数据表 k-匿名化的微聚集算法研究

韩建民[1,2], 岑婷婷[2], 虞慧群[1]
(1. 华东理工大学计算机科学与工程系, 上海 200237; 2. 浙江师范大学数理与信息工程学院, 浙江金华 321004)

摘 要: 数据表的 k-匿名化(k-anonymization)是数据发布时保护私有信息的一种重要方法. 泛化/隐匿是实现 k-匿名的传统技术, 然而, 该技术存在效率低、k-匿名化后数据的可用性差等问题. 近年来, 微聚集(Microaggregation)算法被应用到数据表的 k-匿名化上, 弥补了泛化/隐匿技术的不足, 其基本思想是: 将大量的数据按相似程度划分为若干类, 要求每个类内元组数至少为 k 个, 然后用类质心取代类内元组的值, 实现数据表的 k-匿名化. 本文综述了微聚集算法的基本思想、相关技术和当前动态, 对现有的微聚集算法进行了分类分析, 并总结了微聚集算法的评估方法, 最后对微聚集算法的研究难点及未来的发展趋势作了探讨.

关键词: k-匿名; 泛化/隐匿; 微数据; 微聚集; 隐私保护
中图分类号: TP309.2 文献标识码: A 文章编号: 0372-2112 (2008) 10-2021-09

Research in Microaggregation Algorithms for k-Anonymization

HAN Jian-min[1,2], CEN Ting-ting[2], YU Hui-qun[1]
(1. *Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai 200237, China;*
2. *Mathematics, Physics and Information Engineering College of Zhejiang Normal University, Jinhua, Zhejiang 321004, China*)

Abstract: K-anonymization of tables is a method to prevent private information from disclosure prior to publication, which is achieved traditionally via generalization/suppression techniques. However, these methods have some defects on efficiency, availability, etc. Recently, microaggregation algorithm is proposed as an alternative to generalization/suppression method for k-anonymization whose goal is to cluster a set of records into groups of size at least k such that groups are as homogeneous as possible. Then the records' attribute values in the same group are replaced by the group's centroid. Microaggregation algorithms' core ideas, the state-of-the-art and related techniques are surveyed. The existing algorithms are classified and analyzed. Evaluation methods of microaggregation algorithms are investigated. Finally, some open problems and the research directions in this area are discussed.

Key words: k-anonymization; generalization/suppression; microdata; microaggregation; privacy preservation

1 引言

数据表的 k-匿名化(k-anonymization)是数据发布时保护私有信息的一种重要方法. k-匿名技术是 1998 年由 Samarati 和 Sweeney 提出[1], 它要求发布的数据中存在一定数量(至少为 k)的在准标识符上不可区分的记录, 使攻击者不能判别出隐私信息所属的具体个体, 从而保护了个人隐私, k-匿名通过参数 k 指定用户可承受的最大信息泄露风险. k-匿名化在一定程度上保护了个人的隐私, 但同时会降低数据的可用性. 因此, k-匿名化的研究工作主要集中在保护私有信息的同时提高数据的可用性.

由于 k-匿名技术简单有效, 近年来得到广泛关注. 2001 年, Samarati 采用泛化和隐匿技术实现数据表的 k-匿名化, 并提出 k-最小匿名概念[2]. 2002 年, Sweeney 研究了抵制几种攻击的 k-匿名模型[3]. 同年, 她提出了基于泛化和隐匿的 MinGen 算法[4]. IBM Waston 实验室的 Iyengar 采用基于遗传算法的不完全随机搜索方法, 解决了 k-匿名中的"组合爆炸"问题[5]. 2005 年, Yao Chao 等通过分解视图来解决视图中的 k-匿名的验证问题[6]. 2006 年, Machanavajjhala 等提出了 l-多样性算法, 它保证每个等价类中敏感信息足够多样, 以抵制同质推理攻击和背景知识攻击[7], 其中, 等价类是指匿名表中对于准标识符不可区分的记录组. Li Zude 在文献[7]的基础上, 提出了抵抗推演攻击的(k, l)模型[8], 该模型根据敏感信息的敏感程度, 事先指定每个元组的匿名度和敏感信息的多样度, 根据每个元组的(k, l)约束完成匿名化处理, 达到抵抗推演攻击的目的. Wong 等人提出了(α, k)匿名模型[9], 要求每个等价类的敏感值的频率不大于 α. Xiao XiaoKui 等提出了基于个性化匿名的泛化方

收稿日期: 2007-11-18; 修回日期: 2008-03-31
基金项目: 国家自然科学基金(No. 60773094, No. 60473055); 上海市曙光计划(No. 07SG32); 上海市浦江人才计划(No. 05PJ14030)

法[10], 该方法实现了满足个性需求的最小泛化, 减少了匿名处理的信息损失量. 杨晓春等提出了实现多约束 k-匿名的 Classfly$^+$ 算法[11]. 2007, Li Ninghui 等人提出了 3 种最优泛化模式[12], 并在文献[13] 中指出 l-多样性算法的不足, 提出了 t-closeness 框架, 该方法要求每个等价类的敏感值的分布要接近于其在原始数据表中的分布.

以上研究工作大多集中在采用泛化/隐匿技术实现数据表的 k-匿名化, 但该方法存在明显不足, 主要表现在: (1) 计算复杂性高, Meyerson[14] 和 Aggarwal 等[15] 证明, 采用泛化/隐匿获得最优 k-匿名是 NP-hard 问题. 如何将泛化和隐匿技术最优地组合还处在探讨阶段, 若没有谨慎的组合, 匿名化会损失大量信息; (2) 属性值的泛化是否合理取决于其语义, 如何确定属性的合理泛化域目前还没有可依据的方法; (3) 泛化/隐匿技术较适用于分类型数据(标称型和序数型), 对连续型数据, 泛化往往会丢失较多的数值语义[16].

为了解决泛化/隐匿技术的不足, 近来, 很多学者将 SDC(Statistical Disclosure Control) 技术中的微聚集(Microaggregation) 方法引入到数据表的 k-匿名化上, 其基本思想是: 通过某种启发式方法将数据集划分为若干类, 要求每个类至少包含 k 个元组, 类内数据最大程度地相似, 类间数据最大程度地不同, 然后用类质心来代替类内所有元组, 从而实现数据集的 k-匿名化. 由于微聚集用类质心取代类内元组的值, 故类内同质性越大, 信息的损失量越小. 微聚集算法最初用来处理连续型数据[20], 现已扩展到处理分类型数据[16,17]. 2001 年, Oganian 等证明了多变量数据的最优微聚集是一个 NP-hard 问题[18]. 2002 年, Hansen 等证明了单变量数据的最优微聚集算法可在多项式时间内实现, 其时间复杂度为 $O(k^2n)$[19]. 为降低多变量微聚集算法的时间复杂性, Domingo-Ferrer[20], Laszlo[21], Solanas[22] 等分别提出多种启发式算法. 2006 年, Domingo-Ferrer 提出了满足 (p, k) 约束的微聚集算法[23], 并应用于定位信息保护. Solanas 提出了基于遗传算法的多变量微聚集算法[24] 和可变大小的多变量 V-MDAV 算法[25]. 2007 年, Domingo-Ferrer 提出了基于树的多项式时间的 μ-Approx 算法[26]. Chin-Chen Chang 提出了 TFRP(Two Fixed Reference Points) 算法[27]. 该领域中 Josep Domingo-Ferrer 做出了很多重要的贡献. 目前国内对微聚集算法的研究比较少, 还未见相关的报告和文献.

2 微聚集算法相关技术

2.1 微聚集算法的基本概念

数据表中的属性按其所起的作用可分为四类: (1) 显示标识符, 指能清楚标识个体身份的属性, 如用户身份证号码、社会保险号、姓名等, 为了保护个人私有信息, 常常在数据发布前将这些属性删除或加密; (2) 准标识符 QI(Quasi-Identifier), 同时存在于隐私表与外表中, 可以利用链接来标识个体信息的一组属性称为准标识符[1], 如属性组{Race, Birth, Gender, Zip}. 准标识符不同于显示标识符, 它的定义依赖于攻击者所拥有的外表信息, 故表中的任何属性都有可能成为准标识符; (3) 敏感属性, 该类属性包含了个体的敏感信息, 如薪水、宗教、政治派别、身体状况等; (4) 非敏感属性, 该类属性包含了个体的非敏感信息, 对原始数据进行保护时, 该类属性不能被忽略, 因为该类属性的任意组合都有可能是准标识符. 下面给出本文用到的几个术语的定义[16].

定义 1(k-匿名) 给定数据表 $T(A_1, A_2, \cdots, A_n)$, QI 是 T 的准标识符, $T[QI]$ 为 T 在 QI 上的投影(可重复), 当且仅当在 $T[QI]$ 中出现的每组值至少在 $T[QI]$ 中出现 k 次时, 则称 T 满足 k-匿名.

定义 2(k-划分) 给定数据表 $T(A_1, A_2, \cdots, A_n)$, QI 是 T 的准标识符, QI 含 p 个属性, 表中的元组 $\boldsymbol{X}=(x_1, \cdots, x_p)$ 可视为 p-维空间中的 p-维向量点. 将表 T 基于 QI 划分为 g 个类, n_i 为第 i 类的元组数, 要求对于 $\forall i, n_i \geqslant k$, 且 $n=\sum_{i=1}^{g} n_i$, 则称该划分为数据表 T 基于 QI 的 k-划分.

定义 3(聚集) 给定数据表 $T(A_1, A_2, \cdots, A_n)$, QI 是 T 的准标识符, 基于 QI 的一个 k-划分将 T 划分为 g 个类, 设 C_i 为第 i 类的类质心, 对于所有 $i(i=1, \cdots, g)$, 用 c_i 取代第 i 类中所有元素的操作称为聚集.

最优微聚集基于最优 k-划分, 它要求划分后的类内同质性最大, 文献[20] 证明, 最优 k-划分的等价类的大小应在 $[k, 2k]$ 之间.

数据按其属性的性质可分为三种类型: (1) 连续型: 也称数值型, 指可进行算术运算且直接反映实际的物理或几何意义的属性, 如年龄、收入等; (2) 序数型: 该类属性反映特征的等级, 具有一定次序关系, 如职称, 比赛名次等; (3) 标称型: 该类属性反映某种性质, 比如颜色, 这类属性值既无数量含义, 也无次序关系. 后两类属性属于分类型属性.

不同类型的数据, 类内同质性以及聚类质心的定义不同. 下面分别说明各种类型数据的距离、类质心以及信息损失量的度量方法.

2.2 连续型数据的度量方法

连续型数据, 常用的距离度量方法是欧氏距离, 定义为式(1).

$$d(\boldsymbol{X}, \boldsymbol{Y})=\sum_{i=1}^{n}(X_i-Y_i)^2 \tag{1}$$

其中, X_i, Y_i 表示向量 $\boldsymbol{X}$, $\boldsymbol{Y}$ 第 i 维的属性值.

设 G_i 为 k-划分产生的任一等价类, 则 G_i 的类内同

质性测度 GSE 定义为式(2).

$$GSE(G_i)=\sum_{j=1}^{n_i} d(\boldsymbol{X}_{ij},\overline{\boldsymbol{X}}_i)\quad \text{其中 } \overline{\boldsymbol{X}}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{X}_{ij} \tag{2}$$

式中 n_i 表示类 G_i 中的元组个数, $n_i\geqslant k$, $\boldsymbol{X}_{ij}$ 表示 G_i 类中的第 j 条元组, $\overline{\boldsymbol{X}}_i$ 表示 G_i 类的类质心, GSE 越小, 类内数据的同质性越强.

所有类的同质性测度 SSE 定义为式(3).

$$SSE=\sum_{i=1}^{g} GSE(G_i)=\sum_{i=1}^{g}\sum_{j=1}^{n_i} d(\boldsymbol{X}_{ij},\overline{\boldsymbol{X}}_i) \tag{3}$$

式中 g 为类的个数, $n=\sum_{i=1}^{g} n_i$.

数据表的整体同质性测度 SST 定义为式(4).

$$SST=\sum_{i=1}^{g}\sum_{j=1}^{n_i} d(\boldsymbol{X}_{ij},\overline{\boldsymbol{X}})\quad \text{其中 } \overline{\boldsymbol{X}}=\frac{1}{n}\sum_{l=1}^{n}\boldsymbol{X}_l \tag{4}$$

其中 $\overline{\boldsymbol{X}}$ 表示为整个数据表 T 的平均向量, 即整个数据集的中心.

连续型数据的类质心计算比较简单, 一般可用第 i 类元组的均值向量 $\overline{\boldsymbol{X}}_i$ 来定义该类的类质心.

对给定的数据集 T 而言, SST 固定不变, 但不同的 k-划分, 会导致不同的 SSE, 类内同质性越大, SSE 会越小, 信息损失量也会越小, 因此连续型的数据信息损失量 IL 可用式(5)度量.

$$IL=SSE/SST \tag{5}$$

其中 IL 越小, 匿名表的数据可用性越强. 4.3 节中讨论了信息损失量的其它度量方法.

2.3 分类型数据的度量方法

分类型数据主要包括: 序数型和标称型. Domingo-Ferrer 分别给出了它们距离度量方法[16].

对于序数型属性 A_i 的不同值 a, $b(a\leqslant b)$ 之间的距离定义为式(6).

$$d(a,b)=|\{j\mid a\leqslant j<b\}|/|D(A_i)|,\quad \text{其中 } D(A_i) \text{ 为 } A_i \text{ 的取值域} \tag{6}$$

由式(6)可知, $d(a,b)$ 为属性 A_i 在 $[a,b]$ 之间的值的个数与属性 A_i 所有值个数的比值.

Domingo-Ferrer 采用中位数或凸中位数定义序数型数据的类质心, 计算方法如下:

设序数型属性 A_i 取值集 $C=\{c_1<c_2<\cdots<c_o\}$, n 个数据 $\{a_1,a_2,\cdots,a_n\}(a_i\in C)$ 组成的一个类 S, 中位数指将 S 排序后的中间位置的元素. 如: 某序数型属性的等级为 $C=\{0,1,2,3,4,5,6,7\}$, 则 $S=\{1,2,2,5,6\}$ 的中位数为 2.

凸中位数的确定需要定义频率函数 f', 设 $f(c_j)$ 为等价类 S 中 c_j 出现的频率, 则 f' 定义见式(7).

$$f'(c_i)=\min\left(\max_{c_j\leqslant c_i}(f(c_j)),\max_{c_j\geqslant c_i}(f(c_j))\right) \tag{7}$$

凸中位数的计算基于频率函数 f'. 例如: 设 $C=\{0,1,2,3,4,5,6,7\}$, $S=\{1,2,2,2,7\}$, 由中位数定义知, S 的中位数为 $\{2\}$. 再由式(7)得到 C 中元素 $\{1,2,3,4,5,6,7\}$ 的频率值为: $\{1,3,1,1,1,1,1\}$, 则对应的元素集合为 $\{1,2,2,2,3,4,5,6,7\}$, 故 S 的凸中位数为 3. 凸中位数法比中位数法较多地保留了序数型数据的语义.

对于标称型属性, Domingo-Ferrer[16]给出了简单的距离定义, 令 $\boldsymbol{X}$, $\boldsymbol{Y}$ 为具有 p 个标称型属性的元组, 则二者之间的距离定义为式(8).

$$d(\boldsymbol{X},\boldsymbol{Y})=\sum_{j=1}^{p}\delta(x_j,y_j),\quad \text{其中 } \delta(x_j,y_j)=\begin{cases}0, & x_j=y_j\\ 1, & x_j\neq y_j\end{cases} \tag{8}$$

标称型数据往往有一定的语义层次关系, 式(8)没有考虑到这一点, 因此采用式(8)来度量标称型数据的距离存在不合理的情况, 如邮政编码属性, $\{151400\}$ 与 $\{151411\}$ 的距离比 $\{151400\}$ 与 $\{321004\}$ 的距离更近, 而按式(8)计算, 它们是等距的. 可借鉴文献[28]定义的层次距离来度量标称量的距离, 以改进 k-划分的质量.

标称型数据类质心的选取是基于频率的, 其属性值可确定为该等价类中各属性出现频率最高的值, 即各属性的众数.

分类型数据匿名化的信息损失量的度量方法将在 4.3 节讨论.

2.4 混合型数据的度量方法

混合型数据指数据集中既有连续型属性, 又有分类型属性. 目前面向混合型数据的微聚集算法研究的比较少. 可借鉴文献[29, 30]中混合型数据的距离度量方法来实现混合型数据的微聚集算法.

2.5 k-匿名化微聚集算法的基本步骤

实现 k-匿名化的微聚集算法分为 2 个步骤: 第一步, 对数据表 T 进行 k-划分, 生成由 g 个类组成的数据

表 1　原始数据表 T

Company	Surface	No. emp	Turnover	Net pro
A&A Ltd	790	55	3212334	313250
B&B SpA	710	44	2283340	299876
C&C Inc	730	32	1989233	200213
D&D BV	810	17	984983	143211
E&E SL	950	3	194232	51233
F&F GimbH	510	25	119332	20333
G&G AG	400	45	3012444	501233
H&H SA	330	50	4233312	777882
I&I LLC	510	5	159999	60388
J&J Co	760	52	5333442	1001233
K&K Sarl	50	12	645223	333010

表;第二步,对类进行聚集操作,得到由 g 个等价类组成的新数据表 T'. 例如:原始数据表 T,见表 1,显示标识符为{Company},准标识符 QI = {Surface, No. emp},敏感属性为{Turnover, Net profit}. k-匿名化的过程如下:(1)删除显示标识符 Company,数据被初步匿名化;(2)将数据表 T 的 QI 属性值标准化,再基于 QI 进行 k-划分(k= 3);(3)将标准化的值恢复为原数值,对 k-划分的数据表进行聚集操作(用平均值作为类质心),得到三个等价类,见表 2.

表 2 表 T 微聚集后的匿名表 T', k= 3

Company	Surface	No. emp	Turnover	Net pro
	747.5	46	3212334	313250
	747.5	46	2283340	299876
	747.5	46	1989233	200213
	756.67	8	984983	143211
	756.67	8	194232	51233
	322.5	33	119332	20333
	322.5	33	3012444	501233
	322.5	33	4233312	777882
	756.67	8	159999	60388
	747.5	46	5333442	1001233
	322.5	33	645223	333010

3 微聚集算法的分类

3.1 从 k-划分所依据的属性个数的角度分类

从 k-划分所依据的属性个数的角度,微聚集算法可分为单变量微聚集算法和多变量微聚集算法两大类. 单变量微聚集算法是指以准标识符的单个属性为依据进行 k-划分,方法有单轴排序法[31]、遗传算法[20]等. 多变量微聚集算法以准标识符的多个属性为依据进行 k-划分,尽管多变量数据集的最优 k-划分是 NP 难题,但采用启发式方法[16, 20, 21, 26],也可获得高效的近似解. 文献[20]将 Ward 算法改进,提出 k-Ward 算法来实现 k-划分. 并在 k-Ward 算法基础上提出了 MD(Maximum Distance)算法,用距离最远的两个向量作为初始类的中心,进行聚类,获得了较好的划分效果,但其初始中心选择的时间复杂性较高,MD 算法的时间复杂度为 $O(n^3/k)$. μ-ARGUS[32]软件包中的距中心点最大距离 MDAV(Maximum Distance to Average Vector)算法是 MD 算法的改进,将时间复杂性降为 $O(n^2)$. 文献[16]在 MDAV 算法的基础上,定义了分类型数据的距离度量方法,提出了适用于连续型和分类型两种属性的 MDAV-generic 算法,时间复杂性也为 $O(n^2)$.

排序是实现微聚集较简单的方法,最初是针对单变量微聚集提出的,后扩展到多变量微聚集. 下面说明基于排序思想的微聚集算法[31].

(1)单轴排序算法

单轴排序主要思想:先选择数据表 T 的某属性 A,根据属性 A 取值的大小将表 T 中的记录进行排序,然后依排序顺序进行 k-划分,最后对 k-划分的结果进行聚集操作. 该算法可在多项式时间内实现.

单轴排序算法在处理多变量数据表时,排序变量选择的不同往往导致不同的排序结果,最后 k-划分的结果也就不一样,因此排序变量的选择是一个关键的问题.

(2)第一主成分排序算法

第一主成分排序法解决了单轴排序算法中排序变量的选择问题. 主成分技术最早由 Karl Pearson 提出[33],该技术能对给定的 p 个相关变量 $X_1, X_2, \cdots, X_p$,生成实测变量的线性不相关集合 $Z_1, Z_2, \cdots, Z_p$,其中,Z_1 具有最大方差,随后的 Z_i 方差依次减少,并且彼此线性无关. 第一主成分排序算法采用第一因子 Z_1 对数据表进行排序,其计算方法见式(9).

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \tag{9}$$

其中 $a_{11}^2 + a_{12}^2 + a_{1p}^2 = 1$, $a_{11}, a_{12}, \cdots, a_{1p}$ 为对应特征向量.

该方法依赖于所有可用变量的线性组合,故理论上产生的结果优于单轴排序.

(3)Sum of Z-Scores 方法

Sum of Z-Scores 方法可以实现多变量微聚集,其步骤类似于第一主成分排序法,差别之处是跟据 ZSS 值进行排序. 给定 p 个变量 $X_1, X_2, \cdots, X_p$, ZSS 的计算方法见式(10).

$$ZSS = \sum_{i=1}^{n}\sum_{j=1}^{p}(X_{ij} - \bar{X}_j)/\sigma_j \tag{10}$$

式中 X_{ij} 为第 j 个变量的第 i 个具体值,$\bar{X}_j$ 为第 j 个变量的均值,σ_j 为第 j 个变量的标准偏差.

(4)单独排序法(Individual Ranking)

单独排序法指对多变量数据表进行 k-划分时,认为变量之间相互独立,并对每个变量采用单轴排序法进行微聚集. 尽管处理的结果对每个变量而言是 k-匿名化的,但该结果对多个变量来说未必是 k-匿名的. 该方法效率高,类似于单轴排序法,但只适合多个变量相互独立的情况.

3.2 从等价类大小的角度分类

从 k-划分时各等价类中元素个数的角度,算法可分为定长微聚集算法和变长微聚集算法,Josep Domingo-Ferrer 把变长微聚集算法称为面向数据的微聚集算法. 定长微聚集算法[16, 20, 21, 26, 31]在构建类时,除了最后一个类的大小处于 k 和 $2k$ 之间外,其余类的大小均为 k. 基

于排序思想的算法[31], MD 算法[20], MDAV 算法[32]、MDAV-generic 算法[16], 都属于定长微聚集算法. 其中, MDAV 算法是性能较好的算法, 时间复杂度为 $O(n^2)$, 其描述见图 1. 算法中数据集的中心点为数据集各属性的均值.

MDAV 算法

输入: 包含 n 个记录的数据表 T, 匿名参数 k.

输出: k-匿名化的数据表 T'.

步骤: (1) 计算数据集的中心点 $\bar{x}$, 找到距离 $\bar{x}$ 最远的记录 r, 再找到距离 r 最远的记录 s.

(2) 以 r 为中心, 选择离 r 最近的 $k-1$ 个记录组成一个类; 以 s 为中心, 选择离 s 最近的 $k-1$ 个记录组成一个类.

(3) 若剩下记录数 $\geqslant 2k$, 则对剩余的记录重复执行(1)(2).

(4) 若剩下记录数在 $[k, 2k-1]$ 之间, 则这些记录自成一类. 否则, 将剩余记录分别加入到离各自最近的类.

图 1　MDAV 算法描述

定长微聚集算法尽管执行效率高, 但往往导致错误的聚类结果, 例如: 对图 2 九个点的数据集实现最优 k-划分, 若采用定长微聚类算法, $k=3$, 可得到三个类的数据集, 见图 3. 显然这样的划分结果并不是最优的, 该数据集应划分为两类, 见图 4.

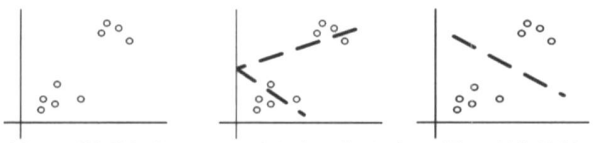

图2　原始数据集　　图3　定长微聚集结果　　图4　最优结果

变长微聚集指数据集进行 k-划分后, 产生类的大小在 $[k, 2k]$ 之间, 变长微聚集的质量高于定长微聚集. 1998 年, Domingo-Ferrer 在研究报告里提出了两种变长微聚集算法: k-Ward 算法和遗传算法, 这两个算法后来发表在文献[20]. 2005 年, Laszlo 和 Mukherjee 提出了变长的最小生成树(MST)算法[21], 尽管实验证明 MST 算法产生的信息损失量和时间复杂度并没有低于 k-Ward 算法, 但 MST 提出了变长划分的新思路. 2006 年, Domingo-Ferrer 将 Hansen-Mukherjee 在 2003 年提出的单变量的最短路径的微聚集算法[19]扩展到多变量[34], 提出了多变量的变长微聚集算法. 同年, 在 MDAV 算法基础上, Solanas 等人提出了 V-MDAV(Variable-MDAV)算法[25], 该算法中类的初次建立类似 MDAV 算法, 建类后按约束条件对类进行扩充, 使类内元素个数处于 $[k, 2k]$ 之间, 以达到较好聚类效果. 2007 年, Chin-Chen Chang[27]提出了 TFRP 算法, 将算法的运行时间缩短到 $O(n^2/k+nk)$, 该算法首先采用定长微聚集快速对数据表进行 k-划分, 然后按类内同质性由低到高有选择地重新分配类内的元素, 来降低信息损失量.

3.3　从聚类方法角度分类

聚类是目前发展比较成熟且应用较为广泛的数据分析方法, 微聚集与聚类很相似, 很多微聚集算法都基于某种聚类思想. 文献[20]提出的 k-Ward 算法是基于层次聚类思想的. 而文献[17]提出的基于 k-modes 算法以及文献[35]提出的模糊 c-均值算法则是基于划分聚类思想的. 文献[21]提出的最小生成树(MST)微聚集算法是基于最小生成树聚类思想的.

下面介绍 MST 微聚集算法[21], 算法中树的结点表示数据集的记录, 边 $e=(org, dest)$ 的权值表示结点 org 与其父结点 $dest$ 的距离. 算法用树来表示类, 通过删除树的边将大树分解为 2 个子树, 且要求两个子树的大小不能小于 k, 算法描述见图 5. 该算法的时间复杂度为 $O(n^2)$.

MST 算法

输入: 包含 n 个记录的数据表 T, 匿名参数 k.

输出: k-匿名化的数据表 T'.

步骤: (1) 最小生成树的构造

将记录看作结点, 定义结点间距离, 采用 Prim 算法构造最小生成树.

(2) 删除边

按边的权值大小循环处理每条边 $e=(org, dest)$

若删除边 e 生成的两个子树的大小均大于 $k-1$, 则删除边 e; 否则保留该边.

(3) 生成类: 每个子树为一个类.

图 5　MST 微聚集算法

聚类与微聚集也有所不同, 主要表现在: (1) 聚类算法可事先确定要生成的类的数目, 而微聚集算法, 无法事前给定; (2) 聚类算法没有类大小的约束, 微聚集则要求类内元组数不能少于 k 个; (3) 聚类算法不需要修改类内元组值, 而微聚集算法最终需要用类质心取代类内元组值.

3.4　从数据类型的角度分类

从数据类型的角度, 微聚集算法可分为连续型数据的微聚集算法、分类型数据的微聚集算法和混合型数据的微聚集算法. 目前微聚集算法的主要研究成果是处理连续型数据[16, 20, 21, 26].

分类型数据的微聚集算法研究较少[16, 17], 可以针对分类型数据定义其距离度量方法, 将基于距离聚类的微聚集算法, 改造后应用到分类型数据的微聚集中. 文献[16]的 MDAV-generic 算法是 MDAV 算法[32]的扩展, 既可处理连续型数据, 也可处理分类型数据, 但该方法没有考虑混合型数据, 其时间复杂度为 $O(n^2)$. Torra V[17]针对分类型数据提出了基于 k-modes 的微聚集算法, 该算法首先构造属性值的频率集, 根据原数据集和频率集计算初始类中心个数, 再利用 k-modes 算法将数据集划分为若干类, 最后调整类大小以保证每个类元素的个数大于等于 k. 该算法的时间复杂度为 $O(n^2)$. 初始类个数的确定以及类大小的调整策略是算法实现的

关键, 这两个问题文献[17]都没有给出较好的解决办法.

混合型数据的微聚集算法更不多见, 将聚类领域中混合型数据的距离度量方法引入到微聚集算法来实现混合型数据的微聚集是一个比较好的途径.

4 微聚集算法评估

4.1 评估指标

微聚集算法的评估指标包括: 算法复杂性、k-匿名化后数据安全性和可用性.

算法的复杂性一般用时间复杂性来度量, 文献[18]研究了最优微聚集算法的时间复杂性, 并证明多变量微聚集是一个NP难题, 本文对这一指标度量方法不多做探讨.

数据安全性指标, 也称泄密风险指标, 用于衡量匿名表对敏感信息的保护程度, 泄密风险可以采用匿名表中的记录与原数据表同一记录的关联程度来度量. 如果攻击者根据匿名表中的记录可较大概率地推导出原数据表中的记录, 则该匿名表是不安全的. 4.2节将研究该指标的度量方法.

数据可用性指标用于衡量匿名表保持原始数据特征的程度. 数据的可用性与 k-匿名化过程中信息的损失量是相对应的, 当信息损失量较大时, 匿名化后数据的可用性就差, 反之亦然. 因此可用信息损失量来度量数据的可用性. 4.3节详细研究该指标的度量方法.

需要说明的是, 匿名表数据的安全性和可用性是两个矛盾的指标, 安全性的加强会增加信息损失量, 降低数据可用性. 如何权衡好这两个指标, 是微聚集算法的一个挑战性课题. 文献[16]研究了数据安全性和可用性的权衡技术, 提出了加权平均法和R-U图法. 但该问题没有得到根本的解决.

4.2 数据安全性度量

Panaretos[31]提出度量数据安全性的3种方法: (1)阈值 k; (2)(n, k)规则; (3)数据扰动指示. 数据扰动指示通过匿名化前后的记录变化程度 $d(j)$ 来度量连续型数据微聚集的安全性, 见式(11).

$$d(j)=\left|\frac{x_j^m-x_j^o}{x_j^o}\right| \tag{11}$$

式中 x_j^m 表示匿名化后记录j 对应的属性x 的值, x_j^o 表示匿名化前记录j 对应的属性x 的值. 数据集的记录个数比较多时, 若 $d(j)$ 均低于一定阈值, 则表明该算法产生的匿名表不够安全. 对于多变量的情况, 可以采用 $d(j)$ 加权平均的方法度量其安全性.

数据安全性度量方法还有[36, 37]: 基于记录链接泄密度量方法、区间泄密度量方法. 基于记录链接的方法又分为基于概率的记录链接方法和基于距离的记录链接方法. 区间泄密方法分为基于分级的区间泄密度量方法和基于标准偏差的区间泄密度量方法. 这些方法中, 基于距离的记录链接方法比较简单、实用. 为说明该方法, 引入以下术语.

定义4(链接成功) 匿名表中的任一记录 R, 计算 R 到原始表所有记录的距离, 得到距离 R 最近的记录集 DS_1、次最近记录集 DS_2, 若 DS_1(或者 DS_2)中存在 R 的原始记录, 则称记录 R 链接成功.

基于距离的记录链接度量方法是用匿名表中链接成功的记录所占的比率来度量泄密风险, 设 $linked\_records$ 为匿名表中链接成功的记录数, $total\_records$ 为匿名表中总的记录数, 则泄密风险 DRL 的度量方法见式(12), 该方法适用于所有数据类型.

$$DRL=\frac{linked\_records}{total\_records} \tag{12}$$

4.3 数据可用性度量

数据可用性可采用基于统计量变化的信息损失量来度量. 不同类型的数据采用不同的统计量: 连续型数据可使用均值、方差和皮尔逊相关系数等; 序数型数据可采用中位数(median)、众数(mode)和信息熵等; 标称型数据则采用众数和信息熵等. 不同的统计量从不同角度反映了数据表匿名前后的变化, 因此可对基于多种统计量的信息损失量进行加权平均, 得到较全面的信息损失量度量结果.

目前, 常用的信息损失量度量方法有以下几种:

(1)连续型数据的信息损失量度量方法

文献[20]给出的信息损失量的度量方法是一种比较简单的连续型数据的度量方法, 见式(5). 该方法是从等价类的同质性角度来度量匿名表的信息损失量.

文献[37, 38]提出了三种基于统计量的信息损失量的度量方法: 均方误差 MSE、绝对平均误差 MAE、平均偏差 MV. 设数据表 T 含 p 个属性, n 个记录, 表 T' 为 T 的匿名表, 表 T' 含 p' 个属性, n' 个记录, $\boldsymbol{X}$, $\boldsymbol{X}'$ 分别为 T, T' 的数据矩阵, x_{ij}, x'_{ij} 分别为数据矩阵 $\boldsymbol{X}$, $\boldsymbol{X}'$ 中对应值, 三种信息损失量度量方法分别见式(13)、(14)、(15).

$$MSE=\frac{1}{pn}\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2 \tag{13}$$

$$MAE=\frac{1}{pn}\sum_{j=1}^{p}\sum_{i=1}^{n}|x_{ij}-x'_{ij}| \tag{14}$$

$$MV=\frac{1}{pn}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{|x_{ij}-x'_{ij}|}{|x_{ij}|} \tag{15}$$

比较的参数量除了可以使用匿名前后的数据矩阵以外, 还可以比较匿名前后数据矩阵的各个属性的均值、协方差、相关系数矩阵等, 其计算方法参见文献[37, 38].

Yancey 等[39]指出当 x_{ij} 接近 0 时, MV 值会急剧增长, 并提出了改进的 MV 测度公式, 见式(16).

$$MV' = \frac{1}{pn}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{x_{ij} - x'_{ij}}{\sqrt{2}S_j} \quad (16)$$

式中 S_j 为原始表第 j 个属性的标准偏差.

文献[40]定义了 5 种基于概率的信息损失量度量方法, 这五种方法分别是基于均值、方差、协方差、皮尔逊相关系数和分位数的.

(2) 分类型数据的信息损失量度量的方法

分类型数据的信息损失量的度量可以采用以下三种方法:

(a) 直接比较法

直接比较法比较匿名化前后数据表之间的距离. 数据表的距离指两数据表中对应记录距离和. 记录之间距离的度量方法参见 2.3 节. 设 T' 为 T 的匿名表, t、t' 分别为 T、T' 中的记录, 即 $t \in T$, $t' \in T'$, 且 t' 是 t 的匿名记录, n 为 T、T' 中记录数, 则信息损失量定义为式(17).

$$IL = \sum_{i=1}^{n} d(t_i, t'_i) \quad (17)$$

(b) 列联表比较法

列联表比较法是比较匿名化前后的数据列联表. 文献[37]给出了匿名前后列联表的差的计算方法 *CTBIL* (Contingency Table Based Infomation Loss measure). 设表 T' 为表 T 的匿名表, N 为其维数, W 是它们的 t 维列联表, 且 $t < N$, 则信息损失量 *CTBIL* 定义为(18)

$$CTBIL(T, T'; W, N) = \sum_{\substack{\{V_{j1}\cdots V_{jt}\} \subseteq W \\ |\{V_{j1}\cdots V_{jt}\}| \leqslant N}} \sum_{i_1\cdots i_t} \left| x^{T}_{i_1\cdots i_t} - x^{T'}_{i_1\cdots i_t} \right| \quad (18)$$

式中 $x^{file}_{subscripts}$ 表示 *file* 对应的列联表在 *subscripts* 位置处的条目.

当数据集的属性个数或属性取值类别数目比较多时, 列联表方法的时间和空间开销都比较大.

(c) 基于信息熵的度量方法

如果把数据集的属性看作随机变量, 那么每个属性的信息熵反映了属性取值的不确定性程度, 匿名化后, 属性取值的不确定性程度发生了变化, 信息熵也会变化. 信息熵比较法是通过比较匿名化前后数据集信息熵的变化程度来度量信息损失量.

文献[31]分别给出了标称型和序数型信息熵的定义. 设数据表 T 含 n 个记录, T 微聚集后产生含 g 个等价类的匿名表 T'. 数据集 T 基于属性 i 可划分为 L 个等价类 $C_i (i = 1, \cdots, L)$, n_i 是类 C_i 中的样本数, 标称型第 j 个属性信息熵定义为式(19), 序数型第 j 个属性信息熵定义为式(20).

$$H_j = \left(-\sum_{i=1}^{L} p_i \log_2 p_i \right) \quad (19)$$

$$H_j = -\sum_{i=1}^{L-1} [\, p_i \log_2 p_i + (1 - p_i)\log_2(1 - p_i)\,] / (L - 1) \quad (20)$$

其中 p_i 为元组属于等价类 C_i 的概率, $p_i = n_i / n$.

若数据表由 t 个属性组成, 则整个数据表的信息熵定义为式(21).

$$H = \sum_{j=1}^{t} H_j \quad (21)$$

基于以上定义, 匿名表的信息损失量可定义为式(22).

$$ILR = \frac{|\, original\ Entropy - new\ Entropy\,|}{original\ Entropy} * 100 \quad (22)$$

5　微聚集算法面临的问题和发展趋势

微聚集算法是近年发展起来实现数据集 k-匿名化的热点技术, 尽管目前已经出现了很多突破性的成果, 但仍存在很多问题亟待解决, 面临的主要问题有:

(1) 如何平衡好匿名表的安全性和可用性. 数据安全性和可用性是两个矛盾的指标, 如何更好地平衡好这两个指标, 得到既安全又可用的匿名表是微聚集算法需要研究的关键问题.

(2) 高维空间的微聚集算法研究. 高维数据给聚类分析带来二大问题: 第一, 不相关属性削弱了数据会聚的趋势. 第二, 低维中很有效的区分数据的标准在高维空间中失效了, 从而导致根据接近度划分类的结果是不可信的. 基于聚类思想的微聚集算法在高维空间中, 也面临同样的问题.

(3) 大样本微聚集算法的研究. 实际应用要求微聚集算法能够处理大样本数据集, 而目前的很多效果比较好的算法无法高效地处理大样本数据集, 故如何将算法拓展到大样本数据集是微聚集算法面临的另一个问题.

目前微聚集算法的研究工作除了围绕以上几个问题以外, 还表现出下面几个发展趋势:

(1) 变长微聚集算法的研究. 定长微聚集往往得到不合理的 k-划分, 变长微聚集算法会大大提高 k-划分的质量, 是微聚集算法的发展趋势.

(2) 混合型数据的微聚集算法的研究. 如何定义混合型数据的距离度量方法、类质心以及信息损失量, 以实现混合型数据的 k-匿名也将是今后研究的热点.

(3) 抵抗同质推理攻击和背景知识推理攻击的微聚集算法的研究、分布式数据集微聚集算法的研究以及增量式数据集微聚集算法的研究都是今后发展的新方向.

参考文献:

[1] Samarati P, Sweeney L. Generalizing data to provide anonymity

when disclosing information (abstract) [A]. Proc of the 17th ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems[C]. Seattle, WA, USA: IEEE press, 1998. 188.

[2] Samarati P. Protecting respondents' identities in microdata release[J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(6): 1010-1027.

[3] Sweeney L. K-anonymity: a model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[4] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression[J]. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 571-588.

[5] Iyengar V. Transforming data to satisfy privacy constraints[A]. Proc of the 12th ACM SIGKDD Conference[C]. Edmonton, Alberta, Canada: ACM Press, 2002. 279-288.

[6] Yao C, Wang X S, Jajodia S. Checking for k-anonymity violation by views[A]. Proc of the 31st International Conference on Very Large Data Bases[C]. Trondheim, Norway: VLDB Endowment, 2005. 910-921.

[7] Machanavajjhala A, Gehrke J, Kifer D. L-diversity: privacy beyond k-anonymity[A]. Proc of the 22nd International Conference on Data Engineering[C]. Atlanta, GA, USA: IEEE Press, 2006. 24-36.

[8] Zude L, Guoqiang Z, Xiaojun Y. Towards an anti-inference (k, l)-anonymity model with value association rules[A]. Conference on Database and Expert Systems Applications[C]. Krakow, Poland: Springer-Verlag, 2006. 883-893.

[9] Wong C R, Li J, Fu A, et al. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing[A]. Proc of the 12th ACM SIGKDD Conference[C]. Philadelphia, PA: ACM Press, 2006. 754-759.

[10] Xiaokui Xiao, Yufen Tao. Personalized privacy preservation [A]. ACM Conference on Management of Data (SIGMOD) [C]. Chicago, Illinois, USA: ACM Press, 2006. 229-240.

[11] 杨晓春, 刘向宇, 王斌, 于戈. 支持多约束的 K-匿名化方法[J]. 软件学报, 2006, 17(5): 1222-1231.
Yang Xiaochun, Liu Xiangyu, Wang Bin, Yu ge. K-Anonymization Approaches for Supporting Multiple Constraints [J]. Journal of Software, 2006, 17(5): 1222-1231. (in Chinese)

[12] Tiancheng Li, Ninghui Li. Towards optimal k-anonymization [J]. Data and Knowledge Engineering. 2008, 65(1): 22-39.

[13] Li Ninghui, Li Tiancheng, Venkatasubramanian S. t-Closeness: privacy beyond k-anonymity and l-diversity[A]. Proc of the 23rd ICDE[C]. Istanbul Turkey: IEEE press, 2007. 81-90.

[14] Meyerson A, Williams R. On the complexity of optimal k-anonymity[A]. Proc of the 23rd ACM Symposium on Principles of Database Systems[C]. Paris, France: ACM Press, 2004. 223-228.

[15] Aggarwal G, Feder T, Kenthapadi K, et al. K-anonymity: algorithms and hardness[R]. Technical Report, Stanford University, 2004.

[16] Domingo-Ferrer J, Torra V. Ordinal, continuous and heterogeneous k-anonymity through microaggregation[J]. Journal of Data Mining and Knowledge Discovery, 2005, 11(2): 195-202.

[17] Torra V. Microaggregation for categorical variables: a median based approach [A]. Workshop on Privacy in Statistical Database[C]. Barcelona, Catalonia, Berlin: Springer-Verlag, LNCS 3050, 2004. 162-174.

[18] Oganian A, Domingo-Ferrer J. On the complexity of optimal microaggregation for statistical disclosure control[J]. Statistical Journal of United Nations Economic Commission for Europe, 2001, 18(4): 345-354.

[19] Hansen S L, Mukherjee S. A polynomial algorithm for optimal univariate microaggregation[J]. IEEE Transaction Knowledge and Data Engineering, 2003, 15(4): 1043-1044.

[20] Domingo-Ferrer J, Mateo-Sanz J M. Practical data-oriented microaggregation for statistical disclosure control[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(1): 189-201.

[21] Laszlo M, Mukherjee S. Minimum spanning tree partitioning algorithm for microaggregation[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(7): 902-911.

[22] Solanas A, Martinez-Balleste A, Domingo-Ferrer J, et al. A 2d-tree-based blocking method for microaggregating very large data sets[A]. Proc of the First International Conference on Availability, Reliability and Security[C]. Vienna, Austria: IEEE press, 2006. 922-928.

[23] Domingo-Ferrer J. Microaggregation for database and location privacy[A]. Proc of Next Generation Information Technologies and Systems[C]. Kibbutz Shefayim, Israel: Springer-Verlag, 2006. 106-116.

[24] Solanas A, Martinez-Basllestе A, Mateo-Sanz J, et al. Multivariate microaggregation based genetic algorithm[A]. Proc of 3rd IEEE Conference on Intelligent Systems[C]. London, UK: IEEE Press, 2006. 65-70.

[25] Solanas A, Martinez-Basllestе A, Domingo-Ferrer J. V-MDAV: a multivariate microaggregation with variable group size[A]. Proc of Computational Statistics[C]. Rome, Italy: Springer-Verlag, 2006. 917-927.

[26] Domingo-Ferrer J, Seb F, Solanas A. A polynomial-time approximation to optimal multivariate microaggregation[J]. Computer and Mathematics with Applications, 2007, 4: 34-53.

[27] Chang Chin-chen, Li Yu-chiang, Huang Wen-huang. TFRP: an efficient microaggregation algorithm for statistical disclosure control[J]. System Software, 2007, 80(11): 1866-1878.

[28] 彭吉, 唐常杰, 程温泉等. 一种基于层次距离计算的聚类算法[J]. 计算机学报, 2007, 30(5): 786-795.
Peng Ji, Tang Changjie, Cheng Wenquan, et al. A Hierarchy distance computing based clustering algorithm[J]. Chinese Journal of Computers, 2007, 30(5): 786-795. (in Chinese)

[29] 李洁, 高新波, 焦李成. 基于特征加权的模糊聚类新算法[J]. 电子学报, 2006, 34(1): 89-92.
Li Jie, Gao Xinbo, Jiao Licheng. A new feature weighted fuzzy clustering algorithm[J]. Acta Electronica Sinica, 2006, 34(1): 89-92. (in Chinese)

[30] 李洁, 高新波, 焦李成. 一种基于 CSA 的混和属性特征大数据集聚类算法[J]. 电子学报, 2004, 32(3): 357-362.
Li Jie, Gao Xinbo, Jiao Licheng. A CSA-based clustering algorithm for large data sets with mixed numeric and categorical values[J]. Acta Electronica Sinica, 2004, 32(3): 357-362. (in Chinese)

[31] Panaretos J, Nikolaos T, . Aspects of estimation procedures at eurostat with some emphasis on over-space harmonization [A]. Proc of the 5th Hellenic-European Conference on Computer Mathematics and its Applications[C]. Athens, Greece: LEA Press, 2001. 853-857.

[32] Hundepool A, Van de Wetering A, Ramaswamy R, et al. μ-ARGUS version 4.1 software and user's manual. Statistics Netherlands, Voorburg NL[EB/OL]. http://neon.vb.cbs.nl/casc, 2007-2-14.

[33] K Pearson. On lines and planes of closest fit to systems of points in space[J]. Phil Mag, 1901, 2: 559-572.

[34] Domingo-Ferrer J, Mateo-Sanz J M. Efficient multivariate data-oriented microaggregation[J]. The VLDB Journal, 2006, 15: 355-369.

[35] Domingo Ferrer J, Torra V. Towards fuzzy c-means based microaggregation[A]. Third international workshop on Soft Methods in Probability, Statistics and Data Analysis[C]. Warsaw, Poland: Physica-Verlag, 2002. 289-294.

[36] Truta T M, Fotouhi F, Barth Jones A. Assessing global disclosure risk in masked microdata[A]. ACM Workshop on Privacy in the Electronic Society[C]. Washington, DC USA: ACM Press, 2004. 85-92.

[37] Domingo-Ferrer J, Mateo-Sanz J M, Torra V. Comparing SDC methods for microdata on the basis of information loss and disclosure risk[A]. Pre-proc of ETK-NTTS[C]. Luxemburg: Eurostat, 2002. 807-826.

[38] Domingo-Ferrer J, Torra V. Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies[M]. USA: Elsevier, 2001. 93-112.

[39] Yancey W E, Winkler W E, Creecy R H. Disclosure risk assessment in perturbative microdata protection[A]. J Domingo-Ferrer(Ed) Inference Control in Statistical Databases[C]. New York: Springer-Verlag, 2002. 135-152.

[40] Mateo-Sanz J M, Domingo-Ferrer J, Sebe F. Probabilistic information loss measures in confidentiality protection of continuous microdata[J]. Data Mining and Knowledge Discovery, 2005, 11: 181-193.

作者简介:

韩建民　男, 1969 年生于辽宁大连, 华东理工大学计算机科学与工程系博士研究生, 副教授, 研究方向为信息安全, 软件工程.
E-mail: hanjm@zjnu.cn

岑婷婷　女, 1985 年生于浙江温州, 浙江师范大学硕士研究生, 研究方向为数据安全, 数据挖掘.

虞慧群　男, 1967 年生于江苏溧阳, 工学博士, 华东理工大学教授, 博士生导师, IEEE 高级会员, ACM 会员, 中国计算机学会高级会员. 主要研究方向为软件工程, 信息安全, 形式化方法.
E-mail: yhq@ecust.edu.cn