DOI: 10.11817/j.issn.1672-7207.2015.10.025

基于聚类相关性约束的(*s*,*l*)-多样性匿名方法

张冰[1,2]，杨静[1]，张健沛[1]，谢静[3]

(1. 哈尔滨工程大学 计算机科学与技术学院，黑龙江 哈尔滨，150001;
2. 哈尔滨理工大学 软件学院，黑龙江 哈尔滨，150080;
3. 武汉纺织大学 管理学院，湖北 武汉，430200)

摘要：针对传统 *l*-多样性模型易形成敏感值高度相关的等价类问题，提出一种约束等价类中敏感值相关性的(*s*,*l*)-多样性模型。该模型在传统 *l*-多样性模型的基础上，以敏感集合中非敏感属性值的分布度量敏感值的相关性，通过等价类中敏感值相关性的约束来降低高相关性敏感值产生的信息泄露。同时，使用属性值间相关性作为距离度量基准，提出一种(*s*,*l*)-多样性聚类算法(SLCA)来实现该匿名模型，以降低数据泛化过程中的信息损失。研究结果表明：SLCA 算法具有较小的信息损失量与较短的运行时间，能够有效地降低等价类中敏感值的相关性，更好地防止个体敏感信息泄露。

关键词：(*s*,*l*)-多样性；相关性约束；匿名；聚类；隐私保护
中图分类号：TP309.2 **文献标志码**：A **文章编号**：1672−7207(2015)10−3733−10

A correlativity constrained (*s*,*l*)-diversity anonymity method based on clustering

ZHANG Bing[1,2], YANG Jing[1], ZHANG Jianpei[1], XIE Jing[3]

(1. School of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China;
2. School of Software, Harbin University of Science and Technology, Harbin 150080, China;
3. School of Management, Wuhan Textile University, Wuhan 430200, China)

Abstract: In allusion to the problem of traditional data anonymity models constructing equivalence class with high correlative sensitive values, (*s*,*l*)-diversity was proposed which limited the correlativity of sensitive values in the equivalence classes. This diversity model was based on traditional *l*-diversity model, and it measured the correlativity of the sensitive attribute values to decrease the information loss by equivalence classes with high corrective sensitive values. At the same time, a (*s*,*l*)-diversity clustering algorithm named SLCA was proposed to achieve (*s*,*l*)-diversity, and the SLCA algorithm measured the distance between tuples by measuring the correlativity of attribute values, which greatly decreased the information loss during data generation. The results show that SLCA algorithm is more effective in terms of both information loss and execution time, and SLCA algorithm can effectively decrease the correlativity of the sensitive values in the equivalence classes to protect the privacy security of the data sets.

Key words: (*s*,*l*)-diversity; correlativity constrained; anonymity; clustering; privacy preservation

收稿日期：2014−10−28；**修回日期**：2014−12−29
基金项目(Foundation item)：国家自然科学基金资助项目(61370083，61073043，61073041)；高等学校博士学科点专项科研基金资助项目(20112304110011，20122304110012)(Projects (61370083，61073043，61073041) supported by the National Natural Science Foundation of China; Projects (20112304110011，20122304110012) supported by the Research Fund for the Doctoral Program of Higher Education)
通信作者：杨静，教授，博士生导师，从事数据的隐私保护、数据挖掘、语义社会网络等研究；E-mail：yangjing@hrbeu.edu.cn

近年来，政府和研究机构发布大量数据以供相关人员进行统计分析，这些数据中往往涉及个体隐私的敏感信息，随之而来的隐私泄露问题便显得尤为突出。因此，在信息共享的同时维护个体敏感信息的私密性便逐渐成为人们研究的热点。目前，学者们研究出大量的隐私保护方法，如数据取样、数据交换、数据抑制、数据扰乱等，但这些技术都会造成大量的信息损失，降低数据的可用性。数据发布中的隐私保护[1−2]的主要目的是破坏个体身份与敏感属性间的对应关系。传统的隐私保护方法仅删除数据表中的身份标识属性(如姓名、身份证号码等)，无法较好地阻止隐私泄露，一旦攻击者获得发布表中的准标识符属性并结合自身的背景知识，便有可能推测出目标个体与敏感信息间的关联。为解决这种攻击者通过推理获得目标个体的敏感信息的链接攻击[3]，Sweeney[4]提出 k-匿名模型，它要求在最终发布的数据中同一等价类不可区分的实体至少为 k 个。虽然，k-匿名模型为数据发布中的隐私保护技术指明了新的发展方向，且能够抵御链接攻击带来的隐私泄露威胁，破坏个体与数据表中元组间的关联，但它并没有破坏个体与敏感值之间的关联，无法抵御来自攻击者的背景知识攻击和同质性攻击。据此，学者们针对 k-匿名模型存在的问题做了很多改进。Wong 等[5]提出(α,k)-匿名模型，它通过约束敏感值在等价类中出现的比例来实现敏感值的多样性，是 k-匿名模型的一种改进模型。Machanavajjhala 等[6]提出 l-多样性模型，该模型要求发布数据的同一等价类中的敏感属性至少有 $l(l\geqslant 2)$个“较好表现”的值，攻击者最高只能以 $1/l$ 的概率获取个体的敏感信息。Li 等[7]提出 t-closeness 模型，它要求敏感属性的不同敏感值在每个等价类中的分布与数据表中的分布间距离不大于阈值 t，以解决敏感属性值的分布倾斜问题。l-多样性模型和 t-closeness 模型是 k-匿名模型的一个飞跃，虽然这 2 种模型能够保证等价类中敏感值的多样性，但其无法抵御相似性攻击和倾斜攻击，仍无法很好地满足人们对隐私保护的需求。据此，学者们对 l-多样性模型进行了许多改进[8−11]。Sun 等[12]提出一种(l,α)-多样性模型，要求等价类中敏感值的权重和不小于 α，以避免高度敏感的敏感值出现在同一等价类中，实现敏感值的均匀分配。Liu 等[13]提出一种 l^+-diversity 模型，通过对数据表列的划分，使每块分区与敏感值 s_i 的关联概率系数大于 p_i，以实现敏感属性的隐私保护。以聚类的思想实现数据表的匿名化能够有效降低发布数据的信息损失。杨高明等[14]提出一种基于聚类的(α,k)-匿名，以聚类的思想分别实现单敏感值和多敏感值的数据表的(α,k)-匿名。王智慧等[15]提出一种基于聚类的 L-clustering 算法，将数据匿名问题转化为带特定约束的聚类问题，提高数据的可用性。以上方法虽能实现等价类中敏感属性的多样性，但均仅考虑敏感值形式上的差异，忽略了敏感值间的相关性，生成的匿名表具有高隐私泄露风险。例如，以“职业”为敏感属性，若某等价类中不同敏感值分别为“治安警察”、“交通警察”与“刑事警察”，通过观察可知该等价类中敏感值为不同类别的警察，具有高度相关性，一旦攻击者认定目标个体属于该等价类，便可得知目标个体的一些敏感信息，如生活规律、教育程度、身体素质和收入范围等。针对此问题，本文作者从保证等价类中敏感值的多样性与低相关性的需求出发，在传统 l-多样性模型的基础上提出一种相关性约束的(s,l)-多样性模型。该模型以准标识符属性的分布度量敏感属性的相关性，并对等价类中敏感值的相关性进行约束，以克服传统匿名模型生成的等价类中敏感值高相关性的缺点。同时，本文使用聚类的思想实现发布表的(s,l)-多样性，以降低泛化过程中产生的信息损失。

1 基本概念

在数据发布时，数据表中的每条记录对应 1 个真实个体，包含多个属性，这些属性可分为显式标识符属性、准标识符属性、敏感属性和其他属性 4 类。其中，显式标识符属性(identifier attribute)是能够唯一标识 1 条记录的属性，例如姓名、身份证号码等，通常用集合{I_1，I_2，…，I_n}表示；准标识符属性 Q(quasi-identifier attribute)是联合起来能近似确定记录身份的属性组，如邮编、年龄、性别的组合，通常用集合{Q_1，Q_2，…，Q_n}表示；敏感属性(sensitive attribute)是涉及个体隐私，需要被保护的属性，如疾病、职业等，通常用集合{S_1，S_2，…，S_n}表示；其他属性(extra attribute)是数据表中与个体敏感信息无关的属性。

定义 1 等价类。数据表 T 中的若干记录的集合，相同等价类中的每条记录在准标识符上具有相同的属性。表 1 所示为初始个人信息数据表。由表 1 可知：前 3 条记录构成 1 个等价类，它们在{Age，Country，Zip Code}上具有相同的属性值。

定义 2 l-多样性。设 D 为匿名表 T'的 1 个等价类，若 D 中敏感属性至少有 $l(l\geqslant 2)$个不同取值，则称等价类 D 满足 l-多样性；若 T'中的所有等价类都是 l-多样性，则匿名表 T'满足 l-多样性。

表 2 所示为满足 2-多样性的匿名表，每个等价类中不同敏感值的个数都不小于 2。由表 2 可见：*l*-多样性模型虽然保证了敏感值的多样性，但没有考虑敏感值的相关性。例如，如果攻击者获取了 Alice 的年龄、国籍和邮编，推测出 Alice 位于匿名表 *T'*的第 1 个等价类中，但无法获知 Alice 具体的工作情况。通过观察得知，Alice 所在等价类中的敏感值分别为 Doctor 和 Anesthetist，可推测 Alice 在医院工作及她的一些隐私信息，如工作时间为朝八晚五、偶尔值夜班、工作强度大、具有医学背景、收入水平处于中上等。

表 1　初始个人信息数据表

Table 1　Initial personal information data table

ID	Name	Age	Country	Zip Code	Occupation
1	Bob	27	USA	14248	Doctor
2	Lily	28	Canada	14207	Nurse
3	Alice	26	USA	14206	Anesthetist
4	Michael	33	China	14304	Nurse
5	Tom	35	Japan	14399	Teacher

表 2　2-多样性匿名表

Table 2　2-diversity anonymization table

ID	Age	Country	Zip Code	Occupation
1	26~28	America	142**	Anesthetist
2	26~28	America	142**	Doctor
3	26~28	America	142**	Nurse
4	33~35	Asia	143**	Teacher
5	33~35	Asia	143**	Nurse

2　度量空间与(*s*,*l*)-多样性模型

2.1　相关性度量

将数据表中元组按敏感值的不同划分为多个集合，若几个集合的敏感值有高相关性，则这些集合在非敏感属性上的属性值分布相似。例如，设职业为人口信息表中的敏感属性，按职业的不同将该表划分为多个集合，若几个集合的敏感值具有高相关性，则这些集合在人员的年龄组成、教育程度、性别、工作时长等属性上的分布都相似。据此，本节提出一种按属性 *S* 划分数据表，依据集合中不同于 *S* 的属性值分布来度量属性 *S* 的属性值间相关性的方法。本文假设数据表中所有属性是相互独立的，属性间不存在函数依赖关系。对于数值型属性，将属性值映射为相应区间，转化为分类型属性处理。

定义 3　属性值的关系矩阵。设数据表 *T* 中属性 *C* 的属性值集合为$\{C_1, C_2, \cdots, C_n\}(n\geqslant 2)$，*S* 为数据表中不同于 *C* 的属性，*S* 的属性值集合为$\{S_1, S_2, \cdots, S_m\}(m\geqslant 1)$，属性 *C* 的值 C_i 在 *S* 上的关系矩阵 $\boldsymbol{E}(C_i, S)$定义为

$$\boldsymbol{E}(C_i, S)=[P_u(C_i, S_1), P_u(C_i, S_2), \cdots, P_u(C_i, S_m)]$$

其中：$P_u(C_i, S_j)(1\leqslant i\leqslant n, 1\leqslant j\leqslant m)$为属性 *C* 上当属性值为 C_i 时属性 *S* 上属性值为 S_j 的记录的联合概率。

表 3 所示为其中 1 个范例，在属性 *S* 上属性值为 Teacher 的记录中，属性 Q_1 上属性值为 Female 的记录的联合概率，即 P_u(Teacher，Female)=1/7，同理可知 P_u(Teacher，Male)=2/7，则属性 *C* 的值 Teacher 在 Q_1 上的关系矩阵为[1/7，2/7]。

表 3　范例数据表 1

Table 3　Example data table1

ID	Q_1	Q_2	*S*
S_1	Female	9	Teacher
	Male	9	Teacher
	Male	12	Teacher
S_2	Female	12	Doctor
	Male	12	Doctor
	Male	10	Doctor
	Female	9	Doctor

夹角余弦法是度量向量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 间相关性 $R(\boldsymbol{a},\boldsymbol{b})$的常用方法，即 $R(\boldsymbol{a},\boldsymbol{b})=\dfrac{(\boldsymbol{a},\boldsymbol{b})}{|\boldsymbol{a}||\boldsymbol{b}|}$。当向量中各元素均不为负时，夹角余弦的取值范围为[0,1]，取值反映了向量间的相似程度，取值越接近 1，意味着 2 个向量间相似度越高。而属性值间的相关性越高，则属性值的关系矩阵越相似。因此，本文将夹角余弦法作为属性值间相关性的度量方法。

定义 4　属性值间相关性。设数据表 *T* 中属性 *C* 的属性值集合为$\{C_1, C_2, \cdots, C_n\}(n\geqslant 2)$，$\{S_1, S_2, \cdots, S_m\}(m\geqslant 1)$为不同于 *C* 的属性，属性值 C_p 与 $C_q(1\leqslant p, q\leqslant n)$的相关性 $R_C(C_p, C_q)$定义为

$$R_C(C_p,C_q)=\frac{1}{m}\sum_{i=1}^{m}R_C^{S_i}(C_p,C_q)=$$

$$\frac{1}{m}\sum_{i=1}^{m}\cos(\boldsymbol{P}(C_p,S_i),\boldsymbol{P}(C_q,S_i))$$

其中：$R_C^{S_i}(C_p, C_q)$为 C_p 与 C_q 在属性 S_i 上的相关性；$\boldsymbol{P}(C_k, S_i)$为在属性 S_i 上的关系矩阵。

由定义 4 可知：属性值间相关性的取值范围为[0,1]，相同敏感值间相关性为 1。

准标识符属性是能够近似确定数据表中记录的真实身份的 1 组属性。相对其他属性，准标识符属性与敏感属性间的关系更密切，因此，本文依据准标识符属性的分布来度量敏感值间相关性。

定义 5 敏感值间相关性。设数据表 T 中敏感属性 C 的属性值集合为$\{C_1, C_2, \cdots, C_n\}(n\geqslant 1)$，准标识符属性为$\{Q_1, Q_2, \cdots, Q_k\}(k\geqslant 1)$，敏感属性 C 的属性值 C_p 与 C_q 间相关性定义为

$$R_C(C_p, C_q)=\frac{1}{k}\sum_{i=1}^{n}R_C^{Q_i}(C_p, C_q)$$

其中：$R_C^{Q_i}(C_p, C_q)$ 为属性 Q_i 上属性值 C_p 与 C_q 的相关性。

定义 6 属性的相关矩阵。设数据表 T 中属性 C 的属性值集合为$\{C_1, C_2, \cdots, C_n\}(n\geqslant 1)$，属性值 C_i 与 C_j 的相关性为 $C_{C_iC_j}$，属性 C 的相关矩阵 $\boldsymbol{R}_C$ 定义为 1 个 n 阶矩阵：

$$\boldsymbol{R}_C=\begin{pmatrix} C_{C_1C_1} & C_{C_1C_2} & \cdots & C_{C_1C_n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{C_nC_1} & C_{C_nC_2} & \cdots & C_{C_nC_n} \end{pmatrix}$$

其中：对角线元素为相同属性值的相似性，值为 1。由于 $C_{C_iC_j}$ 与 $C_{C_jC_i}$ 均代表属性值 C_i 与 C_j 间相关性，因此，$C_{C_iC_j}=C_{C_jC_i}$，即矩阵 $\boldsymbol{R}_C$ 是对角线元素为 1 的对称阵。

在每个满足 l-多样性的等价类中，不同敏感值的个数都不少于 $l(l\geqslant 2)$，这些敏感值彼此间相关性不同，依次以等价类中的每个敏感值为基准，以每个敏感值上等价类的相关性综合分析等价类的整体相关性。

定义 7 等价类的相关性。设 D 为满足 l-多样性的等价类，D 中不同的敏感值集合为$\{S_1, S_2, \cdots, S_n\}(n\geqslant 1)$，敏感值为 S_i 的记录的数量为 $n_i(1\leqslant i\leqslant n)$，等价类 D 的相关性定义为

$$R(D)=R(n_1S_1, n_2S_2, \cdots, n_nS_n)=$$

$$\frac{1}{n}\sum_{i=1}^{n}R(n_1R(S_1, S_i)S_i, \cdots, n_nR(S_n, S_i)S_i)=$$

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}n_iR(S_i, S_j)}{n\sum_{i=1}^{n}n_i}$$

其中：$R(S_i, S_j)$为属性值 S_i 与 S_j 间相关性。

2.2 (*s*,*l*)-多样性模型

本节将给出一种新的数据匿名模型，通过约束满足 l-多样性的等价类的相关性，以降低发布的等价类中的敏感值间相关性。

定义 8 (s,l)-多样性：设等价类 D 是发布表 T'中的 1 个等价类，若果 D 是 l-多样性的，且 D 的相关性不大于阈值 s，则等价类 D 满足(s,l)-多样性；若发布表 T'中的所有等价类都是满足(s,l)-多样性的，则匿名表 T'满足(s,l)-多样性。

s 为预先设置的参数，用来约束等价类中敏感值的相关程度。s 越小说明匿名表中等价类相关性越低，攻击者通过观察敏感值获得的隐私信息越少，数据安全性越高，但是随着 s 降低，数据可用性也随之降低，信息损失会比较严重；s 越大，匿名表的等价类相关性越高，数据安全将得不到保障，但 s 增大的同时，信息损失会随之减小，能够更好地保持数据完整性。因此，在实际应用中应合理选择参数 s，以便在维护数据安全性的同时最大限度地保证数据可用性。

性质 1 泛化性。数据表 T 的 2 个泛化为 G 和 G'，G 的泛化程度高于 G'，若数据表 T 基于 G'是(s,l)-多样性的，则数据表 T 基于 G 也是(s,l)-多样性的。

证明：由性质 1 条件可知，若数据表 T 基于 G'是(s,l)-多样性的，则数据表 T 基于 G'的相关性不大于 s。由于 G 的泛化程度高于 G'，泛化过程中只改变数据数据表 T 的标识符属性值，敏感属性值并未变化。根据定义 7，数据表 T 基于 G 的相关性也不大于 s。因此，数据表 T 基于 G 也是(s,l)-多样性的。

3 (*s*,*l*)-多样性聚类匿名算法

3.1 距离度量

本节使用上文提出的属性值相关性的度量方法，利用标识符属性在敏感属性上的相关性来度量元组间距离，作为聚类的依据。

定义 9 元组间距离。设数据表 T 的准标识符属性为$\{Q_1, Q_2, \cdots, Q_n\}(n\geqslant 1)$，$t_i$ 与 t_j 为 T 中的 2 条不同记录，t_i 与 t_j 在准标识符属性上的属性值分别为$[t_i^1, t_i^2, \cdots, t_i^n]$与$[t_j^1, t_j^2, \cdots, t_j^n]$，则元组 t_i 与 t_j 的距离 $d(t_i,t_j)$ 定义为

$$d(t_i, t_j)=\sum_{p=1}^{n}(1-R_{Q_p}(t_i^p, t_j^p)), i\neq j$$

其中：$R_{Q_p}(t_i^p, t_j^p)$ 为属性 Q_p 的属性值 t_i^p 与 t_j^p 在敏感属性 S 上的相关性。

定义 10 类中心。设数据表 T 的准标识符属性为$\{Q_1, Q_2, \cdots, Q_n\}(n\geqslant 1)$，等价类 D 中属性 Q_i 的不同属性值分别为$\{Q_i^1, \cdots, Q_i^{m_i}\}(1\leqslant m_i\leqslant |T|$，$|T|$为数据表 T

中元组个数)，Q_i 的每个属性值在 D 中出现的次数分别为 $\{n_i^1,\cdots,n_i^{m_i}\}$，等价类的类中心 G 定义为 $\{(n_1^1Q_1^1,\cdots,n_1^{m_1}Q_1^{m_1}),\cdots,(n_n^1Q_n^1,\cdots,n_n^{m_n}Q_n^{m_n})\}$。

表 3 中的前 3 条记录属于同一等价类，准标识符属性为 Q_1 与 Q_2。Q_1 中不同敏感属性为 Female 与 Male，在等价类中出现的次数分别为 2 和 1；Q_2 中不同敏感属性为 Married 和 Never-married，在等价类中出现次数分别为 1 和 2，因此，该等价类的类中心 G={(2Female,Male),(Married,2Never-married)}。

定义 11　元组与等价类间距离。设等价类 D 为数据表 T 中的 1 个等价类，D 的类中心 G 为 $\{(n_1^1Q_1^1,\cdots,n_1^{n_1}Q_1^{n_1}),\cdots,(n_n^1Q_n^1,\cdots,n_n^{n_n}Q_n^{n_n})\}$，元组 $t\in T$ 且 $t\notin D$，t 与 D 间的距离 $d(t, D)$定义为 t 到 D 的类中心的距离。

表 3　范例数据表 2

Table 3　Example data table2

等价类编号	Q_1	Q_2	S
1	Female	Married	Teacher
	Female	Never-married	Doctor
	Male	Never-married	Teacher
2	Female	Divorced	Teacher
	Male	Married	Doctor

3.2　信息损失量度量

数据泛化后会导致元组在准标识符属性上的精度降低，带来一定的信息损失。本文按照文献[15]提出的分析数据信息损失的方法，通过分析数据泛化前后准标识符属性值的不确定性程度的变化，来度量数据泛化后产生的信息损失。同时，将不能与其他元组一起形成满足(*s*,*l*)-多样性的等价类的元组隐匿，并使用隐匿率(suppression ratio)[16]来衡量隐匿的元组在数据表 T 中的比例。

定义 12　隐匿率。设数据表 T 共含有 n 条元组，数据表 T 中隐匿的元组数目为 n_{sr}，隐匿率定义为

$$R=n_{sr}/n$$

显然，隐匿率越小，隐匿的元组数目越少，信息损失也越少，在理想情况下，隐匿率为 0。本文将准标识符属性的信息损失与隐匿率一起作为发布数据质量的度量标准。

3.3　基于聚类的(*s*,*l*)-多样性匿名方法

(*s*,*l*)-多样性聚类匿名算法(SLCA)以聚类的方式实现等价类的(*s*,*l*)-多样性，其基本思想是：1) 按敏感值的不同将数据表划分为多个敏感集合，并构造敏感属性的相关性矩阵，敏感属性的相关性矩阵构造算法(SC_matrix constructing)的具体描述如算法 1 所示；2) 从容量最大的敏感集合 S_i 中任意选取 1 条元组作为聚类质心，根据该聚类的敏感值寻找使聚类相关性最小的敏感值加入聚类中，并更新聚类质心，重复步骤 2)，直至聚类中含有 l 条元组为止；3) 若该聚类的相关性不大于 s，则将该聚类加入待发布表 T_1 中，若相关性大于 s，则说明敏感值为 S_i 的元组无法与其他元组生成满足(*s*,*l*)-多样性的等价类，删除敏感值为 S_i 的敏感集合，重复步骤 2)和 3)，直至非空的敏感集合数目少于 l 为止；4) 将敏感集合中剩余元组分配到加入后相关性不大于 s 且距离最小的聚类中，若不存在这样的聚类，则删除该元组；5) 对每个聚类进行准标识符属性的泛化，形成满足(*s*,*l*)-多样性的匿名表。SLCA 算法的具体描述如算法 2 所示。

算法 1　敏感属性的相关性矩阵构造算法(SC_matrix constructing)

输入：数据表 T，准标识符属性{Q_1，Q_2，…，Q_q}，敏感属性 S

输出：S 的相关性矩阵 $\boldsymbol{R}_S$

步骤：

1) 初始化

{计算 S 的敏感值基数 m;

m 阶矩阵 $\boldsymbol{R}_S=\varnothing$;

按敏感值将表 T 划分为多个敏感集合{$S_1, S_2, \ldots, S_m$};}

2) For (i=1;i<=m;i++) do

{C_{ii}=1;

For (j=i+1;j<=m;j++) do

{C_{ij}=0;

For (k=1;k<=q;k++) do

{计算敏感集合 S_i 与 S_j 在 Q_k 上的关系矩阵 $\boldsymbol{P}(S_i,Q_k)$与 $\boldsymbol{P}(S_i,Q_k)$;

$C_{ij}=C_{ij}+\cos(\boldsymbol{P}(S_i,Q_k),\boldsymbol{P}(S_i,Q_k))$;}

$C_{ji}=C_{ij}$;}}

3) 返回 $\boldsymbol{R}_S$;

算法 2　(*s*,*l*)-多样性聚类匿名算法(SLCA)

输入：数据表 T，参数 s，参数 l，准标识符属性{Q_1，Q_2，…，Q_q}，敏感属性 S

输出：满足(*s*,*l*)-多样性的匿名表 T'

步骤：

1) 初始化

{计算数据表 T 中敏感属性值基数 n_s，若 $n_s<l$，则返回重新设置 l 值;

匿名表 $T'=\varnothing$;

按敏感值将数据表 T 划分为 m 个敏感集合{S_1，S_2，…，S_m};

$S_C(T,S,Q)$;}

2) While (非空集合数目多于 l) do

{初始聚类 $D=\varnothing$;

从容量最大的集合 S_i 中随机选取一条元组 t 加入 D 中;

$S_i=S_i-\{t\}$;

While (D 中元组数目不大于 l) do

{计算所有不同于 S_i 的非空集合的敏感值加入 D 后的 $R(D)$;

选择使 $R(D)$最低的集合 S_j;

在 S_j 中选择 $d(t',D)$最小的元组 t'加入 D;

$S_j=S_j-\{t'\}$;

更新 D 的质心;}

if ($R(D)<=s$) then

{$T'=T'\cup D$;}

Else

{$S_i=\varnothing$;}}

3) $T_1=\cup$ 非空集合 S_i;

4) While ($T_1\neq\varnothing$) do

{随机选取元组 t, 计算 t 加入 T'中每个等价类 EC 后的相关性;

if (存在 t 加入后相关性不大于 s 的 EC) then

{For (每个满足条件的 EC) do

{将 t 加入距离最小的 EC 中;

$T_1=T_1-\{t\}$;}}

else

{$T_1=T_1-\{t\}$;//隐匿该元组}}

5) For (每个 T'中的聚类 D) do

{泛化 D 中准标识符属性;}

6) 返回 T'。

设数据表 T 共有 n 条元组，数据表 T 中共有 q 个 Q 属性，敏感属性 S 共有 m 个不同敏感值。算法 1 敏感属性的相关性矩阵构造须计算 m^2q 次敏感值在准标识符属性上的关系矩阵，每计算 1 次关系矩阵须扫描 1 次相应敏感集合，至多需时间 $O(n)$，因此，生成敏感属性的相关性矩阵可在时间 $m^2qO(n)$内完成，由于 m 为敏感集合个数，q 为准标识符属性个数，m 和 q 都为常数，因此，算法 1 敏感属性的相关性矩阵构造的时间复杂度为 $O(n)$。

算法 2 的步骤 1)为初始化工作，该步骤判断算法的可行性，并将数据表按敏感值的不同划分多个敏感集合，同时计算敏感属性的相关性矩阵，可在 $O(n)$的时间内完成。步骤 2)首先在容量最大的敏感集合中随机抽取 1 条元组作为初始聚类的质心，然后，由生成满足(s,l)-多样性的聚类空间。通过计算将当前非空敏感集合所对应的敏感值加入聚类后的敏感属性相关性，在加入敏感值后相关性最低且不大于 s 的敏感集合中，选择距离该聚类最近的元组加入聚类中，直至聚类中元组数目达到 l 为止。每加入 1 条元组至多须计算 m 次相关性和 n 次距离，生成 1 个聚类须加入 l−1 条元组，此过程可在 $O(n^2)$的时间内完成。若执行完步骤 2)后，还剩余少于 l 个敏感集合，则通过步骤 4)将集合内的元组分配到距离最近且满足(s,l)-多样性的聚类中，每处理 1 条元组，需寻找加入后相关性不大于 s 的聚类，至多须计算 n/l 次相关性及聚类距离，此过程可在 $O(n^2)$的时间内完成。最后，通过步骤 5)泛化每个聚类中的准标识符属性，可在 $O(n)$的时间内完成。综上，算法的时间复杂度为 $O(n^2)$。

4 实验结果

4.1 实验环境

实验采用 UCI 机器学习数据库中的 Adult 数据集作为本次实验的实验数据，该数据集为美国人口普查数据，被广泛应用于基于匿名的隐私保护中。删除该数据集中包含缺失值的数据记录，处理后的数据集共有 30 162 条元组。选用{Age，Workclass，Education，Marital-status，Sex，Hours-per-week}6 个属性作为准标识符候选属性，{Occupation}作为敏感属性，{Occupation}属性具有 14 个不同属性值，各属性的具体描述如表 4 所示。

本实验通过 4 组实验对 l-diversity 模型的实现方法[6]、distinct (l,α)-diversity 模型的实现方法[12]、SLCA 3 种算法形成的高相关性的等价类的比例、算法效率与数据可用性进行比较分析。实验环境为 Inter Pentium(R) 4 CPU，3.00 GHz 处理器，2.00 GB 内存，

表 4 Adult 数据集属性描述

Table 4 Description of attributes of Adult data set

No.	Attribute	Type	Distinct values
1	Age	Numeric	74
2	Workclass	Categorical	8
3	Education	Categorical	16
4	Martial-status	Categorical	7
5	Sex	Categorical	2
6	Hours-per-week	Numeric	99
7	Occupation	Categorical	14

Microsoft Windows XP 操作系统，全部算法均在 VC++ 6.0 与 Matlab 7.0 混合编程环境下实现。

4.2　实验结果分析

4.2.1　高相关性等价类所占比例分析

图 1 所示为随 *s* 增大时 3 种算法形成的等价类中敏感值高相关性的等价类所占的比例对比。将敏感属性按相关性划分为 3 类，任何等价类中的敏感值若全部属于相同相关集合，则意味着该等价类中敏感值高度相关。由图 1 可知：*l*-diversity 与 distinct (*l*,*α*)-diversity 算法生成的高相关性等价类的比例分别为 44%与 17%，SLCA 算法生成的高相关性等价类的比例低于 *l*-diversity 与 distinct (*l*,*α*)-diversity 算法的比例，且随 *s* 增加，SLCA 算法生成的高相关性等价类的比例增加。因为 *l*-diversity 算法未约束等价类的相关性，生成的敏感值高相关性的等价类的数量较多，而 distinct (*l*,*α*)-diversity 算法仅约束敏感值的权重之和，生成的敏感值高相关性等价类数量比 *l*-diversity 算法的低，但仍然较高。*l*-diversity 与 distinct (*l*,*α*)-diversity 算法均不受 *s* 的约束，因此，*s* 的改变与否对 2 种算法生成的等价类的质量并无影响。相对于以上 2 种算法，SLCA 算法约束了等价类中敏感属性的相关性，生成的敏感值高相关性等价类的数量较少。随 *s* 增大，对等价类中敏感值相关性的约束降低，SLCA 算法生成的敏感值高相关性等价类数量增加，高相关性等价类所占的比例增加。

4.2.2　不同 *s* 下数据质量分析

本文将数据表中高相关性的等价类视为失效数据，失效数据中包含的全部元组数记作 n_{inv}，若数据表 *T* 中的元组数为 *n*(*n*>0)，*I* 为数据损失，则匿名表的数据质量度量指标 $D_{\text{QMS}}=(1-n_{\text{inv}}/n)\times(1-I)$。由 D_{QMS} 的定义可知，等价类的相关性越低，匿名表的信息损失越小，D_{QMS} 越大，匿名表的数据质量越高；反之，等价类相关性越高，匿名表的信息损失越大，D_{QMS} 越小，匿名表的数据质量越低。

图 2 所示为当 *l*=6，*Q* 属性个数由 2 增加到 6 时，不同 *s* 下 D_{QMS} 的变化情况。由图 2 可见：随 *Q* 属性个数增加，SLCA 算法的 D_{QMS} 呈下降趋势，*s* 取 0.6 时算法的信息损失与等价类的相关性达到最优平衡。这是因为：随 *Q* 属性数量的增加，算法需在更多属性上进行泛化，数据表的信息损失也随之增大，因此，D_{QMS} 整体呈下降趋势；随 *s* 增大，SLCA 算法对等价类相关性的约束降低，等价类内元组的相关性增加，失效数据也随之增多；当 *s*=0.6 时，SLCA 算法的数据质量保持在较高水平。

图 3 所示为准标识符属性个数为 5，多样性约束 *l* 由 2 增加到 11 时，不同 *s* 下 D_{QMS} 的对比。由图 3 可见：随 *l* 增加，SLCA 算法的 D_{QMS} 呈下降趋势，当 *s*=0.6 时算法达到最优平衡。这是因为：虽然失效数据随 *l* 增加而减少，但等价类中 *Q* 属性值间的差异变大，信息损失也随之增加，D_{QMS} 整体呈下降趋势；而随 *s* 增加，SLCA 算法对等价类的相关性约束降低，失效数据也随之增多；当 *s*=0.6 时，SLCA 算法的数据质量相对较高。

由实验 2 可知：*s* 取值较高时，SLCA 算法生成的等价类相关性较高，对数据隐私保护能力较弱；*s* 取值较低时，算法虽然能够有效地降低等价类的相关性，但产生的信息损失较大。因此，需合理设置 *s* 才能有效地保证数据质量。当 *s*=0.6 时，能够有效地平衡等价类的相关性与信息损失间的矛盾，得到较高质量的

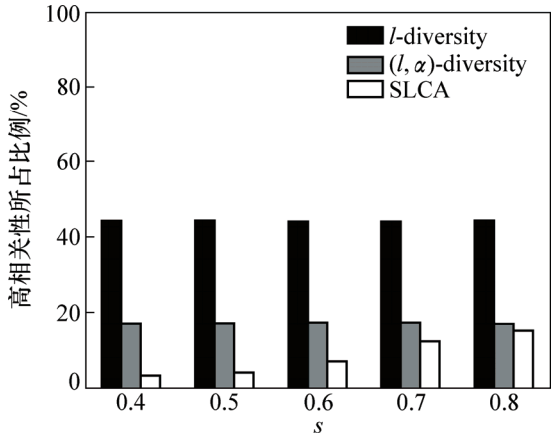
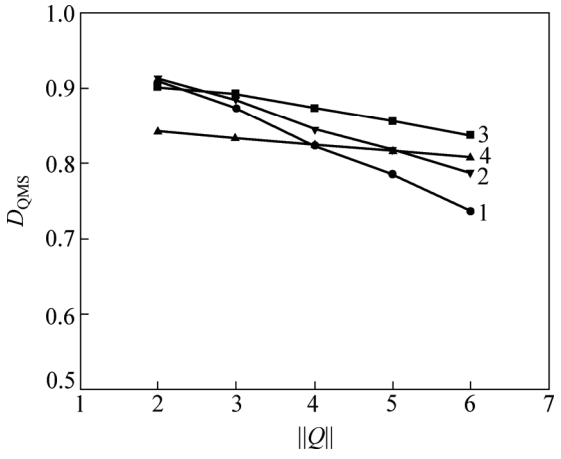


图 1　不同 *s* 下高相关性等价类的概率比较

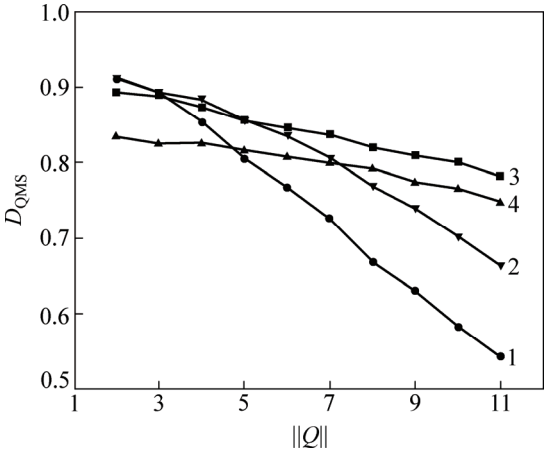
Fig. 1　Comparisons of probability of high relative equivalence class under different *s* values



s：1—0.4；2—0.5；3—0.6；4—0.7

图 2　不同 *s* 与 *Q* 属性值下数据质量比较

Fig. 2　Comparisons of data qualitative under different *s* and *Q* values



s：1—0.4；2—0.5；3—0.6；4—0.7

图 3 不同 *s* 与 *l* 下数据质量比较

Fig. 3 Comparisons of data qualitative under different *s* and *l* values

数据，以下实验中 *s* 均设置为 0.6。

4.2.3 算法效率分析

图 4 所示为 *l*=6，*Q* 属性个数由 2 增加到 6 时 3 种算法运行时间的对比。由图 4 可知：3 种算法的运行时间都随 *Q* 属性个数的增加而增加；随 *Q* 属性个数的增多，*l*-diversity 与 distinct (*l*,*α*)-diversity 算法的运行时间急剧增加，SLCA 算法的运行时间增长较缓慢。因为随 *Q* 属性个数的增多，3 种算法需泛化的属性增多，运行时间也随之增加。由于 *l*-diversity 算法通过检测 *Q* 属性的子属性集合上泛化属性值组合，以寻找实现匿名化的泛化方案，distinct (*l*,*α*)-diversity 算法将所有元组泛化到同一等价类中再采用迭代的方式逐级细化，这 2 种算法在最坏情况下的运行时间将随 *Q* 属性的增加呈指数增长。而 SLCA 算法通过度量元组与

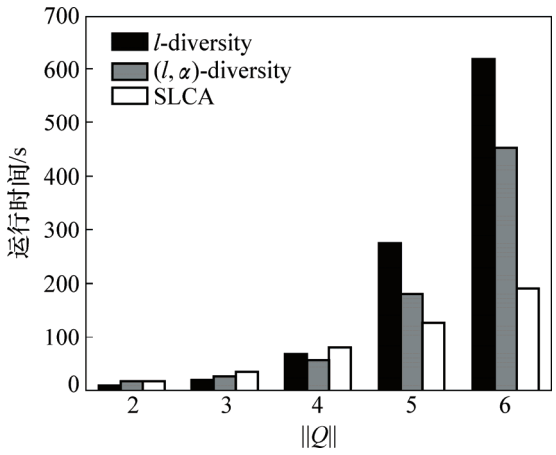


图 4 不同 *Q* 属性值下运行时间比较

Fig. 4 Comparisons of running time under different *Q* values

等价类间距离寻找最佳泛化方案，运行时间随 *Q* 属性的增加呈线性增长。因此，随 *Q* 属性个数的增多，SLCA 算法运行时间的增长幅度要比另 2 种算法的低。

图 5 所示为 *Q* 属性个数为 5，多样性约束 *l* 由 2 增加到 11 时 3 种算法运行时间的对比。由图 5 可知：*l*-diversity 算法与 distinct (*l*,*α*)-diversity 算法的运行时间随 *l* 增大而降低，SLCA 算法的运行时间随 *l* 增加而缓慢增加；当 *l*>9 时，SLCA 算法所需的运行时间比 distinct (*l*,*α*)-diversity 算法的长。这是因为 *l*-diversity 算法采取自底向上的泛化策略，*l* 的增加减少了高维子属性集上候选集数目和低维子属性集上满足需求的泛化属性值数目，因此，*l*-diversity 算法运行时间随 *l* 增加有降低的趋势。distinct (*l*,*α*)-diversity 算法采取自顶向下的泛化策略，随 *l* 增加等价类中不同敏感值增加，权值也随之增加，因此，运行时间也有降低的趋势。而 SLCA 算法每生成一个等价类需计算 *l*−1 次元组与等价类间的距离，以构成满足(*s*,*l*)-多样性的等价类，因此，构建等价类时的运行时间随 *l* 增加而增加；同时，虽然剩余的待分配元组数量随 *l* 增加而增多，但生成的等价类数量将显著减少，极大减少了元组分配时计算距离的时间，此部分的运行时间呈先增大后减小的趋势。因此，综合等价类生成与剩余元组分配 2 部分的运行时间，随 *l* 增加，SLCA 算法的整体运行时间缓慢增长。

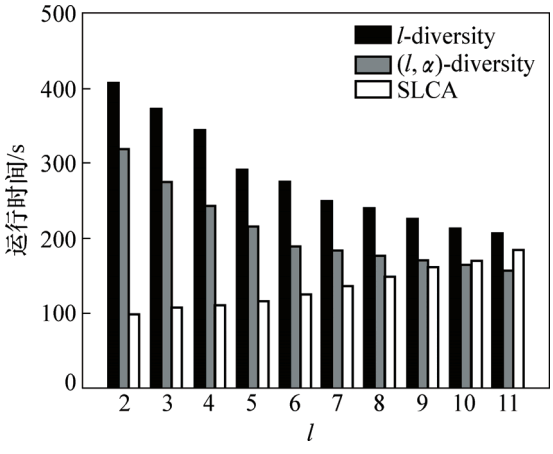


图 5 不同 *l* 时运行时间比较

Fig. 5 Comparisons of running time under different *l* values

4.2.4 信息损失分析

SLCA 算法在构建满足(*s*,*l*)-多样性的等价类时会隐匿一些元组，按照定义，将隐匿的元组数目与数据表中全部元组数目的比作为算法信息损失 *I* 的度量依据。图 6 所示为当 *l*=6，*Q* 属性个数由 2 增加到 6 时 SLCA 算法隐匿率的变化。由图 6 可知：SLCA 算法

的隐匿率不受 *Q* 属性的个数的影响。这是由于 SLCA 算法的隐匿过程只与数据表中的敏感属性有关，与 *Q* 属性无关，因此，SLCA 算法的隐匿率一直保持在 0.53%。

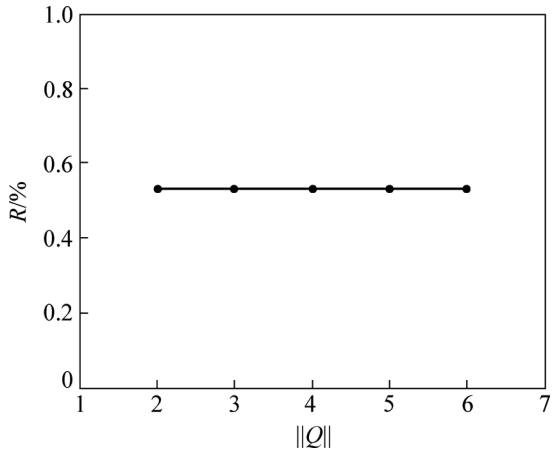


图 6　不同 *Q* 属性下隐匿率比较

Fig. 6　Comparisons of suppression ratio under different *Q* values

图 7 所示为当 *Q* 属性个数为 5，多样性约束 *l* 由 2 增加到 11 时 SLCA 算法隐匿率的对比。由图 7 可知：SLCA 算法的隐匿率随 *l* 的增加而降低。这是由于 *l* 增加，等价类的相关性降低，被隐匿的元组数量减少，因此，SLCA 算法的隐匿率随 *l* 增加而降低。

图 8 所示为当 *l*=6，*Q* 属性个数由 2 增加到 6 时，3 种算法信息损失的对比。由图 8 可知：3 种算法的信息损失都随 *Q* 属性个数的增加而增加。由于 *Q* 属性增多，3 种算法都需在更多属性上进行泛化，因此，算法的信息损失也逐渐增大。由于 *l*-diversity 算法与

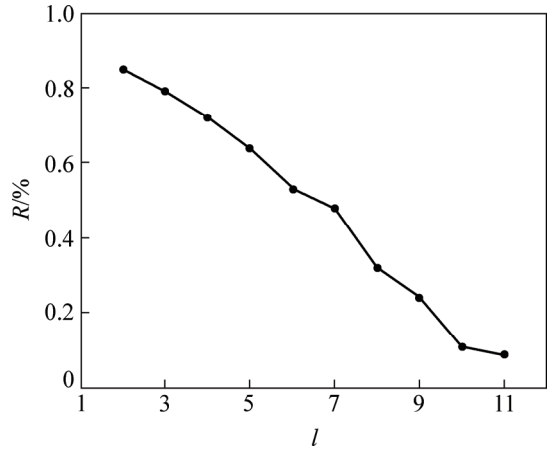
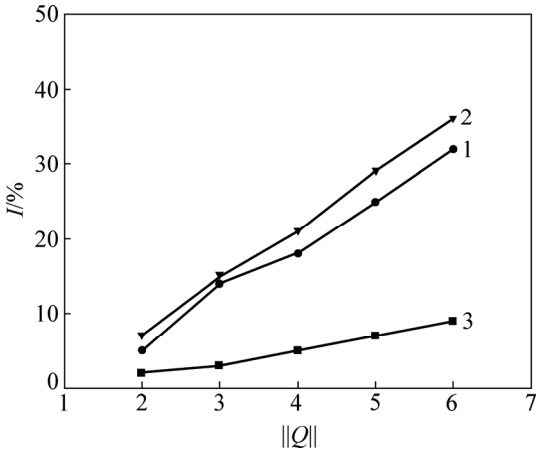


图 7　不同 *l* 下隐匿率 *R* 比较

Fig. 7　Comparisons of suppression ratio under different *l* values



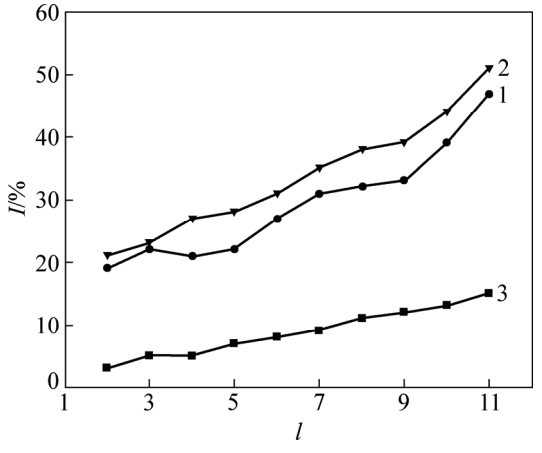
1—*l*-diversity；2—(*l*,*α*)-diversity；3—SLCA

图 8　不同 *Q* 属性下信息损失 *I* 比较

Fig. 8　Comparisons of information loss under different *Q* values

distinct (*l*,*α*)-diversity 算法采用的是全值域泛化策略，会产生不必要的过度泛化，而 SLCA 算法以聚类的方式生成等价类，能够有效降低信息损失，因此，SLCA 算法的信息损失要比另 2 种算法的低。

图 9 所示为当 *Q* 属性个数为 5，多样性约束 *l* 由 2 增加到 11 时 3 种算法信息损失的对比。由图 9 可知：3 种算法的信息损失都随 *l* 的增加而增加。这是由于 *l* 增加，3 种算法都要求等价类中包含更多的敏感值，等价类中 *Q* 属性值间的差异性增大，所产生的信息损失也随之增大，因此，3 种算法的信息损失都随 *l* 增加而增大。由于 *l*-diversity 算法与 distinct (*l*,*α*)-diversity 算法采用全局泛化策略会导致过度泛化，产生的信息



1—*l*-diversity；2—(*l*,*α*)-diversity；3—SLCA

图 9　不同 *l* 下信息损失 *I* 比较

Fig. 9　Comparison of information loss under different *l* values

损失要比 SLCA 算法的高。

5 结论

1) 针对传统 *l*-多样性模型易形成敏感值高度相关的等价类的问题，提出一种通过敏感集合中非敏感属性分布的相似性度量敏感值的相关性的方法，并在传统 *l*-多样性模型的基础上提出一种敏感值相关性约束的(*s*,*l*)-多样性模型。该模型能够有效降低等价类中敏感值的相关性，更好地防止个体敏感信息的泄露。

2) 提出(*s*,*l*)-多样性聚类匿名算法(SLCA)实现(*s*,*l*)-多样性模型。该算法通过属性相关性度量元组间距离，有效地降低泛化过程中产生的信息损失。在多样性约束 *l* 一定大时，SLCA 算法所需运行时间比 *l*-diversity 与 distinct (*l*,*α*)-diversity 算法的高；同时，SLCA 算法在生成满足(*s*,*l*)-多样性的等价类的过程中会隐匿部分元组，造成一定信息损失，但损失较低。SLCA 算法不仅具有较低的信息损失与较小的时间代价，而且能够更好地保护信息安全。

3) 在本文提出的多样性模型中，仅考虑数据集中存在单一敏感属性的问题，但现实中的数据集大多具有多个敏感属性，因此，下一步的工作是研究相关性限制的多敏感属性隐私保护问题。

参考文献：

[1] Fung B C M, Wang Ke, Chen R, et al. Privacy preserving data publishing: A survey of recent developments[J]. ACM Compute Survey, 2010, 42(4) : 1−53.

[2] Kenig B. Tassa T. A practical approximation algorithm for optimal *k*-anonymity[J]. Data Mining and Knowledge Discovery, 2012, 25 (1): 134−168.

[3] Sweeney L. Computational disclosure control: A primer on data privacy protection[D]. Massachusetts: Cambridge. Massachusetts Institute of Technology, 2001: 67−82.

[4] Sweeney L. *k*-Anonymity: A model for protecting privacy[J]. International Journal of Uncertainty Fuzziness and Knowledge Based Systems, 2002, 10(5): 557−570.

[5] Wong R C W, LI Jiuyong, Fu A W C, et al. (*α*, *k*)-anonymity: An enhanced *k*-anonymity model for privacy preserving data publishing[C]//Proceedings of the 12th ACM/SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2006: 754−759.

[6] Machanavajjhala A, Gehrke J, Kifer D. *l*-diversity: privacy beyond *k*-anonymity[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 24−36.

[7] LI Ninghui, LI Tiancheng, Venkatasubramanian S. *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity[C]//23rd International Conference on Data Engineering. Istanbul, Turkey: IEEE, 2007: 106−115.

[8] SUN Xiaoxun, SUN Lili, WANG Hua. Extended *k*-anonymity models against sensitive attribute disclosure[J]. Computer Communications, 2011, 34(4): 526−535.

[9] WANG Yunlin, CUI Yan, GENG Liqiang, et al. A new perspective of privacy protection: unique distinct *l*-SR diversity[C]//Proceedings of the Eighth Annual International Conference on Privacy Security and Trust. Ottawa, Canada: IEEE, 2010: 110−117.

[10] SHOU Lidan, SHANG Xuan, CHEN Ke. Supporting patten-preserving anonymization for time-series data[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4): 877−892.

[11] 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.
WANG Bo, YANG Jing. A personalized privacy anonymous method based on inverse clustering[J]. Acta Electronica Sinica, 2012, 40(5): 883−890.

[12] SUN Xiaoxun, LI Min, WANG Hua. A family of enhanced (*L*,*α*)-diversity models for privacy preserving data publishing[J]. Future Generation Computer Systems, 2011, 27(3): 348−356.

[13] LIU Junqiang, WANG Ke. On optimal anonymization for l^+-diversity[C]//26th International Conference on Data Engineering. New York, USA: IEEE Computer Society, 2010: 213−224.

[14] 杨高明, 杨静, 张健沛. 聚类的(*α*,*k*)-匿名数据发布[J]. 电子学报, 2011, 39(8): 1941−1946.
YANG Gaoming, YANG Jing, ZHANG Jianpei. Achieving (*α*,*k*)-anonymity via clustering in data publishing[J]. Acta Electronica Sinica, 2011, 39(8): 1941−1946.

[15] 王智慧, 许俭, 汪卫, 等. 一种基于聚类的数据匿名方法[J]. 软件学报, 2010, 21(4): 680−693.
WANG Zhihui, XU Jian, WANG Wei, et al. Clustering-based approach for data anonymization[J]. Journal of Software, 2010, 21(4): 680−693.

[16] 杨晓春, 王雅哲, 王斌, 等. 数据发布中面向多敏感属性的隐私保护方法[J]. 计算机学报, 2008, 31(4): 574−587.
YANG Xiaochun, WANG Yazhe, WANG Bin, et al. Privacy preserving approaches for multiple sensitive attributes in data publishing[J]. Chinese Journal of Computers, 2008, 31(4): 574−587.

(编辑 刘锦伟)