doi:10.3772/j.issn.1002-0470.2015.03.002

FVS k-匿名:一种基于 k-匿名的隐私保护方法①

王 良② 王伟平 孟 丹

(信息内容安全技术国家工程实验室 中国科学院信息工程研究所 北京 100093)

摘 要 为了确保数据发布应用环节中个人敏感隐私数据信息的安全,深入研究了 k-匿名技术的机制及性能,针对其不能完全有效地防止敏感属性数据信息泄漏的问题,通过引入真子树的概念和全新的敏感属性值选择手段,在实验探索的基础上,提出了一种基于 k-匿名隐私保护模型的新的数据发布隐私保护方法——FVS k-匿名隐私保护方法。这种隐私保护方法继承了 k-匿名技术实现简单、处理数据便捷的优点,而且弥补了其保护个人敏感隐私数据信息不完全、不充分的缺点。优化后的 FVS k-匿名方法能有效地防止个人敏感隐私数据信息的泄漏,确保个人敏感隐私数据信息的安全。

关键词 k-匿名,信息安全,隐私保护,敏感属性

0 引 言

当前,许多特定的组织机构需要将组织机构内的原始数据(如医院医疗数据、民意调查数据等)发布出去,以供其它组织机构或科研团体进行研究分析(或进行其它目的的应用)。发布出去的原始数据集中可能会包含着敏感的个人隐私信息(如疾病、收入、存款等)。如果攻击者利用数据组织机构发布的数据集与其它渠道获取的数据集进行相互连接,那么就能准确地推断出某些特定个体的敏感隐私信息,Sweeney 将这种攻击方式定义为连接攻击[1]。到目前为止,连接攻击已经成为非法获取个人敏感隐私数据信息最常见和最通用的方法。

为了有效地防止发布的数据集中敏感隐私数据信息的泄漏,数据发布组织机构需要将发布数据集中能够唯一准确识别出个体的标识信息(如姓名、身份证号码、地址信息等)移除。但是经过移除标识信息处理的发布数据集,并不能完全有效地防止敏感隐私数据信息的泄漏。攻击者仍可以利用发布数据集中包含的出生日期、性别、邮政编码等属性信息,通过连接攻击方式获取某些特定个体的敏感隐私数据信息。针对上述情况,Sweeney 和 Samarati 等人提出了 k-匿名隐私保护模型[1-3]。k-匿名隐私保护模型能够避免连接攻击发生,对隐私数据信息起到有效的防护作用,但是对于敏感属性信息并没有采取有效的防护手段,仍然存在着隐私数据信息泄漏的风险。在发生同质攻击[4]、背景知识攻击[5]、相似性攻击[6]等情况下,k-匿名隐私保护模型并不能有效地保护敏感属性信息的安全。针对敏感属性信息易引起隐私泄漏的问题,我们提出了一种新的隐私保护方法,并将其命名为 FVS k-匿名隐私保护方法,这种新的隐私保护方法建立在 k-匿名隐私保护模型的基础上,继承了 k-匿名隐私保护模型的优点,又弥补了在保护敏感属性信息方面的不足,能有效地防止常见的典型攻击案例的发生。

1 k-匿名隐私保护模型

1.1 k-匿名(k-anonymity)

定义 1 k-匿名[1](k-anonymity):给定一个数据

① 863 计划(2013AA013204),中国科学院先导专项(XDA06030200),核高基项目(Y3M001105)和新疆维吾尔自治区科技专项(201230121)资助项目。

② 男,1975 年生,博士生;研究方向:数据和隐私安全保护研究;联系人,E-mail:wangliang@iie.ac.cn

(收稿日期:2014-09-11)

表 T(A_1,…,A_n) 及与其相关联的准标识符 QI_T = (A_i,…,A_j)(A_i,…,$A_j \subseteq A_1$,…,A_n),如果表 T 满足 k 匿名,当且仅当 T[QI_T]中的每一个元组至少在 T[QI_T]中出现 k 次。

例 1　我们通过一个具体的实例来说明 k-匿名隐私保护模型。表 1 是要进行数据发布的原始病例数据记录表。表中包含 9 个元组,每个元组对应一条具体的个人病例诊疗结果记录信息。表中第一列为序号字段,表示当前元组在数据表中的相对存储位置;第二列为病人姓名(Name)属性信息,是显示标识符;第三列为病人年龄(Age)属性信息;第四列为病人居住地的邮政编码(Zip Code)属性信息,准标识符 QI_T = {Age,ZIP Code};第五列为疾病(Disease),是敏感属性信息,表示病人的医疗诊断结果。表 2 是原始病例表表 1 经过 3-匿名化处理后的数据结果发布表。根据等价类的定义,表 2 中一共有 3 个等价类,等价类 1 = {R_1,R_2,R_3}、等价类 2 = {R_4,R_5,R_6}、等价类 3 = {R_7,R_8,R_9}。每一个等价类中分别包含着 3 个元组,等价类 1 中的元组 R_1[QI_T] = R_2[QI_T] = R_3[QI_T] = {[0 - 25], 945**},等价类 2 中的元组 R_4[QI_T] = R_5[QI_T] = R_6[QI_T] = {[26 - 35],946**},等价类 3 中的元组 R_7[QI_T] = R_8[QI_T] = R_9[QI_T] = {[36 - 60], 945**}。所以,攻击者利用连接攻击方式获取敏感隐私信息的概率仅为 $1/k$ = 1/3。因此,经过 k-匿名化处理后的数据结果表(表 2)可以有效地防止连接攻击。

表 1　原始病例记录表

序号	姓名	年龄	邮编	疾病
1	Mee	33	94623	Bronchitis
2	Thomas	27	94622	Pneumonia
3	Astin	18	94505	Angina Pectoris
4	Abel	26	94616	Flu
5	Eddy	24	94534	Angina Pectoris
6	Bob	21	94582	Angina Pectoris
7	Sam	37	94508	Angina Pectoris
8	Vega	59	94509	Stomach Cancer
9	Andy	44	94503	Stomach Cancer

表 2　表 1 经过 3-匿名化处理后的数据

序号	年龄	邮编	疾病
1	[0 - 25]	945**	Angina Pectoris
2	[0 - 25]	945**	Angina Pectoris
3	[0 - 25]	945**	Angina Pectoris
4	[26 - 35]	946**	Flu
5	[26 - 35]	946**	Pneumonia
6	[26 - 35]	946**	bronchitis
7	[36 - 60]	945**	Angina Pectoris
8	[36 - 60]	945**	Stomach Cancer
9	[36 - 60]	945**	Stomach Cancer

1.2　k-匿名攻击

k-匿名隐私保护模型虽然能有效地防止连接攻击的发生,却不能完全确保个人敏感隐私信息的安全。我们通过 6 种最典型的攻击案例来说明 k-匿名隐私保护模型固有的缺陷和不足。

(1) 同质攻击[5](homogeneity attack)又称为一致性攻击,是指在经 k-匿名化处理后发布的数据集中,某个等价类所包含元组的敏感属性值都相同的情况下所发生的攻击行为。

例 2 假设 Alice 知道同学 Bob 的医疗记录在表 2 中,并且她也知道 Bob 的年龄在 20 岁到 25 岁之间,居住地的邮政编码为 94582,Alice 根据表 2 中的数据记录信息可以准确地推断出 Bob 的医疗记录在等价类 {R_1,R_2,R_3} 中,又因为这三个元组的敏感属性值相同,都是 Angina Pectoris(心绞痛),所以,Alice 能进一步推断出 Bob 已经得了心脏病。

(2) 背景知识攻击[5](background knowledge attack)是指攻击者对被攻击者的生活和工作背景信息有一定了解的情况下所发生的攻击行为。

例 3 假设 Alice 知道同事 Andy 的医疗记录在表 2 中,并且她也知道 Andy 的年龄在 40 岁到 45 岁之间,居住地的邮政编码为 94503,Alice 根据表 2 中的数据记录信息可以准确地推断出 Andy 的医疗记录在等价类 {R_7,R_8,R_9} 当中,同时,Alice 也从其他同事那里了解到,Andy 和 Andy 家族没有得心脏病的疾病家族史,而且他得心脏病的概率也非常低,所以,Alice 能进一步推断出 Andy 得了 Stomach Cancer(胃癌)。

(3) 相似性攻击[6](similarity attack) 是指在经 k-匿名化处理后发布的数据集中,如果在某一个等价类中,存在所有元组敏感属性值都非常接近(或含义相似或含义相近)的情况下所发生的攻击行为。

例 4 假设 Alice 知道同事 Thomas 的医疗记录在表 2 中,并且她也知道 Thomas 的年龄在 25 岁到 30 岁之间,居住地的邮政编码为 94622。Alice 非常想知道 Thomas 到底得了什么疾病,她通过研究发现,Thomas 的医疗记录一定在等价类 $\{R_4, R_5, R_6\}$ 中,由于 Flu(流行性感冒)、Pneumonia(肺炎) 、bronchitis(支气管炎)这三种疾病同属于呼吸道感染疾病,因此,Alice 能够推断出 Thomas 已经得了呼吸道感染疾病。

(4) 非对称性攻击[6](skewness attack) 是指在经 k-匿名化处理后发布的数据集中,所有等价类中全部元组的敏感属性值稀疏并且分布不均匀的情况下所发生的攻击行为。

例 5 在某种特殊情况下,k-匿名发布的数据集中所有元组的敏感属性值可能只有两个(如在艾滋病群体数据单一发布的数据集中,艾滋病只有阴性和阳性两个敏感属性值),如果两个敏感属性值所对应的元组数目差值巨大可能会造成 k-匿名处理后的数据集反而没有处理前的隐私保护安全性高。

(5) 对等性攻击(isometric attack) 是指在经 k-匿名化处理后发布的数据集中,如果在某一个等价类中,所有元组敏感属性值都具有相对同等重要地位(即在同一敏感属性分类中具有相对同等重要的权重值[见 2.3 节])的情况下所发生的攻击行为。

例 6 假设 Alice 知道同事 Martin 的医疗记录在表 4 中,并且她也知道 Martin 的年龄在 30 岁到 35 岁之间,居住地的邮政编码为 94502。Alice 根据表 4 中的数据记录信息可以准确的推断出 Martin 的医疗记录在等价类 $\{R_4, R_5, R_6\}$ 中,由于 Stomach Cancer(胃癌)、Pneumonia(肺炎)、Angina Pectoris(心绞痛)这三种疾病在图 1 的疾病分类层次树上都具有同等重要的权重值,它们分别是1.0、0.8和0.8,因此 Alice 能够推断出 Martin 已经得了十分严重的疾病。

表 3 原始病例记录表

序号	姓名	年龄	邮编	疾病
1	Taylor	23	94534	Digestive
2	Martin	31	94502	Pneumonia
3	White	28	94507	Angina Pectoris
4	Clark	18	94509	Stomach disease
5	Green	21	94532	Gastric ulcer
6	Robin	35	94505	Stomach Cancer

表 4 表 3 经过 3-匿名化处理后的数据

序号	年龄	邮编	疾病
1	[0 – 25]	945 **	Digestive
2	[0 – 25]	945 **	Stomach disease
3	[0 – 25]	945 **	Gastric ulcer
4	[26 – 35]	945 **	Pneumonia
5	[26 – 35]	945 **	Stomach Cancer
6	[26 – 35]	945 **	Angina Pectoris

(6) 包含性攻击(inclusion attack) 是指在经 k-匿名化处理后发布的数据集中,如果在某一个等价类中,存在超过半数元组的敏感属性值都被包含在某一特定域值时所发生的攻击行为。

例 7 假设 Alice 知道同事 Taylor 的医疗记录在表 4 中,并且她也知道 Taylor 的年龄在 20 岁到 25 岁之间,居住地的邮政编码为 94534。Alice 通过研究发现,Taylor 的医疗记录一定在等价类 $\{R_1, R_2, R_3\}$ 中,由于 Digestive(消化系统疾病)包含 Stomach disease(胃部疾病),并且 Stomach disease 包含 Gastric ulcer(胃溃疡),即 Gastric ulcer $\in$ Stomach disease $\in$ Digestive,因此 Alice 能够推断出 Taylor 已经得了消化系统疾病。

通过上述 6 种最典型的攻击案例可以充分说明,k-匿名隐私保护模型在特定的应用场景下不能完全地防止敏感属性信息的泄漏,不能有效地保护个人隐私信息,存在一定的缺陷与不足。

2 FVS k-匿名((f,v)-敏感属性 k-匿名隐私保护方法)

由于 k-匿名隐私保护模型存在 1.2 节中所描述的敏感属性信息易受攻击的安全风险,Machanava-

jjhala 等人在 k-匿名隐私保护模型的基础上,针对同质攻击[5]和背景知识攻击[5],提出了一种新的隐私保护模型,称为 ℓ-多样性模型[5]。ℓ- 多样性模型虽然考虑到敏感属性值的多样性,但是并没有考虑到敏感属性值之间可能存在的内在联系,它忽略了不同敏感属性值之间可能存在的包含、层级、相似关系。Li 等人在 ℓ-多样性的基础上,针对相似性攻击和非对称性攻击,提出了 t-接近模型[6]。t-接近模型要求每个等价类内敏感属性值具有与全体敏感属性值成比例的线性分布特性,所以全体敏感属性值的分布状态直接影响了 t-接近模型处理后的结果。在实际应用中,t-接近模型应用效果并不理想。Wong 等人提出了(a,k)-匿名模型[7],该模型通过限制敏感属性值在等价类中出现次数的比重来降低敏感属性信息泄漏的安全风险,在实际应用中,有一定的应用条件要求和特定的限制,具有一定的局限性。在 k-匿名隐私保护模型的基础上,发展起来的隐私保护模型和隐私保护方法还有很多种[8-16],每种隐私保护模型和隐私保护方法都有特定的应用场景和应用参数限制,在实际应用过程中都具有一定的局限性。本文主要阐述了我们在静态数据发布领域提出的一种新的通用的隐私保护方法-FVS k-匿名隐私保护方法,它是在 k-匿名隐私保护模型的基础上发展而来,既能有效地防止 1.2 节中所描述的攻击案例的发生,又没有任何特定参数限制,克服了 ℓ-多样性、t-接近等模型的缺点,在实践应用中具有更高的安全性和更广的应用性。

2.1　FVS K-匿名隐私保护方法描述

定义 2 真子树:树 T 是一棵高度为 h 的树,第 $f(1 \leqslant f < h)$ 层结点的子结点自身以及其子结点所组成的树,被称为第 f 层结点的真子树,记作 LSubTree$[f]$。

在敏感属性值分类层次树 T 中,根结点为第一层,第一层结点的子结点所在的层为第二层,第二层结点的子结点所在的层为第三层,依此类推。若第 f 层 $n(n \geqslant 1)$ 个结点一共有 $v(v \leqslant n)$ 棵真子树,每棵真子树分别记作为 LSubTree$[f_1]$,…, LSubTree$[f_v]$,用$|$ LSubTree$[f]|$表示第 f 层 $n(n \geqslant 1)$ 个结点的真子树的总数目,则 $v = |$ LSubTree$[f]|$。

引理 1 在敏感属性 $SA_i(1 \leqslant i \leqslant n)$ 值分类层次树 T 中,第 $f(1 \leqslant f < h, h$ = 树 T 的高度$)$层结点以及其祖先结点,不会被选择到发布数据集中。

在进行敏感属性值选择发布时,只能选择敏感属性值分类层次树 T 中第 f 层以下的结点或叶结点,第 f 层结点及其祖先结点不在发布数据集中。

引理 2 在敏感属性 $SA_i(1 \leqslant i \leqslant n)$ 值分类层次树 T 中,第 $f(1 \leqslant f < h, h$ = 树 T 的高度$)$层结点的任何两棵真子树,它们都不存在交集,即不存在公共的父结点或祖先结点。

在数据组织发布的数据集中,敏感属性值大多数(如疾病、收入存款等)都可以利用树型结构分类层次树来表示。在本文中,如果没有特殊说明,我们假设 f = 1,默认为第一层结点的真子树,即根结点的真子树。例如在图 1 Disease(疾病)的分类层次树中,第一层结点也就是根结点的真子树一共有 3 棵,分别记作 LSubTree$[1_1]$, LSubTree$[1_2]$, LSubTree$[1_3]$,其中 LSubTree$[1_1]$是以 respiratory system(呼吸系统)结点为根结点的真子树,LSubTree$[1_2]$是以 digestive system(消化系统)结点为根结点的真子树,LSubTree$[1_3]$是以 cardiovascular system(心血管系统)结点为根结点的真子树。

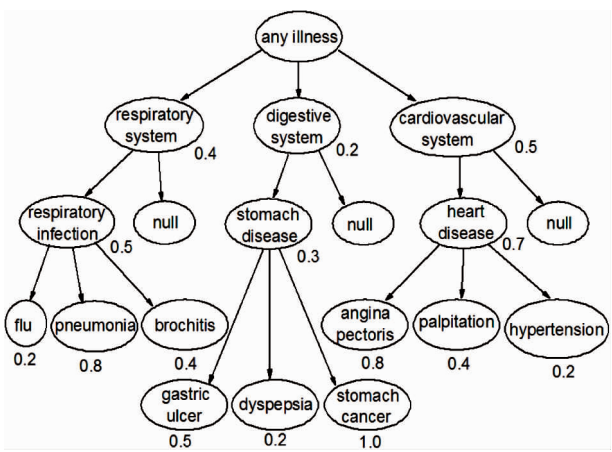

图 1　Disease 敏感属性值的分类层次树

定义 3 (f,v)-敏感属性 k-匿名(FVS k-匿名):如果一个等价类中的所有敏感属性值都来源于该敏感属性值分类层次树中的第 $f(1 \leqslant f < h, h$ 代表该敏感属性值分类层次树的树高$)$层 $v(v \geqslant 2)$ 棵不同真子树中的结点或叶结点,那么该等价类满足$(f,$

v)-敏感属性;若数据表中所有的等价类都满足(f, v)-敏感属性,那么该数据表满足(f,v)-敏感属性。

从定义可以推论出,在敏感属性值的分类层次树中,若第 f 层的真子树总数目越大,每棵真子树 LSubTree$[f_i]$ ($1 \leqslant i \leqslant v$) 树高越高、树中结点越丰富,则每一个等价类中敏感属性值的可选择的范围越广,种类越丰富,那么敏感属性可能存在泄漏的安全风险越小。参数 f 的选择主要由树 T 的高度、树中结点的丰富程度、发布数据的安全级别要求决定的。当树 T 的高度越高、树中的结点越丰富、安全级别要求越低时,参数 f 可以选择大于 1 并且小于 h 的相应层级数值。下面我们通过具体的例子来说明 FVS k-匿名隐私保护方法。

例 8 原始病例记录表表 1 中敏感属性 Disease 值的分类层次树如图 1 所示。图 1 中根结点一共有 3 个子结点,每个子结点自身与它的子孙结点构成一棵真子树,所以第一层的真子树总数目一共有 3 棵;第二层的真子树一共有 6 棵,其中 3 棵真子树为空树(null),属于无效真子树,实际有效的真子树总数目为 3 棵。依此类推,第三层的有效真子树一共有 9 棵。根据 FVS k-匿名隐私保护方法,我们采用默认值 $f = 1$, 选择第一层的真子树总数目 3 作为参考发布 v 值,即 $v = 3$,发布 $k = 3$ 匿名数据集。在每一个等价类中,敏感属性值分别从第一层的不同真子树的结点(或叶结点)中选择。表 5 是(1,2)-敏感属性 3-匿名化原始病例记录表表 1。在等价类 $\{R_4, R_5, R_6\}$、$\{R_7, R_8, R_9\}$ 中,每个等价类的敏感属性值都是从第一层的 3 棵不同真子树中选择的叶结点,所以 $V_2 = 3$, $V_3 = 3$。在等价类 $\{R_1, R_2, R_3\}$ 中,敏感属性值仅从第一层的 2 棵真子树中选择,则 $V_1 = 2$, 所以最终实际发布值 $v = \text{Min}(V_1, V_2, V_3) = \text{Min}(2,3,3) = 2$。

在表 5 中出现了等价类映射值交叉(例如 Age 映射值[0 - 30]与[20 - 45]出现交集[20 - 30]),由此可能会引起发布数据集精确度下降的情况,这是由于举例用的原始病例记录表总记录特别少、包含多种常见攻击案例场景引起的,在我们真实的实验数据处理过程中,这种情况基本不会发生。

表 5 表 1 经过(1,2)-敏感属性 3-匿名化处理后的数据

序号	年龄	邮编	疾病
1	[0 - 30]	94***	AnginaPectoris
2	[0 - 30]	94***	Angina Pectoris
3	[0 - 30]	94***	Flu
4	[20 - 45]	94***	Stomach Cancer
5	[20 - 45]	94***	Pneumonia
6	[20 - 45]	94***	Angina Pectoris
7	[36 - 60]	94***	Angina Pectoris
8	[36 - 60]	94***	Stomach Cancer
9	[36 - 60]	94***	Bronchitis

定理 1 在 FVS k-匿名隐私保护方法发布的数据集中,任何等价类中的敏感属性值都不相同,不满足同质攻击发生的前提条件是属性值相同。

证明:因为在 FVS k-匿名隐私保护方法发布的数据集中,每一个等价类中的敏感属性值都是从该敏感属性值分类层次树第 $f(1 \leqslant f < h)$ 层 $v(v \geqslant 2)$ 棵不同的真子树 $\{\text{LSubTree}[f_1], \cdots, \text{LSubTree}[f_v]\}$ 中选取,所以在任意一个等价类中,任意选取的两个敏感属性值结点 Node_x 和 Node_y ($1 \leqslant x, y \leqslant$ 所属真子树结点最大范围),必存在 LSubTree$[f_i]$ 和 LSubTree$[f_j]$ ($1 \leqslant i, j \leqslant v$, $i \neq j$), 使得 $\text{Node}_x \in$ LSubTree$[f_i]$, $\text{Node}_y \in$ LSubTree$[f_j]$, 同时又因为 LSubTree$[f_i] \cap$ LSubTree$[f_j] = \varnothing$, 所以 $\text{Node}_x \cap \text{Node}_y = \varnothing$。即每一个等价类中敏感属性值都不存在相同的情况,不满足同质攻击发生的前提条件。

定理 2 在 FVS k-匿名隐私保护方法发布的数据集中,发生敏感属性泄漏的可能性为 $1/v$。

证明:因为在 FVS k-匿名隐私保护方法发布的数据集中,每一个等价类中的敏感属性值都是从该敏感属性值分类层次树第 $f(1 \leqslant f < h)$ 层的 $v(v \geqslant 1)$ 棵不同的真子树 $\{\text{LSubTree}[F_1], \cdots, \text{LSubTree}[F_V]\}$ 中选取,任意两棵真子树的交集为空,即 LSubTree$[f_i] \cap$ LSubTree$[f_j] = \varnothing$ ($1 \leqslant i, j \leqslant v$, $i \neq j$), 发生敏感属性泄漏的可能性为 $1/v$。假设还存在任意结点 $\text{Node}_x \in$ LSubTree$[f_z]$ 能够降低发生敏感属性泄漏发生的可能性,由于敏感属性值层次树第 f 层,最多只有 v 棵不同的真子树,根据鸽笼原理,LSubTree$[f_z]$ 只能存在于 $\{\text{LSubTree}[f_1], \cdots, \text{LSub-}$

Tree$[f_v]$}中,故存在 LSubTree$[f_m]$ $(1 \leqslant m \leqslant v)$,使得 LSubTree$[f_z]$ = LSubTree$[f_m]$或 LSubTree$[f_z] \in$ LSubTree$[f_m]$,所以 $\text{Node}_x \in$ LSub Tree$[f_z] \in$ LSubTree$[f_m]$,即 $\text{Node}_x \in$ LSubTree$[f_m]$,所以不存在结点 Node_x能够减少敏感属性泄漏发生的可能性。

定理 3 在 FVS k-匿名隐私保护方法发布的数据集中,任何等价类的敏感属性值都不相似,不满足相似性攻击发生的前提条件是属性值相似。

证明:因为在 FVS k-匿名隐私保护方法发布的数据集中,每一个等价类中敏感属性值都是从该敏感属性值分类层次树第$f(1 \leqslant f < h)$层的$v(v \geqslant 2)$棵不同的真子树{LSubTree$[f_1]$,…,LSubTree$[f_v]$}中选取,任意两棵真子树的交集为空,即 LSubTree$[f_i] \cap$LSubTree$[f_j] = \varnothing$ $(1 \leqslant i, j \leqslant v,\ i \neq j)$,不存在$n(n \geqslant 2)$个敏感属性值共享同一父结点或祖先结点 LSubTree$[f_m]$ $(1 \leqslant m \leqslant v)$ 的情况,即每个等价类中敏感属性值都不相似,所以 FVS k-匿名隐私保护方法发布的数据集不满足相似性攻击发生的前提条件。

定义 4 在 FVS k-匿名隐私保护方法发布的数据集中,当敏感属性值分类层次树只有两层并且第一层只有两棵真子树,每一棵真子树所对应的元组数目差值巨大时,则选择按第一层结点值进行数据发布。

该定义是对 k-匿名隐私保护模型、ℓ-多样性模型中参数值 k、ℓ必须大于等于 2 的限制的扩展,即 FVS k-匿名隐私保护方法中 $v \geqslant 2$,当 $v = 2$ 时,是有一定条件限制的。这种限制可以有效地避免非对称性攻击。

定理 4 在 FVS k-匿名隐私保护方法发布的数据集中,任何等价类都不满足对等性攻击发生的前提条件是属性值具有相对同等重要权重值(即在同一敏感属性分类中具有相对同等重要的权重值)。

证明:因为在 FVS k-匿名隐私保护方法发布的数据集中,每一个等价类中敏感属性值都是从该敏感属性值分类层次树第$f(1 \leqslant f < h, h$ =该敏感属性层次树的高度)层的$v(v \geqslant 2)$棵不同的真子树{LSubTree$[f_1]$,…,LSubTree$[f_v]$}中选取,真子树中每个结点都具有相对全局属性值的权重值,使得在任意等价类中,敏感属性值由低到高分布,不存在结点权重值相等(或接近)的情况,所以 FVS k-匿名隐私保护方法发布的数据集不满足对等性攻击发生的前提条件。

定理 5 在 FVS k-匿名隐私保护方法发布的数据集中,任何等价类都不满足包含性攻击发生的前提条件是属性值之间具有包含关系。

证明:因为在 FVS k-匿名隐私保护方法发布的数据集中,每一个等价类中的敏感属性值都是从该敏感属性值分类层次树第$f(1 \leqslant f < h)$层$v(v \geqslant 2)$棵不同的真子树{LSubTree$[f_1]$,…,LSubTree$[f_v]$}中选取,任意两棵真子树的交集为空,即 LSubTree$[f_i] \cap$LSubTree$[f_j] = \varnothing$ $(1 \leqslant i, j \leqslant v,\ i \neq j)$,所以在任意一个等价类中,任意选取的两个敏感属性值结点 Node_x和 Node_y($1 \leqslant x, y \leqslant$ 所属真子树结点最大编号),必存在 LSubTree$[f_i]$和 LSubTree$[f_j]$ $(1 \leqslant i, j \leqslant v,\ i \neq j)$,使得 $\text{Node}_x \in$ LSubTree$[f_i]$,$\text{Node}_y \in$ LSubTree$[f_j]$,同时又因为 LSubTree$[f_i] \cap$ LSubTree$[f_j] = \varnothing$,所以 $\text{Node}_x \cap \text{Node}_y = \varnothing$。即每一个等价类中敏感属性值不存包含的情况,不满足包含性攻击发生的前提条件。

2.2　FVS k-匿名隐私保护方法多敏感属性扩展

定义 5 $((f_1, v_1), \cdots, (f_n, v_n))$-敏感属性:$k$-匿名如果一个等价类中的 n 个敏感属性值都分别来源于该 n 个敏感属性值分类层次树中的第f_i层$(1 \leqslant i < h,\ h$ = 第 i 个敏感属性值分类层次树的高度)$v_j(v_j \geqslant 2)$棵不同真子树中的结点(或叶结点),那么该等价类满足$((f_1, v_1), \cdots, (f_n, v_n))$-敏感属性;若数据表中所有的等价类都满足$((f_1, v_1), \cdots, (f_n, v_n))$-敏感属性,那么该数据表满足$((f_1, v_1), \cdots, (f_n, v_n))$-敏感属性。

利用 FVS k-匿名隐私保护方法处理发布数据集时,不仅能有效地保护敏感属性信息,而且方法更简单,手段更灵活。对于 n 个敏感属性,每一个敏感属性值的选择,只需要按单敏感属性值的选择方法进行选择,然后再将这些被选择的敏感属性值组合在一起形成最后要发布数据集中的敏感属性值。若 n 个敏感属性对应的真子树数目分别为|LSubTree$[f]_1$|, …, |LSubTree$[f]_n$|,则最后发布数据集

中每一个等价类的敏感属性值的总数目 $v = \sum_{i=1}^{n} |\mathrm{LSubTree}[f]_i|$。

2.3 FVS k-匿名隐私保护方法优化

FVS k-匿名隐私保护方法选择敏感属性值时，也存在两方面的缺陷和不足。

(1) 选择敏感属性值的随意性

FVS k-匿名隐私保护方法生成的每一个等价类中，敏感属性值都是从 v 棵不同的真子树结点中随机选择，具有很大的随意性，例如，随机选择敏感属性发布的数据集容易发生对等性攻击[1.2 节，例 6]。我们通过为敏感属性值分类层次树中的结点(或叶结点)增加权重值(敏感属性值的权重值是比照全体敏感属性值进行设定的)，图 1 中结点(或叶结点)下面的数值表示该结点(或叶结点)的权重值，该方法避免了在发布的数据集中，某一个等价类中敏感属性值的权重值不存在非常接近的情况，因此避免了对等性攻击案例的发生，增加了攻击者分析敏感属性的复杂性，有效地防止了敏感属性信息泄漏，提高了发布数据的安全性，对抑制其他攻击(例如相似性攻击[6]、同质攻击[5])也具有重要的意义。FVS k-匿名隐私保护方法在进行真子树结点(或叶结点)选择时，充分考虑了结点(或叶结点)的权重值，使得每个等价类中选取的敏感属性值的权重值由低到高均匀分布，从而增加了发布数据集中敏感属性的安全性、减少了敏感属性泄漏的风险、提高了敏感属性的不可推测性。

(2) 敏感属性值分类层次树中第 f 层 v 棵不同真子树所对应的元组总数目分布不均匀

如果敏感属性值分类层次树第 f 层 v 棵不同真子树所对应的元组总数目分布均匀，那么每一个等价类中敏感属性值的选择就比较容易，也比较方便。但在实际应用过程中，v 棵不同真子树所对应的元组总数目分布往往是不均匀的，此时为了保证隐私安全保护的效果，就不可以简单地直接进行选择。而是采用按单个真子树所对应元组的总数目与全体真子树所对应的元组总数目的比值，来选择每一个等价类中该真子树结点的数目。比值的计算方法是每一个等价类中第 i 棵真子树选择结点的数目 = (第 i 棵真子树所对应元组总数目/第 f 层全部真子树所对应元组总数目) × v，($1 \leqslant i \leqslant v$，$v$ 代表第 f 层真子树总数)。例如：在某个敏感属性值分类层次树中一共有 4 棵真子树，其中 LSubTree[1] 真子树所有结点对应的元组总数目为 23 个，4 棵真子树全部结点所对应的元组总数据目为 64 个，那么发布数据集中每一个等价类的敏感属性值从 LSubTree[1] 真子树所对应结点中选取的敏感属性值总数为 (23/64) ×4≈2 个。该方法能够保证在同一棵真子树中结点(或叶结点)均匀分布，提高了攻击者分析敏感属性值的难度，降低了敏感属性信息泄漏的风险。

3 实 验

我们根据实验测试结果，评估了 FVS k-匿名隐私保护方法的算法运行时间性能，分析了经该算法处理后发布数据集泄漏的可能性，并分别与 k-匿名隐私保护模型、ℓ-多样性模型、t-接近模型进行了运行时间性能、隐私安全性能的对比，并对算法的参数 f 和 v 组合的多样性进行了评估，同时也对发布数据集的质量进行了对比分析。

本实验中，我们采用美国 UCI(University of California，Irvine)所提供的机器学习库中的成人数据集。数据集由美国人口普查数据组成，共计 32561 个元组。在该数据集中一共选取了 9 个属性字段：Age，Workclass，Education，Race，Gender，Occupation，Marital-status，Country，Income。实验中所使用的软硬件参数如下：(a)操作系统：Windows7 x64 Professional Edition；(b)硬件参数：Intel CoreTM i3-370M 2.4GHz CPU，6GB DDR 内存；(c)编译环境：Microsoft Visual Studio 2012 C + +。此外，FVS k-匿名隐私保护方法的实现算法是在 Incognito[8] 算法的基础上扩展实现的。

3.1 FVS k-匿名隐私保护方法隐私泄漏的可能性分析

实验中，我们采用了差别度量方法，将 FVS k-匿名隐私保护方法 (f = 1，v = 5) 与 k-匿名隐私保护模型、ℓ-多样性模型(ℓ = 5，c = 4)、t -接近模型(t = 0.2)进行了数据泄漏可能性对比分析。原始数据

集元组大小为 100kb,按元组增量每 10kb 为一个度量点进行了敏感属性泄漏可能性比较,对比分析结果如图 2 所示。其中原始数据集在 0kb ~2.5kb 区间进行细分比较,元组增量为 0.5kb,对比分析结果如图 3所示。根据四种模型方法实验测试的数据结果,我们对比分析了这四种模型方法在数据泄漏可能性方面的差异,可以得出,FVS k-匿名隐私保护方法明显优于其它三种隐私安全保护模型。参数 ℓ, t,c 的选择参考了隐私安全保护领域的经典文献[6,15],具有重要的参考衡量标准。FVS k-匿名隐私保护方法能够对个体的敏感属性信息提供更安全的防护。

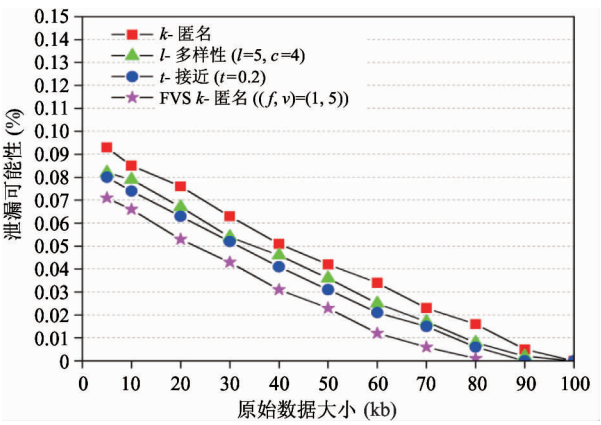

图 2　数据泄漏可能性分析(全部测试数据)

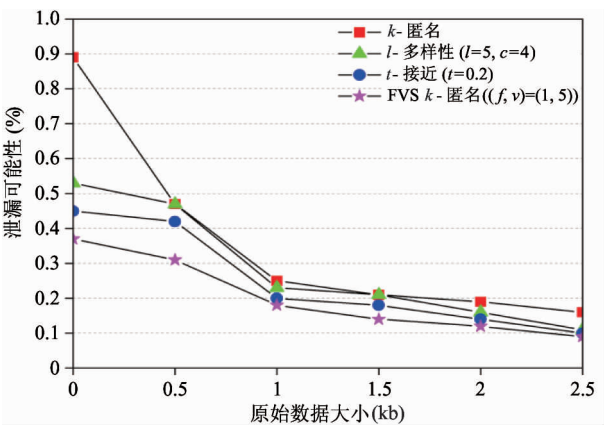

图 3　按准标符组成元素个数进行性能比较

3.2　FVS k-匿名隐私保护方法运行时间性能分析

通过实验,我们将 FVS k-匿名隐私保护方法 (f = 1, v = 5) 与 k-匿名隐私保护模型、ℓ-多样性模型 (ℓ=5, c =4)、t-接近模型(t =0.2)按准标识符组成元素的个数由小到大进行了运行时间的对比分析,对比分析结果如图 4 所示。当 k,ℓ分别取不同数值时的运行时间对比分析结果如图 5 所示。通过运行时间的对比分析,我们发现 FVS k-匿名隐私保护方法运行时间相对较长,究其原因是由于该实现算法在 k-匿名隐私保护模型的实现算法的基础上,增加了对敏感属性值分类层次树的层级结点的检索处理,同时还增加了对该分类层次树中结点权重值的平衡选取的判断。FVS k-匿名隐私保护方法在总体运行时间上并不优于其它隐私安全保护模型算法,它的总体运行时间相对较长,主要因为它在优化 k-匿名隐私保护模型的过程中,增加了对敏感属性信息的安全保护处理环节,但是整体时间开销的差别不大,在系统使用用户可接受范围,它虽然牺牲了系统性能(运行时间),却提高了发布数据集的个人隐私信息安全性。

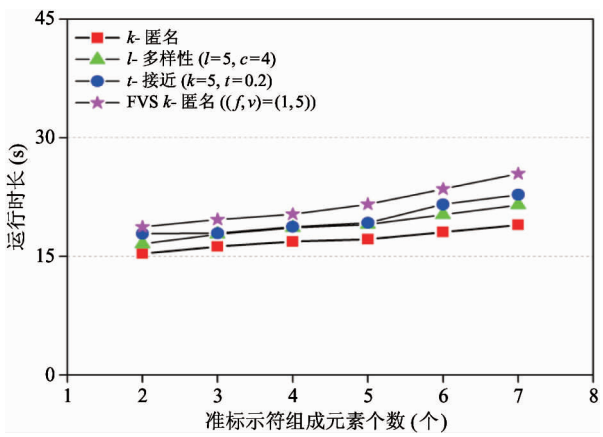

图 4　数据泄漏可能性分析(部分测试数据)

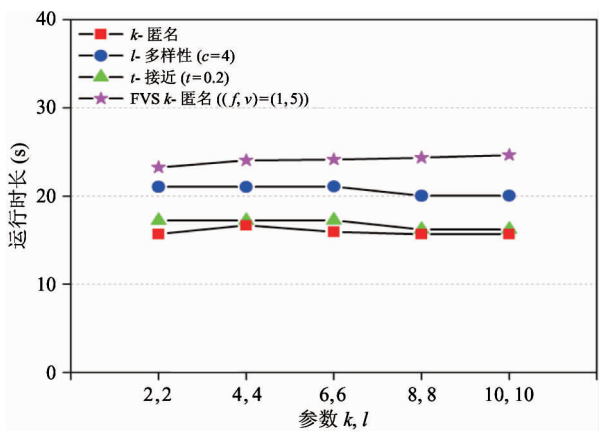

图 5　按 k,l 参数不同数值进行性能比较

3.3　FVS k-匿名隐私保护方法发布数据质量对比分析

在发布隐私保护的数据集中，发布数据集的精确度也是衡量隐私保护方法优劣的一个重要指标。在本实验中，测试数据中的敏感属性 Income 值采用层次分类树的表示方法，一共分成 5 个层次等级。我们采用直接比较的方法[17,18]，对原始数据经过匿名化处理后的数据信息丢失率进行测算。为了对比分析匿名化前后数据的真实对应关系，我们对原始数据进行了改造，增加了唯一标示信息，使得匿名化处理后的数据，能够通过唯一标示信息找到原始数据，然后进行信息丢失率计算。对隐私保护模型中等价类的数目采用 k = (51520253035404550) 进行多级数据信息丢失率对比，结果如图 6 所示。从图中可以得出随着 k 值的逐渐增大，匿名化处理后的数据信息损失程度越大。但 FVS k-匿名数据精确要高于使用其他方法的数据精度，这是由于在匿名化的过程中，元组数据先按敏感属性分类层次进行排序和归类，使得在生成 k 值等价类过程中，元组数据不需要做更多的概化处理。

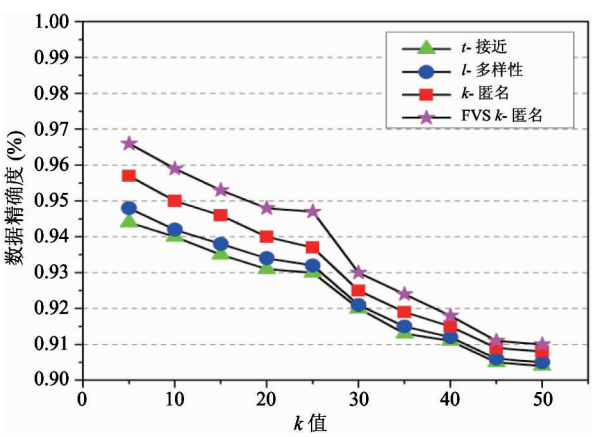

图 6　发布数据精确度对比

3.4　FVS k-匿名隐私保护方法参数 (f,v) 选择对隐私安全的影响

在 FVS k-匿名隐私保护方法中参数 (f,v) 选择对发布数据集的隐私安全有着重要的影响。f 的可选择值在 1 和敏感属性分类层次树的树高之间，选择的 f 值越接近 1，说明敏感属性值在分类层次树中的位置越接近根结点，敏感数据的隐私保护安全性越高，隐私泄漏的可能性越低，反之敏感属性值与实际的真实值越远，数据的真实性和可信性越低。v 值表示在敏感属性分类层次树中同一层的真子树的数量，v 值越大，表示可选择的 LSubTree 越多，敏感属性值的可选择范围越广，敏感属性值泄漏的可能性越低，数据隐私保护安全性越高。f 和 v 的组合的相对安全性评估值用 v/f 来表示，在实验中，选取第一、二、三层不同组合的 LSubTree 数目进行隐私安全性对比实验。对比结果如图 7 所示。实验数据表明，v/f 的比值越大，数据泄漏的可能性越低，反之，数据泄漏的可能性越高。当 v 值相同时，f 值越小，意味着敏感属性值所在的层级越高，数据值被概化的粒度越大，信息泄漏的可能性越低。当 f 值相同时，v 值越大，意味着敏感属性值的可选择范围越大，信息泄漏的可能性越低。

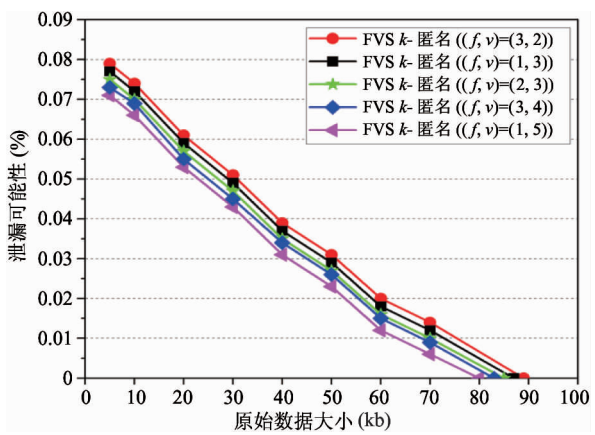

图 7　参数 (f,v) 选择对安全性的影响

4　结　论

本文提出了一种新的隐私保护方法——FVS k-匿名隐私保护方法，该方法通过引入真子树的概念，进而建立了全新的敏感属性值的选择方法，针对系统选择敏感属性值的随机性缺陷，我们通过为真子树中不同结点设定不同权重值，来标识该结点在树中的权重，同时又根据真子树所对应元组总数目，进行权衡分析选择，最终确立敏感属性值的选择方法。FVS k-匿名隐私保护方法能有效地防止 k 匿名隐私保护模型中敏感属性信息的泄漏，对多敏感属性具有良好的扩展性。l-多样性模型，t-接近模型可以

看作是 FVS k-匿名隐私保护方法中的两种特殊案例。在 ℓ-多样性、t-接近等模型中参数 ℓ,t 选择比较困难,对不同的隐私安全衡量标准,参数 ℓ,t 会有不同的选择标准。在 FVS k-匿名隐私保护方法中,参数 f,v 的选择与敏感属性值的分类层次树相关,选择方法比较直观,不同的应用者不会产生太大的歧异性,而且该方法在个人隐私数据信息保护方面具有更高的安全性和可靠性。

参考文献

[1] Sweeney L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002, 10(05): 557-570

[2] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002, 10(05): 571-588

[3] Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression [Technical report]: SRI International, 1998

[4] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information. In: ACM Symposium on Principles of Database Systems, 1998, 98: 188

[5] Machanavajjhala A, Kifer D, Gehrke J, et al. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2007, 1(1): 3

[6] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. In: Proceedings of the IEEE 23rd International Conference on Data Engineering, 2007. 106-115

[7] Wong R C W, Li J, Fu A W C, et al. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge discovery and data mining. ACM, 2006. 754-759

[8] LeFevre K, DeWitt D J, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. ACM, 2005. 49-60

[9] LeFevre K, DeWitt D J, Ramakrishnan R. Mondrian multidimensional k-anonymity. In: Proceedings of the 22nd International Conference on Data Engineering, 2006. 25-25

[10] Xiao X, Tao Y. Personalized privacy preservation. In: Proceedingsof the 2006 ACM SIGMOD International Conference on Management of Data. ACM, 2006. 229-240

[11] 杨晓春, 刘向宇,王斌等. 支持多约束的 K-匿名化方法. 软件学报, 2006,17(5):1222-1231

[12] Xiao X, Wang G, Gehrke J. Interactive anonymization of sensitive data. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. ACM, 2009. 1051-1054

[13] Wang K, Fung B. Anonymizing sequential releases. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006. 414-423

[14] Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation. In: Proceedings of the 32nd International Conference on Very Large Data Bases. VLDB Endowment, 2006. 139-150

[15] Machanavajjhala A, Kifer D, Gehrke J, et al. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2007, 1(1): 3

[16] Soria-Comas J, Domingo-Ferrer J, Sánchez D, et al. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *The International Journal on Very Large Data Bases*, 2014, 23(5): 771-794

[17] Navarro-Arribas G, Torra V, Erola A, et al. User k-anonymity for privacy preserving data mining of query logs. *Information Processing & Management*, 2012, 48(3): 476-487

[18] Lixia W, Jianmin H. Utility evaluation of k-anonymous data by microaggregation. In: Proceedings of the 2009 ISECS International Colloquium on Computing, Communication, Control, and Management. 2009. 381-384

FVS *k*-anonymity: an anonymous privacy protection method based on *k*-anonymity

Wang Liang, Wang Weiping, Meng Dan
(Institute of Information Engineering, Chinese Academy of Science, Beijing 100093)

Abstract

The study aimed to ensure the safety of individul sensitive information when publishing data. The mechanism and performance of the *k*-anonymity technique for anonymous privacy protection were deeply studied, and aiming at its problem of incapable of complete protection of sensitive attributes against discloure, a new privacy protection approach based on the *k*-anonymity model, called the FVS *k*-anonymity, was put forward after experimental investigations. The new approach adopts the concept of true subtree and a wholly-new way for selection of sensitive attribute values, so it can effectively remedy the above-mentioned shortcoming to protect the sensitive from disclosure.

Key words: *k*-anonymity, information security, privacy protection, sensitive attribute