隐私保护中K-匿名模型的综述

岑婷婷[1],韩建民[1,2],王基一[1],李细雨[1]
CEN Ting-ting[1],HAN Jian-min[1,2],WANG Ji-yi[1],LI Xi-yu[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 计算机系,上海 200237
1.Mathematics,Physics and Information Engineering College,Zhejiang Normal University,Jinhua,Zhejiang 321004,China
2.Department of Computer Science and Engineering,East China University of Science and Technology,Shanghai 200237,China
E-mail:02190202@zjnu.net

CEN Ting-ting,HAN Jian-min,WANG Ji-yi,et al.Survey of K-anonymity research on privacy preservation.Computer Engineering and Applications,2008,44(4):130-134.

Abstract: K-anonymity is a highlighted topic of privacy preservation research in recent years.In this paper,the concepts of K-anonymity and K-Minimal anonymity are described.Then,generalization & suppression,K-anonymity evaluation criterion,and many different algorithms proposed previously are presented.Finally,the future directions in this field are discussed.

Key words: K-anonymity;privacy preservation;generalization & suppression;data mining;evaluation

摘 要:K-匿名是近年来隐私保护研究的热点,介绍了K-匿名、K-最小匿名化的基本概念,阐述了泛化与隐匿技术,总结了K-匿名的评估标准,并分析了现有的K-匿名算法。最后对该领域的发展方向作了展望。

关键词:K-匿名;隐私保护;泛化和隐匿;数据挖掘;评估
文章编号:1002-8331(2008)04-0130-05 **文献标识码**:A **中图分类号**:TP309

1 引言

随着网络信息技术的发展，人类大量信息被政府部门、商业机构等存储、发布,这极大激发了各部门从海量数据中挖掘有用信息的需求。但事物往往具有两面性,当数据挖掘用于公开分析大量的私人信息(如购物习惯、犯罪记录、病史、信用记录等)时,它在为人们提供强大知识发现功能的同时,也对个人隐私带来威胁。目前,解决该问题的主要方法有:(1)匿名保护,有的机构为了保护个人信息,常常对姓名、社会保险代码等能清楚标识用户信息的显式标识符进行删除或加密,但攻击者往往通过发布数据中的其他信息如种族、生日、性别和邮编等和从其他渠道获得的数据进行链接,推演出隐私数据。(2)在数据清理时对初始数据进行扭曲、扰乱、随机化后再挖掘,这类方法尽管能保持结果的整体统计特性,但通常是以破坏数据的完整性和真实性为代价[5,21]。(3)基于密码学的隐私保护技术。有些数据挖掘算法采用密码学技术解决实际的隐私问题。如安全两方或多方计算问题[3]等,该方法需要过多的计算资源。

为了解决以上3种方法的不足,1998年Samarati P和Sweeney L[15,16]提出K-匿名技术,它要求公布后的数据中存在一定数量的不可区分的个体,使攻击者不能判别出隐私信息所属的具体个体,从而防止了个人隐私的泄密。K-匿名自提出以来,得到了学术界的普遍关注,国内外很多学者都从不同层面研究和发展了该技术。2001年,Samarati P[15]采用泛化和隐匿技术实现K-匿名来保护个体隐私信息,并给出了K-最小匿名化定义;IBM Waston实验室的Iyengar V[17]提出基于遗传算法的不完全随机搜索方法,解决了K-匿名中的数据挖掘分类问题;Yao C等[19]给出视图中的K-匿名验证问题;Machanavajjhala A等[14]的*l*-多样性算法是K-匿名的改进模型,它保留了数据表中足够多的信息,但不足之处是该算法只适用于单个个体对应单条元组的数据表,对单个个体对应多条元组的数据表的隐私信息不能保证,而且该算法一次只能处理一个敏感属性;于戈等[2]对单一约束K-匿名化方法进行扩展,提出了多约束K-匿名方法Classfly+,Classfly+继承了Classfly的元组泛化思想,减少了信息损失,保证了匿名精度。在2006年SIGMOD会议上,XiaoKui Xiao等[18]提出了基于人性化匿名的新泛化框架,该方法实现了满足个体需求的最小泛化，保留了原数据中大量信息。Aggarwal[1]讨论了高维情况下K-匿名的灾难。

本文对近几年K-匿名的工作进行总结和分析，并对将来K-匿名的发展方向作了展望。

2 K-匿名的概念

显示标识符指能清楚标识用户信息的属性,如用户身份证号码、社会保险号、姓名等,在隐私表中删除显示标识符可以在某种程度上达到保护个人隐私的目的。但事实上,原始数据中往往还包含邮编、性别、生日等非显示标识符,攻击者可将非显示标识符和其他渠道获得的数据进行链接，识别出主体身份。据统计,美国约87%的人口可通过邮编、性别、生日等非显示标

作者简介:岑婷婷(1984-),女,研究生,主要研究领域为信息安全;韩建民(1969-),男,博士生,副教授,主要研究领域为信息安全;王基一(1953-),男,教授,硕士生导师,主要研究领域为智能计算;李细雨(1985-),男,研究生,主要研究领域为信息安全。

收稿日期:2007-05-30 **修回日期**:2007-08-09

识符来唯一确定其个体身份[10]。

例如表 1,隐私表 PT 中,通过删除姓名和社会保险代码,数据被初步匿名化。但隐私表中还包含 Race、Birth、Sex、Zip、Marital status 等属性,这些属性能够被链接到表 2 的投票清单从而泄露个体的姓名、地址等隐私信息。如某一个体是离异女性,出生于 64/04/12,居住在 94121 区。如果该记录是唯一的,通过对表 1、表 2 的链接,易推得"*Sue*,900 *Market Street*,*San Francisco*,64/04/12,*F*,94142,*divorced*,*hypertension*",人们就可以知道该女性是Sue,得了高血压,但 Sue 却并不希望她得高血压的事情被泄露。

为了解决该问题,人们提出了 K-匿名模型。

表 1 "匿名化"隐私表 PT(医疗信息)

SSN	Name	Race	Birth	Sex	Zip	Marital	Disease
		asian	64/04/12	F	94142	divorced	hypertension
		asian	64/09/13	F	94141	divorced	obesity
		asian	64/04/15	F	94139	married	chest pain
		asian	63/03/13	M	94139	married	obesity
		asian	63/03/18	M	94139	married	short breath
		black	64/09/27	F	94138	single	short breath
		black	64/09/27	F	94139	single	obesity
		white	64/09/27	F	94139	single	chest pain
		white	64/09/27	F	94141	window	short breath

表 2 非"匿名化"的公开数据表

Name	Address	City	Zip	Dob	Sex	Status
……	……	……	……	……	……	……
……	……	……	……	……	……	……
Sue	900 Market st.	San Francisco	94142	64/04/12	F	divorced
……	……	……	……	……	……	……

表 3 K-匿名实例表,K=2,QI={*Race*,*Birth*,*Sex*,*Zip*,*Marital status*}

Num	Race	Birth	Sex	Zip	Marital	Disease
t1	asian	64/0*/**	F	941**	been-married	hypertension
t2	asian	64/0*/**	F	941**	been-married	obesity
t3	asian	64/0*/**	F	941**	been-married	chest pain
t4	asian	63/0*/**	M	941**	been-married	obesity
t5	asian	63/0*/**	M	941**	been-married	short breath
t6	black	64/0*/**	F	941**	never-married	short breath
t7	black	64/0*/**	F	941**	never-married	obesity

2.1 K-匿名定义

定义 1(准标识符) 给定实体集 U、实体表 $T(A_1,A_2,\cdots,A_n)$,$f_c:U\to T$ 以及 $f_g:T\to U'$,其中 $U\subseteq U'$。表 T 的准标识符 QI(Quasi Identifier)[9]为属性组$(A_i,\cdots,A_j)$,$(A_i,\cdots,A_j)\subseteq(A_1,A_2,\cdots,A_n)$,其中 $\exists p_i\in U$ 且满足 $f_g(f_c(p_i)[QI])=p_i$。

换言之,同时存在于隐私表与外表中,利用推演来标识个体信息的一组属性称作准标识符,如属性组{*Race*,*Birth*,*Gender*,*ZIP*}。准标识符的定义依赖于攻击者所获得的外表信息,即外表的关联属性,同一个隐私表对于不同的外表,可能存在不同的准标识符。为了简化问题,本文假设隐私表中的单个个体对应单条元组,并假定给出的准标识符是被最优定义的。

定义 2(K-匿名) 给定数据表 $T(A_1,A_2,\cdots,A_n)$,QI 是与 T 相关联的准标识符,当且仅当在 $T[QI]$中出现的每个值序列至少要在 $T[QI]$中出现 K 次,则 T 满足 K-匿名。$T[QI]$表示 T 表的元组在准标识符 QI 上的投影。

该定义满足 K-匿名的充分条件,如果外表的一系列属性组出现在与 T 表相关联的准标识符中(假定 T 表满足 K-匿名的定义),则外表与 T 表的链接就不会允许攻击者推导出任何少于 K 个个体的元组。如表 1,准标识符为 {*Race*,*Date of birth*,*Sex*,*Zip*,*Marital status*},该表满足 K-匿名当且仅当 K=1。表 3 是表 1 进一步匿名化后发布的数据表,准标识符同上,K=2。根据定义 2 得,$T[QI]$中出现的任一有序元组值在 $T[QI]$中重复两次以上,$t_1[QI]=t_2[QI]=t_3[QI]$,$t_4[QI]=t_5[QI]$,$t_6[QI]=t_7[QI]$。

推论 给定数据表 $T(A_1,A_2,\cdots,A_n)$,$QI(A_i,\cdots,A_j)$是与 T 关联的准标识符,$A_i,\cdots,A_j\in A_1,A_2,\cdots,A_n$,若 T 满足 K-匿名,则在 $T[A_x]$中出现的每一个值序列至少要在 $T[QI]$中出现 K 次,其中 $x=i,\cdots,j$。

表 3 满足 K-匿名要求,准标识符为{*Race*,*Date of birth*,*Sex*,*Zip*,*Marital status*},K=2。根据推论可得,表 3 中与准标识符 QI 某一属性相关联的值至少出现 2 次。$|T[Race=\text{"}asian\text{"}]|=5|$,$|T[Race=\text{"}black\text{"}]|=2$,$|T[Birth=\text{"}1964\text{"}]|=5$,$|T[Birth=\text{"}1963\text{"}]|=2$,$|T[Sex=\text{"}F\text{"}]|=5$,$|T[Sex=\text{"}M\text{"}]|=2$,$|T[Zip=\text{"}941**\text{"}]|=7$,$|T[Marital\ status=\text{"}been\text{-}married\text{"}]|=5$,$|T[Marital\ status=\text{"}never\text{-}married\text{"}]|=2$。

易证,如果发布的数据表 T 满足 K-匿名,那么以准标识符 QI 为基础的外部源与表 T 的结合,不能够联接到 QI 或 QI 任一子集上,来匹配出少于 K 个数目的个体。该特性能够预防表中信息与外部源的链接推理。

2.2 泛化和隐匿

K-匿名主要通过泛化和隐匿技术实现,这两种技术不同于原先提到的扭曲、扰乱、随机化等方法,它们能保持数据的真实性。在典型的关系型数据库系统中,经常用域来描述属性值的集合,比如,邮政编码域、数值域、时间域等。泛化主要涉及两个概念:域泛化和值泛化。域泛化通常是将一个给定的属性值集合概括成一般值集合,比如原始邮政编码域{94138,94139,94141,94142},通过去掉最右数字,泛化成{9413*,9414*},使得语义上指示了一个较大范围,该范围称为泛化域(Dom),泛化域包含各自原先的泛化值,并且两者之间存在一一对应关系,标记为 $\leqslant_D$。

给定隐私表 T 关于属性 A 的两个域 D_i、D_j,$D_i,D_j\in Dom$,$D_i\leqslant_D D_j$ 表明域 D_j 中的值是域 D_i 中的泛化值。泛化关系 $\leqslant_D$ 在一系列域上定义了一个偏序关系,该关系要求满足以下两个条件:

C_1:$\forall D_i,D_j,D_z\in Dom$;$D_i\leqslant_D D_j$,$D_i\leqslant_D D_z\Rightarrow D_j\leqslant_D D_z\vee D_z\leqslant_D D_j$。

C_2:所有 Dom 的最大元素都是单个值。

条件 C_1 说明,对于任意域 D_i,其域泛化集是有序的,故域 D_i 至多只有一个直接泛化域 D_j。条件 C_2 保证每个域中的所有值总能被泛化到单个值。泛化关系的定义决定了域泛化层的存在,标记为 DGH_D。

值泛化关系,标记为 $\leqslant_V$,它对应于域 D_i 中每个值直接泛化成域 D_j 中的唯一值。值泛化关系同样决定了值泛化层的存在,标记为 VGH_D。值泛化层 VGH_D 可以以树形结构表示,其叶子节点是域 Dom 中的值,根节点为 DGH_D 中的最大元素。图 1 说明了 ZIP 域的域泛化层、值泛化层以及两者之间的包含关系。ZIP 泛化关系指定为泛化 5 位有效数字的邮政编码,先泛化成 4 位有效数字,之后进一步泛化到 3 位有效数字。其他属性泛化构建同理。

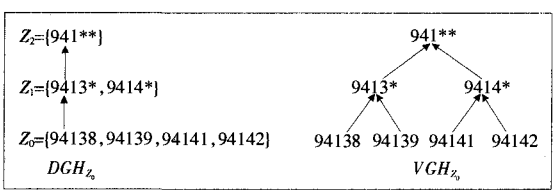


图 1 ZIP 泛化层

给定数据表 T 的属性域集合 $DT=\{D_1,\cdots,D_n\}$,D_i 为属性 i 的泛化域,$i=1,\cdots,n$,DT 对应的域泛化层是按属性排列顺序形成的笛卡尔积 $DGH_{DT}=DGH_{D_1}*\cdots*DGH_{D_n}$,该表可能存在的泛化数目为:

$$\prod_{i=1}^{n}(|DGH_{D_i}|+1) \tag{1}$$

在满足 K-匿名约束情况下,数据泛化后还存在一定有限数量的元组不符合 K-匿名,如进一步泛化则会导致信息大量丢失,故一般采用隐匿来减轻泛化程度。如表 3 通过隐匿元组 $t8$、$t9$ 来达到 2-匿名的要求。

2.3 K-最小匿名化定义

并不是所有的泛化和隐匿都满足用户需求的,不同程度的泛化和隐匿会导致不同程度的数据损失,信息损失越大,匿名化后数据的实用性越小。所以隐私表的匿名化一般都要求 K-最小泛化隐匿。

定义 3(K-最小匿名化) 给定数据表 T_i、T_j,T_j 为 T_i 的匿名表,标记为 $T_i\leqslant T_j$,$MaxSup$ 为可以接受的最大隐匿数。T_j 是数据表 T_i 的最小匿名表,当且仅当满足:(1)T_j 满足 K-匿名,且其隐匿的元组数要少于 $MaxSup$,即 $|T_i|-|T_j|\leqslant MaxSup$。(2)不存在泛化表 T_z,T_z 满足 K-匿名,隐匿数少于 $MaxSup$,且泛化的程度小 T_j。

表 4 是 K-最小匿名化的实例,准标识符 $QI\{Race,Zip\}$,$K=2$。根据定义可得,GT_1、GT_2 均是 T 的 K-最小匿名表。

表 4 隐私表 T 和其 2-最小泛化,假设 $MaxSup=2$

Race	Zip	Race	Zip	Race	Zip
asian	94142	***	***	asian	9414*
asian	94141	person	94141	asian	9414*
asian	94139	person	94139	asian	9413*
asian	94139	person	94139	asian	9413*
asian	94139	person	94139	asian	9413*
black	94138	***	***	black	9413*
black	94139	person	94139	black	9413*
white	94139	person	94139	***	***
white	94141	person	94141	***	***
T		GT_1		GT_2	

2.4 K-匿名的评估

K-匿名的评估主要是建立合适的评价标准及相关的参考标准。在实际的应用中,不可能用一个标准衡量所有的 K-匿名算法,只能说某一算法可能在某一个特定的应用中,在某个标准上优于其它算法。因而,应该向用户提供一套度量准则,让用户根据自身的需求选择合适的 K-匿名算法。通常,K-匿名的评价指标为:性能、可测量性、数据实用性、不确定性水平、隐匿失败率、匿名程度和耐久性。

评价性能的方法是估计该算法的时间复杂性或算法中基本操作的平均次数。可测量性用于评价算法在数据容量增大时的效率趋势,它要求 K-匿名算法对大型甚至超大型的数据库而言是可升级的。数据实用性指数据 K-匿名化后的信息丢失量,它主要通过精确度来体现。

定义 4(K-匿名精度) 给定 $T(A_1,A_2,\cdots,A_{N_a})$,表中含 N_a 个属性,N 条记录。RT 为 T 的匿名表,DGH_A 表示属性 A 的域泛化层,则基于匿名表 RT 的 K-匿名精度 $Pre(RT)$ 为:

$$Quality(RT)=1-\frac{\sum_{i=N_a}^{i=N_a}\sum_{j=1}^{j=N}\frac{h}{DGH_{A_i}}}{|T|*|N_a|} \tag{2}$$

式中:h 表示属性 A 的域泛化层的高度,h/DGH_{A_i} 指每个数据单元匿名化后的信息丢失量。如果 $T=RT$,即给定表 T 等于匿名表 RT,则 $h=0$,$Quality(RT)=1$。反之,如果 RT 中的值为域泛化层中的最大值,则 $h=|DGH_{A_i}|$,$Quality(RT)=0$。

尽管不同的 K-匿名方法可以对信息进行匿名隐藏,但由于不确定性的存在,一些隐藏的信息仍然可以被推理出来。隐匿失败率指匿名化后敏感信息存在的比率,大部分算法的设计目标是为了获得零失败率,于是隐藏了尽量多的敏感信息,但随着匿名化程度升高,非敏感数据的丢失量也随之变大。所以目前的算法一般设定匿名程度来达到隐私和知识发现之间的平衡。

匿名程度也是评估 K-匿名算法的重要参数,针对数据挖掘的分类问题,Shannon 提出了信息熵的概念来对匿名程度进行量化分析:X 为随机变量,$P(x)$为任意实数 x 属于某一类的分布概率。则 X 的信息熵为:

$$h(X)=-\sum p(x)\mathrm{lb}(p(x))\text{或}\ h(X)=-\int f(x)\mathrm{lb}(f(x))\mathrm{d}x \tag{3}$$

式中 $f(x)$表示连续随机变量的密度函数。

信息熵表示数据的平均信息容量,熵值越小,子集划分的纯度越高,故匿名化后的信息熵值要大于匿名化之前。信息增益 $Gain$ 是由于知道随机变量 X 的值而导致的熵的期望压缩,它反映了划分的最小随机性或"不纯性"。Ke Wang[8]等定义了属性域中的 c 值属于 cls 类的信息熵公式为:

$$Info(T_c)=-\sum_{cls}\frac{freq(T_c,cls)}{|T_c|}*\mathrm{lb}\frac{freq(T_c,cls)}{|T_c|} \tag{4}$$

式中 c 为某属性的具体值,$freq(T_c,cls)$表示分类结果属于 cls 并且具有属性值 c 的元组个数。

其信息丢失公式为:

$$I(G)=info(RT_p)-\sum_c\frac{T_c}{RT_p}Info(T_c) \tag{5}$$

式中 RT 表示表 T 的匿名表,p 是属性值 c 的泛化值。

其最小匿名程度为:

$$IP(G)=I(G)/P(G) \tag{6}$$

$$P(G)=x-T[QI] \tag{7}$$

当 $RT[QI]\leqslant K$ 时,$x=RT[QI]$,否则 $x=K$,$T[QI]$表示 T 表的元组在准标识符 QI 上的投影。

K-匿名算法的最终目的是反对链接攻击产生的隐私泄露,攻击者往往利用各种数据挖掘工具威胁隐私,一个针对具体挖掘技术而设计的 K-匿名算法是不可能适用所有其他的挖掘算法,所以耐久性也是评估该类算法的重要参数。

3 K-匿名算法

目前已出现很多算法来实现 K-匿名,本文从不同的角度对其进行划分和分析:

(1)从泛化约束角度,可将 K-匿名算法划分成两类:"全域"泛化(full-domain generalization)[15]和"全子树编码"(full-subtree recoding)。"全域"泛化指准标识符中的整列属性均同时被泛化成一般域,它要求给定属性的所有值都属于同一个域,并且都被泛化为 DGH_D 中的值。如图 1,邮政编码 94141、94142 属于同一个域,并都泛化成 9414*。Samarati P[15]设计了一种寻找最小全域泛化的二叉搜索算法。文献[9]描述了实现全域泛化的贪婪启发式算法 Datafly,尽管其泛化结果能达到 K-匿名,但它并不能保证得到的是最小匿名化表。文献[13]提出了一种高效率全域 K-匿名化方法,减少了执行时间,但文献没有针对信息损失给出有效的解决方案。简而言之,"全域"泛化

常常导致不必要的信息丢失。“全子树编码”又称局部泛化(Local recoding)指在域泛化层的某一子树中搜索局部最小匿名,即仅泛化属性列中的某些单元,并非整个属性列。文献[17]采用遗传算法搜索局部最小匿名,它使用单维全子树编码模型和单维有序集划分模型分别来泛化描述型属性(如性别等)和数值型属性(如年龄等)。它将匿名表标准化后的个体惩罚总值定义为分类机制 CM:

$$CM=\frac{\sum_{all\ rows} penalty(row\ r)}{N} \tag{8}$$

式中:当元组 r 被隐匿或错误分类时,$penalty(row\ r)=1$,否则 $penalty(row\ r)=0$。

(2)从数据分布方式角度,可分为分布式数据匿名化和集中式数据匿名化。大多数 K-匿名算法主要基于集中式数据库进行设计。然而,当在双方或多方合作进行数据挖掘时,由于某种原因,参与者往往不愿将数据与他人共享,而只愿共享数据挖掘的结果。这就要求人们对此类数据库的匿名化提出相应的算法。目前针对分布式数据,主要的算法有:Jiang[7]提出了一种通信协议,数据所有者遵循该协议合并分布式的数据来获得匿名表。Zhong[20]等通过构建一个横向划分表来取代分布式 K-匿名方法。

(3)从维数角度,可分为单维 K-匿名和多维 K-匿名两大类。单维 K-匿名指每个值只有一种泛化类型。如图 3,邮政编码 94142 只泛化为 9414*,但事实上邮政编码 94142 还可以泛化成 941*2 或 94*42。如果假定泛化层不是一颗树而是一个图,那么该泛化问题将变得更加复杂。为了解决该问题,LeFevre[11,12]等提出了 K-匿名的多维模型,他指出这类泛化问题属于 NP 难题,提出了一种贪婪近似值算法,其时间复杂性为 $O(n\log n)$,n 为隐私表中的元组个数。

表 5 2 匿名化的另一实例表 $K=2$,$QI=\{Race, Birth, Sex, Zip, Marital\ status\}$

Num	Race	Birth	Sex	Zip	Marital	Disease
t1	asian	64	F	941**	divorced	hypertension
t2	asian	64	F	941**	divorced	obesity
t3	asian	64	F	941**	divorced	chest pain
t4	asian	63	M	941**	married	obesity
t5	asian	63	M	941**	married	obesity
t6	black	64	F	941**	single	short breath
t7	black	64	F	941**	single	short breath
t8	white	64	F	941**	single	chest pain
t9	white	64	F	941**	single	short breath

(4)从推理攻击角度,Machanavajjhala A[14]等描述了两种可能的推理攻击,即同质推理攻击和背景知识推理攻击。同质推理攻击指攻击者根据 K-匿名表能 100%地推导出某个体的敏感信息:假设攻击者试图推导 Carol 的疾病情况,已知 Carol 是一位黑人女性,邮政编码 94141,从公开的数据表 5 中,可以推导出 Coral 对应 t_6 或 t_7,但该两条记录都与 short breath 相关,故攻击者可以很确定的得到 Coral 患上了 short breath。因此该匿名表是不安全的。

RT 表示隐私表 T 基于准标识符 QI 的匿名表,表 T 中的非敏感属性 A 含某具体值 c,敏感属性 S 含某具体值 s,c^* 为 RT 表中属性值 c 的泛化值。$n(c^*,s')$ 表示元组 t^* 的数目,其中 $t^*\in RT$,$t^*[QI]=c^*$,$t^*[S]=s'$。$f(s'|c^*)$ 表示在非敏感属性 A 泛化成 c^* 成立的条件下,敏感属性值 s' 的条件概率。同质推理攻击的判定参数 $\beta(c,s,RT)$ 为:

$$\beta(c,s,RT)=\frac{n(c^*,s)\dfrac{f(s|c)}{f(s|c^*)}}{\sum_{s'\in S} n(c^*,s')\dfrac{f(s'|c)}{f(s'|c^*)}} \tag{9}$$

背景知识推理攻击指攻击者利用预先知道的某些额外信息来进行攻击。比如攻击者知道 Hellen 是白人女性,通过表 5,可以推导出 Hellen 可能患了 chest pain 或 short breath,假设攻击者知道 Hellen 每天跑步两小时,如果 Hellen 患的是 short breath,那她不可能长时间跑步,显而易见,Hellen 患了 chest pain。

为了防止此类攻击,Machanavajjhala A 等[14]提出了 l-多样性的概念:给定隐私表 T 和对应的泛化表 RT,将 RT 中基于相同准标识符的元组集组成一个 q-块,如果每个 q-块中的敏感属性至少包含 l 个不同的值,则称该 q-块是 l-多样性的,该算法形成了具有 l-多样性的 K-匿名表。

(5)从具体算法的研究方法来看,K-匿名除了对泛化和隐匿技术进行改进外,还与其它方法相混合,如 SDC(Statistical Disclosure Control)。Domingo-Ferrer 等[5]提出了微聚类法(Microaggregation),该方法将类的范围设定在[K,2K]之间,要求将数据集划分成多个类,每个类中至少包含 K 个元组。最优 K-划分准则为:在同一类中的元组之间尽可能“接近”或相关,而不同类中的元组之间尽可能“远离”或不同,使得信息的丢失量会尽可能的少。最优 K-划分的目标是使得 SST 最小化:

$$SST=\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})'(x_{ij}-\bar{x}) \tag{10}$$

式中:隐私表含 p 个属性,n 条元组。划分 n 条元组为 g 个类,每个类含 n_i 条记录,x_{ij} 表示第 i 个类中的第 j 个属性值,$\bar{x}$ 表示 n 个元组的平均属性值

4 K-匿名发展趋势与展望

K-匿名经过多年的研究,逐渐成熟和完善。虽然各种方法还存在一定的缺陷,但随着一些关键问题的解决,K-匿名将越来越多地应用到各个领域。今后的主要的发展方向有以下几个方面:基于微聚类的 K-匿名算法的研究;同质推理攻击和背景知识推理攻击的研究;多维度情况下 K-匿名的研究;K-匿名局部隐私保护和 K-匿名通讯协议的研究:如 Bettini 设计了评价用户身份隐私的框架,它保证了一组个体能在相同的时空背景下发送消息;Aggarwal 指出,当准标识符中的属性个数增加时,K-匿名表的信息丢失量也随之增加。如何确定最小的准标识,如何在数据满足 K-匿名约束的同时使得数据精度最高,也是今后 K-匿名研究的新方向;K-匿名技术广泛应用于集中式数据,而对不同连接团体之间的分布式数据研究甚少,故分布式数据匿名将会引起人们的注意;目前存在的 K-匿名主要研究对所有个体进行相同程度保护的通用方法,并没有关注个体自身的需求,致使该类方法对某些个体提供的保护不足,而对另一些个体应用的隐私控制却过多,所以人性化的 K-匿名也将是今后研究的热点。

参考文献:

[1] Aggarwal C.On k-anonymity and the curse of dimensionality[C]// Proc of the 31st International Conference on Very Large Data Bases,Trondheim,Norway,2005:901-909.

[2] 杨晓春,刘向宇,王斌,等.支持多约束的 K-匿名化方法[J].软件学报,2006:1222-1231.

[3] Du Wenliang,Attalah M J.Secure multi-problem computation problems and their applications:A review and open problems,CERIAS Tech Report 2001-51[R].Center for Education and Research in Information Assurance and Security and Department of Computer Sciences,Purdue University,West Lafayette,IN 47906,2001.

[4] Bettini C,Wang X S,Jajodia S.Protecting privacy against location based personal identification[C]//Proc of the Secure Data Management,Trondheim,Norway,2005:185-199.

[5] Domingo-Ferrer J,Mateo-Sanz J M.Practical data-oriented microaggregation for statistical disclosure control[C]//IEEE Transactions on Knowledge and Data Engineering,2002:189-201.

[6] Kargupta H,Datta S,Wang Qi,et al.On the privacy preserving properties of random data perturbation techniques[C]//Proc of ICDM'03,Washington,DC,USA,IEEE Computer Society,2003:99.

[7] Jiang W,Clifton C.Privacy-preserving distributed k-anonymity[C]//Proc of the 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security,Storrs,CT,USA,2005.

[8] Wang Ke,Yu P,Chakraborty S.Bottom-up generalization:a data mining solution to privacy protection[C]//Proc of the 4th International Conference on Data Mining,Brighton,UK,2004:249-256.

[9] Sweeney L.Achieving k-anonymity privacy protection using generalization and suppression[J].International Journal on Uncertainty,Fuzziness and Knowledge-based Systems,2002:571-588.

[10] Sweeney L.k-Anonymity:a model for protecting privacy[J].International Journal of Uncertainty,Fuzziness and Knowledge-Based Systems,2002,10(7):557-570.

[11] LeFevre K,DeWitt D J,Ramakrishnan R.Multidimensional k-anonymity,Technical Report 1521[R].Department of Computer Sciences,University of Wisconsin,Madison,USA,2005.

[12] LeFevre K,DeWitt D J,Ramakrishnan R.Mondrian multidimensional k-anonymity[C]//Proc of the International Conference on Data Engineering(ICDE'06),Atlanta,GA,USA,2006.

[13] LeFevre K,DeWitt D,Ramakrishnan R.Incognito:Efficient full-domain k-anonymity[C]//Proc of the SIGMOD'05 on Management of Data.New York:ACM,2005:49-60.

[14] Machanavajjhala A,Gehrke J,Kifer D.l-diversity:Privacy beyond k-anonymity[C]//Proc of the International Conference on Data Engineering(ICDE'06),Atlanta,GA,USA,2006:24.

[15] Samarati P.Protecting respondents' identities in microdata release[C]//Proc of the TKDE'01,2001:1010-1027.

[16] Samarati P,Sweeney L.Generalizing data to provide anonymity when disclosing information (Abstract)[C]//Proc of the 17th ACM-SIGMODSIGACT-SIGART Symposium on the Principles of Database Systems,Seattle,WA,USA,1998:188.

[17] Iyengar V.Transforming data to satisfy privacy constraints[C]//SIGKDD,2002:279-288.

[18] Xiao Xiaokui,Tao Yufen.Personalized privacy preservation[C]//SIGMOD,Chicago,Illinois,USA,2006:229-240.

[19] Yao C,Wang X S,Jajodia S.Checking for k-anonymity violation by views[C]//Proc of the 31st Int'l Conf on Very Large Data Bases.Trondheim:ACM,2005:910-921.

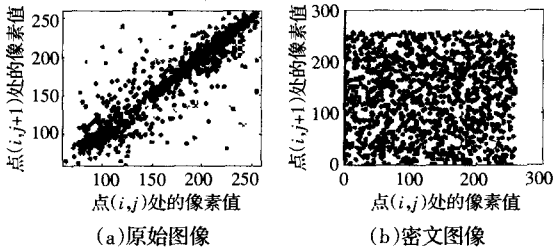
[20] Zhong Sheng,Yang Zhiqiang,Wright R N.Privacy-enhancing k-anonymization of customer data[C]//Proc of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems,Baltimore,Maryland,USA,2005:139-147.

[21] Huang Zhengli,Du Wenliang,Chen Biao.Deriving private information from randomized data[C]//Proc of ACM SIGMOD'05,Baltimore,Maryland,2005:37-48.

(上接126页)

密钥具有高度的敏感性。

(4)相邻像素点的相关性分析



(a)原始图像 (b)密文图像

图4 图像水平方向相邻像素点的相关性

图4描述了水平方向明文和密文相邻像素的相关性。由结果可知,原始明文图像的相邻像素是高度相关的,相关系数接近于1,而加密图像的相邻像素相关系数接近于0,相邻像素已基本不相关,说明明文的统计特征已被扩散到随机的密文中。

5 结束语

本文利用统一混沌系统模型,提出了基于统一非线性混沌系统同时实现图像像素置换和替代加密的新算法。算法具有以下主要优点:(1)像素的位置置换和像素值的替代均基于复杂非线性高维混沌系统,克服了低维混沌系统不能抵御相空间重构攻击的缺点。(2)以三维统一混沌系统的系统参数和初值为密钥,及引入了辅助密钥,大大拓展了密钥空间,使算法具有抵御穷举攻击的能力。(3)统一混沌系统具有复杂的非线性混沌行为,因此生成的密钥具有较高的复杂性;且每次随机产生的密钥不同,具有一次一密特性,导致密文具有在整个取值空间均匀分布的特性,相邻像素具有近似于零的相关性。

混沌加密和解密的速度是较快的;但由于求解高维混沌系统微分方程的时间开销较大。如果将其用于对网络通信的实时图像加密,可以离线生成混沌序列,然后进行在线加密。这样,算法既具有图像加密的实时性,又确保了加密效果的高安全性。

参考文献:

[1] Cheng Howard,Li Xiaobo.Partial encryption of compressed images and videos[J].IEEE Transcations on Signal Processing,2000,48(8):2439-2551.

[2] Dang P P,Chan P M.Image encrption for secure Internet multimedia applications[J].IEEE Transcations on Consumer Electronics,2000,46(3):395-403.

[3] 龙翔.Liuux中检查点(Checkpoint)的核心支持:ckpt文件系统的设计[J].计算机工程与应用,2002,38(6):120-122.

[4] 孙克辉,刘巍,张泰山.一种混沌加密算法的实现[J].计算机应用,2003,23(1):15-17.

[5] Scnarmger J.Fast encryption of image data using chaotic:Kolomogorov flows[C]//Proceedings of the Symposium on Electronic Imaging:Science and Technology_Storage and Retrieval for Image and Video Dtabase V,San Jose,California,1997,3022:278-289.

[6] Yang T,Yang L B,Yang C M.Application of neural networks to unmasking chaotic secure communication[J].Phys D,1998:124:248.

[7] Lv Jin-hu,Chen Guan-rong,Zhang Suo-chun.The compound structure of a new chaotic attractor[J].Chaos,Solitons and Fractals,2002,14(5):669-672.