中的某个元素.易知,线段$(q(i,x),q(i,y))$中共有 $y-x+1$ 个 Q_i 对应的实域序列元素.此时,设置 $y-x+1$ 个对应的计数器,遍历 P_X 一遍,根据每个点在第 i 维上的投影来改变当前计数器的数值.遍历 P_X 完成后,按顺序遍历计数器,即可找到满足步骤 2 的 j 分割线.

基于取整划分函数的 k 匿名算法总是试图将一个临时匿名组(多维矩形区域)划分为两个更小的多维矩形区域,这两个区域相交于算法中所提到的 j 分割线.若 j 分割线上没有元素,则分割结束;若 j 分割线上有元素,则将这些元素分配到两个多维矩形区域中,从而形成两个新的子临时匿名组.

例如,给定 k=2.图 4 所示的临时匿名组共含有 7 个点,7=3×2+1.从左到右共有 6 个计数器,其计数值分别为{0,3,1,0,1,2}.其中,最左边的是 $q(i,x)$,最右边的是 $q(i,y)$.j 分割线处于倒数第 2 个计数器处(其对应计数器值为 1).此时,$0+3+1+0+1\geqslant\left\lceil\frac{3}{2}\right\rceil\times 2+\left\lceil\frac{1}{2}\right\rceil=5$,且$1+2\geqslant\left\lfloor\frac{3}{2}\right\rfloor\times 2+\left\lfloor\frac{1}{2}\right\rfloor=2$,满足算法步骤 2 的条件.两个矩形区域有一个交点,而这个交点最后分配给左边的矩形区域.

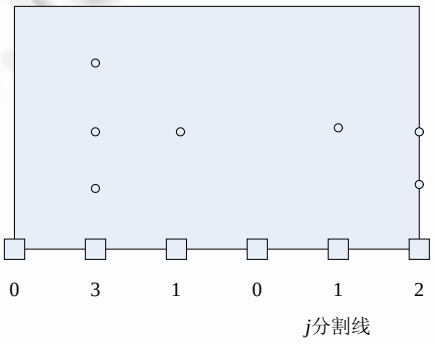


Fig.4 j partitioning line

图 4 j 分割线

若令 $n=|T(d)|$,则根据之前对寻找 j 分割线的描述可知,对于某个临时匿名组,在寻找其 j 分割线时,最多需要遍历该临时匿名组内的所有记录.因此,其时间耗费为 $O(n)$;另外,在 k 匿名限制下,表格 $T(d)$最多能分成$\left\lfloor\frac{n}{k}\right\rfloor$个多维矩形区域,因此,表格最多被划分$\left\lfloor\frac{n}{k}\right\rfloor-1$次.因而,对于固定的 k,整个算法的时间开销是 $O\left(\frac{n^2}{k}\right)$.

然而,许多临时匿名组的规模实际上都远小于 n.因此,上述时间复杂度分析是松弛的.第 2.4 节还将对算法时间复杂度进行分析.

另外,以上算法只描述了划分的部分.当将Ω划分为无法再细分的众多矩形区域时,需要将所有这些区域输出,从而形成一张完整的匿名表.这也就是分治中“合”的部分.一般可使用合适的数据结构(例如 KD 树)来存储和管理分割过程中形成的每一个矩形区域,而在分割结束后再输出重构成完整的匿名表[10,13].

2.3 基于取整划分函数的*k*匿名算法的匿名组规模上界

本节讨论第 2.2 节中算法所产生的匿名化数据在匿名组数量以及匿名组规模的上界.为便于证明,首先给出两个定义.

定义 1(*k* 系数). 给定表格或临时匿名组 X 和 k,若$|X|=\alpha k+\beta$,则称α为 X 的 k 系数.

定义 2(剩余 *k* 系数). 给定表格或临时匿名组 X 和 k,若$|X|=\alpha k+\beta$,则称β为 X 的剩余 k 系数.

定理 1. 给定表 $T(d)$和 k,若$|T(d)|=\alpha k+\beta$,则基于取整划分函数的 k 匿名算法所产生的匿名化数据恰好包含α个匿名组.

证明:首先,当α等于 1 时,结论显然成立.当α大于等于 2 时,$T(d)$会被分成两个部分,每个部分的大小分别是$\left\lfloor\frac{\alpha}{2}\right\rfloor k+\left\lfloor\frac{\beta}{2}\right\rfloor$和$\left\lceil\frac{\alpha}{2}\right\rceil k+\left\lceil\frac{\beta}{2}\right\rceil$.这两个部分的 k 系数之和等于α.同样,对于某个临时匿名组 X,当它被划分之后,其两个子匿名组的 k 系数之和等于 X 的 k 系数.如此,记某时刻所有临时匿名组的 k 系数之和为 Tp,当任意多临时匿名组进行划分之后,所得的所有新匿名组的 k 系数之和必然仍为 Tp.此外,任意一个临时匿名组必然可以递归地划分成为 k 系数为 1 的子临时匿名组的并.因此,定理 1 得证. □

定理 2. 给定表 $T(d)$和 k,且$|T(d)|=\alpha k+\beta$.若$[\log_2\alpha]=x$,则基于取整划分函数的 k 匿名算法得到的所有最终匿名组规模均不超过$k+\left\lfloor\frac{\beta}{2^x}\right\rfloor$.

为了证明定理 2,下面先给出若干定义和引理.

定义 3(匿名组层次). 若称原始表格 $T(d)$是第 0 层的匿名组,则称由第 i–1(i>0)层的临时匿名组划分后形成的子临时匿名组为第 i 层匿名组.

引理 1. 若第 i 层的某个匿名组的剩余 k 系数不超过$\left\lfloor\frac{\beta}{2^i}\right\rfloor$,则由它产生的第 i+1 层的匿名组的剩余 k 系数必然不超过$\left\lceil\frac{\beta}{2^{i+1}}\right\rceil$.

证明:不妨设$\frac{\beta}{2^i}=\theta_1+\theta_2$.其中,$\theta_1$ 是整数部分,$0\leqslant\theta_2<1$.显然,$\left\lceil\frac{\beta}{2^i}\right\rceil\leqslant\theta_1+1$.根据本文的取整划分函数,由该匿名组所产生的第 i+1 层的两个匿名组的剩余 k 系数必然都不超过$\left\lceil\left\lceil\frac{\beta}{2^i}\right\rceil\Big/2\right\rceil$.

- 如果θ_1是奇数,则$\left\lceil\frac{\beta}{2^{i+1}}\right\rceil=\left\lceil\frac{\theta_1+\theta_2}{2}\right\rceil=\left\lceil\frac{\theta_1-1}{2}+\frac{1+\theta_2}{2}\right\rceil=\frac{\theta_1-1}{2}+1=\frac{\theta_1+1}{2}$,而$\frac{\theta_1+1}{2}=\left\lceil\frac{\theta_1+1}{2}\right\rceil\geqslant\left\lceil\left\lceil\frac{\beta}{2^i}\right\rceil\Big/2\right\rceil$;
- 若θ_1是偶数,则$\left\lceil\frac{\beta}{2^{i+1}}\right\rceil=\left\lceil\frac{\theta_1+\theta_2}{2}\right\rceil=\left\lceil\frac{\theta_1}{2}+\frac{\theta_2}{2}\right\rceil=\frac{\theta_1}{2}+1=\frac{\theta_1+2}{2}$,而$\frac{\theta_1+2}{2}=\left\lceil\frac{\theta_1+1}{2}\right\rceil\geqslant\left\lceil\left\lceil\frac{\beta}{2^i}\right\rceil\Big/2\right\rceil$.

因此,无论θ_1是奇数还是偶数,新产生的两个第 i+1 层的匿名组的剩余 k 系数大小必然都不超过$\left\lceil\frac{\beta}{2^{i+1}}\right\rceil$.引理 1 得证. □

引理 2. 给定表 $T(d)$和 k,且$|T(d)|=\alpha k+\beta$.若$[\log_2\alpha]=x$,则任意一个第 i 层匿名组 X 的 k 系数α_i必然满足如下的条件:$2^{x-i}\leqslant\alpha_i<2^{x-i+1}$.

证明:下面采用数学归纳法来证明.

当 i 等于 0 时,显然有$2^x=2^{[\log_2\alpha]}\leqslant 2^{\log_2\alpha_i}=\alpha_i<2^{[\log_2\alpha_i]+1}=2^{x+1}$.因此,引理 2 成立.

不妨假设结论对于第 i 层总是成立.对于第 i 层的某个匿名组 X,其划分出的某个第 i+1 层匿名组的 k 系数为α_{i+1},则$\left\lfloor\frac{\alpha_i}{2}\right\rfloor\leqslant\alpha_{i+1}\leqslant\left\lceil\frac{\alpha_i}{2}\right\rceil$.而$\left\lfloor\frac{\alpha_i}{2}\right\rfloor\geqslant\left\lfloor\frac{2^{x-i}}{2}\right\rfloor=2^{x-(i+1)}$,并且有$\left\lceil\frac{\alpha_i}{2}\right\rceil<\left\lceil\frac{2^{x-i+1}}{2}\right\rceil=2^{x-(i+1)+1}$.因而,仍然有 $2^{x-(i+1)}\leqslant\alpha_{i+1}<2^{x-(i+1)+1}$ 成立.引理 2 得证. □

利用引理 1、引理 2,可构造出定理 2 的证明如下:

定理 2 证明:根据引理 2,易知所有的最终匿名组实际上都是第 x 层的匿名组.而根据引理 1 可知,第 x 层的匿名组的剩余 k 系数必然不会超过 $\left\lceil \frac{\beta}{2^x} \right\rceil$.因此,第 x 层匿名组的大小也必然不会超过 $k+\left\lceil \frac{\beta}{2^x} \right\rceil$.定理 2 得证. □

根据定理 1 和定理 2,可以得到几个简单的推论.

推论 1. 当$|T(d)|\geqslant 2k$并且$k>3$时,基于取整划分函数的 k 匿名算法所产生的匿名组,其规模在最坏情况下小于 $2k-1$,换言之,是一个比 $2k-1$ 更优的上界.

推论 2. 当$|T(d)|\geqslant 2k^2$时,基于取整划分函数的 k 匿名算法所产生的匿名组,其规模在最坏情况下为 $k+1$,且在$|T(d)|$可以被 k 整除时为 k.

推论 3. 给定 k,当$|T(d)|$足够大时,基于取整划分函数的 k 匿名算法所产生的所有匿名组的平均规模将足够趋近于 k.

以上3个推论说明了如下事实:在非平凡条件下,基于取整划分函数的k匿名算法所产生的匿名化数据的匿名组规模上界总是优于之前的最好结果 $2k-1$;而在海量数据条件下,基于取整划分函数的 k 匿名算法将产生在最大匿名组规模的上界上更优的匿名数据.

2.4 基于取整划分函数的k匿名算法(划分部分)时间复杂度分析

本节进一步讨论本文算法的时间复杂度.

定理 3. 将所有第 i 层的临时匿名组按照基于取整划分函数的策略进行划分,其时间耗费是 $O(n)$.这里,$n=|T(d)|$,而 $T(d)$是给定的原始数据表.

证明:根据第 2.2 节的分析,对于某个临时匿名组 X 进行划分,其时间复杂度是 $O(|X|)$,而所有第 i 层的临时匿名组规模之和必然小于 $n=|T(d)|$.因此,将所有第 i 层的临时匿名组按照基于取整划分函数的策略进行划分,其时间复杂度必然是 $O(n)$.定理得证. □

定理 4. 基于取整划分函数的 k 匿名算法(划分部分)的时间复杂度是 $O\left(n\log_2\frac{n}{k}\right)$.

证明:若 $n=|T(d)|=\alpha k+\beta$,而$[\log_2\alpha]=x$,则根据引理 2,$T(d)$最多可以划分到第 x 层;而根据定理 3,产生每层的所有临时匿名组的时间复杂度是 $O(n)$.因而,整个算法的总时间复杂度是 $O(nx)$,也即 $O\left(n\log_2\frac{n}{k}\right)$.定理得证. □

3 实 验

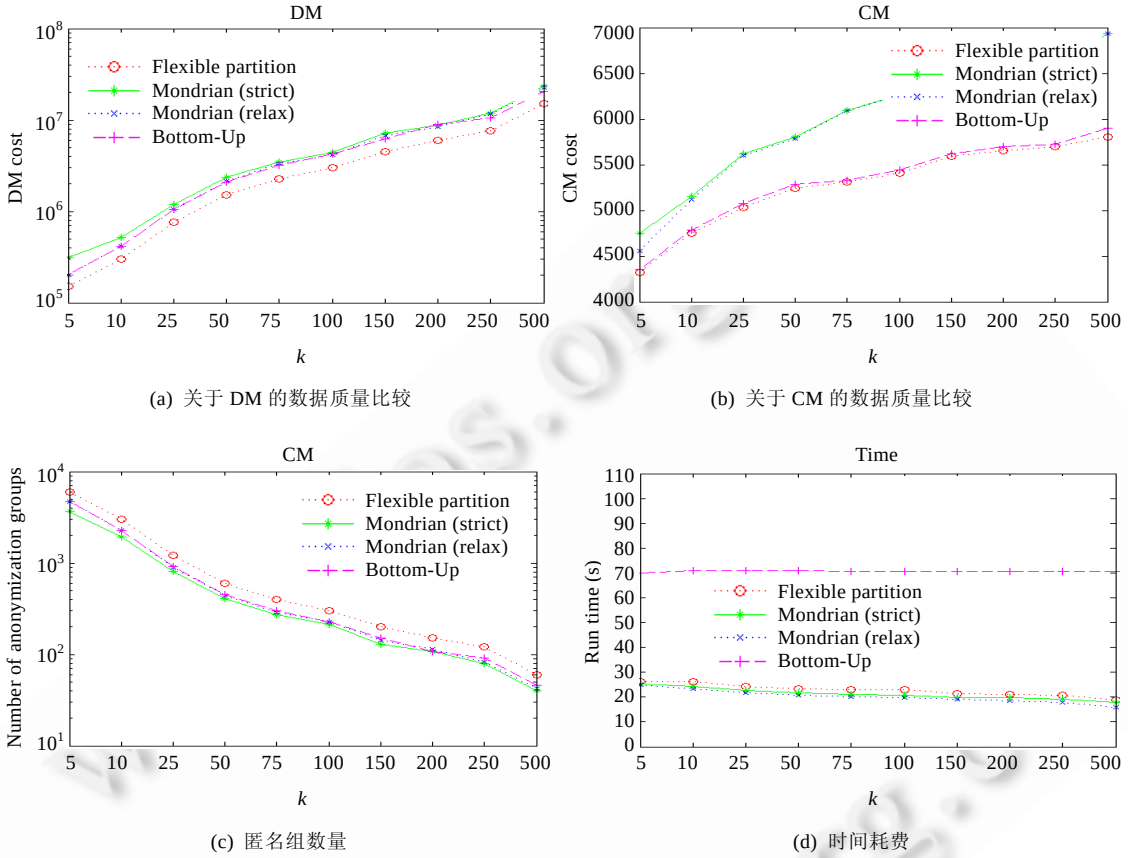
本节介绍我们的新算法与几个相关的著名算法(包括严格 Mondrian 算法[13]、松弛 Mondrian 算法[13]以及 Bottom-Up 算法[19])的实验比较情况.采用经常被使用的 DM 和 CM 评价函数来评价上列算法所产生的匿名化数据的质量,并且比较各种算法的时间开销,其目的是验证基于取整划分函数的 k 匿名算法不但拥有期望的理论上界,同时拥有可接受的时间耗费,而且其所产生的匿名化数据在常用评价函数(DM,CM)下也是比较好的.

本文中的所有实验都是在统一的平台下进行的:奔腾 4 双核 2.2Hz 处理器,4G 内存,Windows XP 操作系统.所使用的数据来自美国 Adult 数据库[21],这是隐私保护领域最常被引用的数据库,之前的相关研究几乎都是将其作为实验中的对比数据[1−4,7−14,19].在删除该数据库中所有不完整的数据记录以后,共有 30 162 条记录.我们选择其中的 7 个常规属性(年龄、工作类别、教育背景、婚姻状况、职业、种族、性别、国籍)作为实验中的 QI,并且使用工资属性为 CM 的分类标识.图 5 给出了比较分析的实验结果.

在图 5 中,Flexible Partition 表示基于取整划分函数的 k 匿名算法,Mondrian(strict)表示严格 Mondrian 算法,而 Mondrian(relax)表示松弛 Mondrian 算法.图 5(a)给出了 4 种算法所产生的匿名化数据关于 DM 的变化情况.可以看出,Flexible Partition 曲线处于所有曲线的下方,也即它受到的信息损失惩罚最小,有较好的信息可用性.另外,根据前述证明结论,随着 k 的增大,匿名组大小上界也随之上升.而在图 5(a)中,4 条曲线都随着 k 的增大而上升,受到更多的信息惩罚.图 5(b)给出了 4 种算法所产生的匿名化数据关于 CM 的变化情况.其上升趋势同图

5(a)一致,而 Flexible Partition 所代表的曲线仍然处于所有曲线的最下方.

图 5(c)是 4 种算法所产生的匿名化数据所含有的匿名组数量的比较.之前已经证明,我们提出的算法将产生匿名组规模上界更低的数据,即产生更多的匿名组.而在实验中,Flexible Partition 也确实处于其他 3 条曲线的上方.图 5(d)给出了 4 种算法的时间耗费.可以看出,基于取整划分函数的 k 匿名算法在时间开销上也是可以接受的.



(a) 关于 DM 的数据质量比较 (b) 关于 CM 的数据质量比较

(c) 匿名组数量 (d) 时间耗费

Fig.5 Experimental results

图 5 实验结果

4 结 论

本文设计了一种有效的基于上下取整划分函数的 k 匿名算法,并从理论上证明了该算法在非平凡数据集中总能取得比 $2k-1$ 更低的上界;且当数据集的大小超过 $2k^2$ 时,算法所产生的匿名化数据的匿名组规模必然不会超过 $k+1$.此外,在面对海量数据时,算法所产生的数据的平均匿名组规模可以足够趋近于 k.仿真实验结果表明,基于取整划分函数的 k 匿名算法是有效而可行的.

在今后的研究工作中,我们将继续探讨 k 匿名机制应用在移动数据保护时的数据质量,以及其他相关的隐私保护机制的数据质量的理论上界.

References:

[1] Bayardo RJ, Agrawal R. Data privacy through optimal *k*-anonymization. In: Aberer K, Franklin M, Nishio S, eds. Proc. of the 21st IEEE Int'l Conf. on Data Engineering. Washington: IEEE Computer Society, 2005. 217−228. [doi: 10.1109/ICDE.2005.42]

[2] Samarati P, Sweeney L. Protecting privacy when disclosing information: *k*-Anonymity and its enforcement through generalization and suppression. Technical Report, SRI Int'l, 1998.

[3] Sweeney L. Achieving *k*-anonymity privacy protection using generalization and suppression. Int'l Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems, 2002,10(5):571−588. [doi: 10.1142/S021848850200165X]

[4] Sweeney L. *k*-Anonymity: A model for protecting privacy. Int'l Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002,10(5):557−570. [doi: 10.1142/S0218488502001648]

[5] Xu Y, Wang K, Fu AWC, Yu PS. Anonymizing transaction databases for publication. In: Li Y, Liu B, Sarawagi S, eds. Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2008. 767−775. [doi: 10.1145/1401890.1401982]

[6] Terrovitis M, Mamoulis N, Kalnis P. Anonymity in unstructured data. Technical Report. Hong Kong: Hong Kong University, 2008.

[7] Fung BCM, Wang K, Yu PS. Top-Down specialization for information and privacy preservation. In: Aberer K, Franklin M, Nishio S, eds. Proc. of the 21st IEEE Int'l Conf. on Data Engineering. Washington: IEEE Computer Society, 2005. 205−216. [doi: 10.1109/ICDE.2005.143]

[8] Fung BCM, Wang K, Chen R, Yu PS. Privacy-Preserving data publishing: A survey on recent developments. ACM Computing Surveys, 2010,42(4):1−53. [doi: 10.1145/1749603.1749605]

[9] Iyengar VS. Transforming data to satisfy privacy constraints. In: Hand D, Keim D, Ng R, eds. Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2002. 279−288. [doi: 10.1145/775047.775089]

[10] Lefevre K, Dewitt DJ, Ramakrishnan R. Mondrian multidimensional *k*-anonymity. In: Liu L, Reuter A, Whang K, Zhang J, eds. Proc. of the 22nd IEEE Int'l Conf. on Data Engineering. Washington: IEEE Computer Society, 2006. 25−25. [doi: 10.1109/ICDE.2006.101]

[11] Wang K, Yu PS, Chakraborty S. Bottom-Up generalization: A data mining solution to privacy protection. In: Wu X, ed. Proc. of the 4th IEEE Int'l Conf. on Data Mining. Washington: IEEE Computer Society, 2004. 249−256. [doi: 10.1109/ICDM.2004.10110]

[12] Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation. In: Dayal U, Whang K, Lomet DB, Alonso G, Lohman GM, Kersten ML, Cha SK, Kim Y, eds. Proc. of the 32nd Very Large Data Bases. New York: Association for Computing Machinery, 2006. 139−150.

[13] Dewitt DJ, Lefevre K, Ramakrishnan R. Incognito: Efficient full-domain *k*-anonymity. In: Ozcan F, ed. Proc. of the 2005 ACM SIGMOD Int'l Conf. on Management of Data. New York: Association for Computing Machinery, 2005. 49−60. [doi: 10.1145/1066157.1066164]

[14] Hore B, Ch R, Jammalamadaka R, Mehrotra S. Flexible anonymization for privacy preserving data publishing: A systematic search based approach. In: Proc. of the 7th SIAM Int'l Conf. on Data Mining. Philadelphia: Society for Industrial and Applied Mathematics, 2007. 497−502.

[15] Li N, Li T, Venkatasubramanian S. *t*-Closeness: Privacy beyond *k*-anonymity and *l*-diversity. In: Proc. of the 23rd IEEE Int'l Conf. on Data Engineering. Washington: IEEE Computer Society, 2007. 106−115. [doi: 10.1109/ICDE.2007.367856]

[16] Truta TM, Campan A, Meyer P. Generating microdata with *p*-sensitive *k*-anonymity property. In: Jonker W, Petkovic M, eds. Proc. of the 4th VLDB Workshop on Secure Data Management. Berlin: Springer-Verlag, 2007. 124−141.

[17] Xiao X, Tao Y. *m*-Invariance: Towards privacy preserving re-publication of dynamic datasets. In: Chan CY, Ooi BC, Zhou A, eds. Proc. of the 2007 ACM SIGMOD Int'l Conf. on Management of Data. New York: Association for Computing Machinery, 2007. 689−700. [doi: 10.1145/1247480.1247556]

[18] Kifer D. Attacks on privacy and definettis theorem. In: Çetintemel U, Zdonik SB, Kossmann D, Tatbul N, eds. Proc. of the 2009 ACM SIGMOD Int'l Conf. on Management of Data. New York: Association for Computing Machinery, 2009. 127−138. [doi: 10.1.1.173.4501]

[19] Xu J, Wang W, Pei J, Wang X, Shi B, Fu AWC. Utility-Based anonymization using local recoding. In: Eliassi-Rad T, Ungar LH, Craven M, Gunopulos D, eds. Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2006. 785−790. [doi: 10.1145/1150402.1150504]

[20] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. *l*-Diversity: Privacy beyond *k*-anonymity. In: Liu L, Reuter A, Whang K, Zhang J, eds. Proc. of the 22nd IEEE Int'l Conf. on Data Engineering. Washington: IEEE Computer Society, 2006. 24–24. [doi: 10.1145/1217299.1217302]

[21] Frank A, Asuncion A. UCI machine learning repository. Irvine: School of Information and Computer Science, University of California, 2010. http://archive.ics.uci.edu/ml

吴英杰(1979－),男,福建安溪人,博士,副教授,主要研究领域为数据挖掘,数据安全隐私保护.

倪巍伟(1979－),男,博士,副教授,CCF 会员,主要研究领域为数据挖掘,数据安全隐私保护.

唐庆明(1985－),男,硕士,主要研究领域为数据挖掘,数据安全隐私保护.

孙志挥(1941－),男,教授,博士生导师,CCF 高级会员,主要研究领域为复杂信息系统集成,数据库知识发现.