软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(2):341−351 [doi: 10.13328/j.cnki.jos.005015]

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

面向表数据发布隐私保护的贪心聚类匿名方法*

姜火文[1,2,3], 曾国荪[1,3], 马海英[1,3]

[1](同济大学 计算机科学与技术系,上海 200092)

[2](江西科技师范大学 数学与计算机科学学院,江西 南昌 330038)

[3](嵌入式系统与服务计算教育部重点实验室(同济大学),上海 200092)

通讯作者: 曾国荪, E-mail: gszeng@tongji.edu.cn

摘 要: 为了防范隐私泄露,表数据一般需要匿名处理后发布.现有匿名方案较少分类考察准标识属性概化,并缺少同时考虑信息损失量和时间效率的最优化.利用贪心法和聚类划分的思想,提出一种贪心聚类匿名方法:分类概化准标识属性,并分别度量其信息损失,有利于减小并合理评价信息损失.对元组间距离和元组与等价类距离,建立与最小合并概化信息损失值正相关的距离定义,聚类过程始终选取具有最小距离值的元组添加,从而保证信息损失总量趋于最小.按照 k 值控制逐一聚类,实现等价类均衡划分,减少了距离计算总量,节省了运行时间.实验结果表明,该方法在减少信息损失和运行时间方面是有效的.

关键词: 数据发布;隐私保护;聚类匿名;信息损失

中图法分类号: TP309

中文引用格式: 姜火文,曾国荪,马海英.面向表数据发布隐私保护的贪心聚类匿名方法.软件学报,2017,28(2):341–351. http://www.jos.org.cn/1000-9825/5015.htm

英文引用格式: Jiang HW, Zeng GS, Ma HY. Greedy clustering-anonymity method for privacy preservation of table data-publishing. Ruan Jian Xue Bao/Journal of Software, 2017,28(2):341−351 (in Chinese). http://www.jos.org.cn/1000-9825/5015.htm

Greedy Clustering-Anonymity Method for Privacy Preservation of Table Data-Publishing

JIANG Huo-Wen[1,2,3], ZENG Guo-Sun[1,3], MA Hai-Ying[1,3]

[1](Department of Computer Science and Technology, Tongji University, Shanghai 200092, China)

[2](School of Mathematics and Computer Science, Jiangxi Science & Technology Normal University, Nanchang 330038, China)

[3](Key Laboratory of Embedded System and Service Computing, Ministry of Education (Tongji University), Shanghai 200092, China)

Abstract: To prevent privacy disclosure, table data generally needs to be anonymized before being published. Existing anonymity methods seldom distinguish different types of quasi-identifier in generalization, and also lack investigation into optimization of both information loss and time efficiency. In this paper, a greedy clustering-anonymity method is proposed using the ideas of greedy algorithm and clustering algorithm. The method makes distinct generalizations according to the type of quasi-identifier to conduct different calculations on information loss, and this providing reduction and reasonable estimate on information loss. Moreover, with regard to distance between tuples, or distance between a tuple and an equivalence class, two definitions are put forward in order to achieve minimum information loss in merging generalization. When establishing a new cluster, the tuple with the minimum distance in the ongoing cluster is always chosen to add. It ensures that the total information loss is close to minimum. Since the number of tuples in establishing each cluster is subject to k and the size of every cluster is equal to or just greater than k, the amount of calculation on

* 基金项目: 华为创新研究计划(IRP-2013-12-03); 高效能服务器和存储技术国家重点实验室开放基金(2014HSSA10); 江西科技师范大学重点科研项目(2016XJZD002)

Foundation item: Huawei Innovation Research Project (IRP-2013-12-03); Program of State Key Laboratory of High-End Server & Storage Technology (2014HSSA10); Key Research Project of Jiangxi Science and Technology Normal University (2016XJZD002)

收稿时间: 2015-02-11; 修改时间: 2015-11-06; 采用时间: 2015-12-22

distances and therefore the running time are reduced. Experimental results show that the proposed method is effective in reducing both information loss and running time.

Key words: data-publishing; privacy preservation; clustering-anonymity; information loss

当今社会,每天都在产生并流通海量数据,大量数据的消费与共享给人们的工作和生活带来便利,也增加了涉及个体或组织隐私信息泄露的风险.因此,实现具有隐私保护的数据共享,成为人们的期盼."隐私"意指当事人不愿意被外人知道的敏感信息,它与公共利益、群体利益无关,具有隐藏特性[1,2].数据发布中的隐私保护是对公开发布的数据,采取隐私保护举措,防止他人通过知识推断、数据挖掘、链接攻击等手段获取到隐私数据.例如,表 1 为一张待发布的就诊记录表,医院可能出于方便相关机构研究病情分布特点、犯病规律等原因,将这些医疗数据公开发布.显然,这些数据不作防护地直接发布,会暴露个体疾病隐私.如何对数据记录进行处理,以防止患者疾病隐私泄露?最简单的方法就是将其中"Name"属性数据隐去,这样就不会直接暴露"某人患有某病"这种隐私,也不影响数据供相关分析或研究利用.但攻击者在知道患者某些背景知识的情况下,根据(Age,Sex,Zip)属性组合,也可能推测出某个人,导致个人隐私泄露.例如,攻击者如果知道"某社区(对应 zip 属性)42 岁的男同志赵四,曾在那段时间到所在医院看病",就可推断出赵四患有肝炎.另外,美国卡耐基梅隆大学的 Sweeney 研究指出[3],通过与发布的关系表(文献中举例为选民登记表)作连接运算,能够以高概率推断出个体隐私.

对于涉及隐私保护的关系表数据,可将其属性归纳为 4 类:标识(explicit identifier)属性、准标识(quasi-identifier)属性、敏感(sensitive)属性及其他属性.为方便讨论,一般忽略其他属性.标识属性也称为身份属性,或简称标识符,是能够直接区分个体身份的属性,如,表 1 中的"Name"属性.准标识属性也称为非敏感属性,或简称为准标识符,是联合起来可能推断出个体身份的多个属性,如,表 1 中"Age"、"Sex"、"Zip"均为准标识属性.敏感属性,也称为隐私属性,是包含隐私数据的属性,如,表 1 中的"Disease"属性.由上例可见,单纯隐匿标识属性并不能保证隐私安全,因为攻击者可能根据准标识属性,借助背景知识或链接攻击等方式推断出标识属性,从而暴露隐私.为了保护隐私,非常简单的做法是把准标识属性和标识属性一起隐去,这样就没有可能根据准标识属性推断出标识属性与隐私属性的一对一关联,但这样仅剩隐私属性的表数据将变得毫无用处.隐私保护本质上就是要切断敏感属性与标识属性的联系,防止两者间建立一一对应.然而,隐私保护既要保护数据隐私,又要维持数据可用,即保持隐私保护程度与数据可用性的平衡.

Sweeney 提出的 k-匿名(k-anonymity)模型可以防范如上所述的基于背景知识的攻击和链接攻击,是一种经典的隐私保护方法[3].该方法将表数据划分为至少包含 k 个元组的等价类,分别对其匿名,使每个等价类全部元组在准标识属性上取值相同,即对表中任一元组 t_i,至少存在 k–1 个元组与 t_i 在准标识属性上取值相同,以此保证最多只能以 1/k 的概率,通过准标识属性关联出标识属性.概化(generalization)和压缩(suppression)技术被广泛应用于实现 k-匿名[4,5].概化是将具体的属性值用值域更广的概括值来取代,例如,将年龄 45、55 概括为[45,55].表 2 显示了表 1 的 2-匿名概化结果.2006 年,Aggrawal 等人首次提出利用聚类方法实现数据匿名[6],此后陆续可见一些聚类匿名的研究成果.

Tabel 1 Original table

表 1 原始表

Name	Age	Sex	Zip	Disease
张一	30	男	23208	感冒
王二	35	男	23200	哮喘
李三	45	女	23085	胃炎
赵四	42	男	23220	肝炎
陈五	55	女	23050	风湿

Tabel 2 Anonymity table

表 2 匿名表

Age	Sex	Zip	Disease
(30,42)	男	232**	感冒
(30,42)	男	232**	哮喘
(30,42)	男	232**	肝炎
(45,55)	女	230**	胃炎
(45,55)	女	230**	风湿

针对表数据发布中隐私的匿名保护问题,本文提出一种贪心聚类匿名法.该方法区别于多数统一依据概化层次树概化和评价信息损失的方法,区分准标识属性类型,分别概化并度量信息损失.合理定义两个元组间和元组与等价类间的距离,使其正向反映概化信息损失.借用贪心法和聚类思想将 n 个元组按距离最小化进行贪心聚类划分,分别实现匿名,保证了总体信息损失和匿名时间趋于最优.本文第 1 节简介相关研究工作.第 2 节介绍

k-匿名和概化的概念,重点给出概化信息损失的度量方法.第 3 节主要阐述并分析 GAA-CP 匿名算法.第 4 节实验对本文算法与相关匿名算法进行对比分析.第 5 节总结全文,并指出下一步的研究工作.

1 *k*-匿名技术相关研究

作为一种简单而有效的隐私保护手段,*k*-匿名技术自提出以来,得到了广泛研究.Sweeney 最早提出的 MinGen 算法[4]每一步都完全搜索概化空间,以便选取最优概化操作,直到数据满足 *k*-匿名原则.算法简洁、有效,但由于时间复杂度高,实用性受限.文献[4]提出的 Datafly 算法则基于 MinGen 算法,引入压缩与启发式概化思想,提升了算法效率.2005 年,LeFevre 等人提出了 Incognito 匿名算法[7],采用全局重编码技术,自下而上对原始数据进行概化,但 Datafly 和 Incognito 都容易过度概化.2006 年,LeFevre 等人又提出了 Mondrian 匿名算法[8],将原始数据映射到一个多维空间,通过对空间多维数据的优化划分来实现数据 *k*-匿名.2010 年,韩建民等人提出了面向敏感值的个性化(*a*,*k*)-匿名隐私保护模型[9].该方法不利于处理数值属性及多个敏感属性的情况.2012 年,王波等人提出了一种基于逆聚类的个性化隐私匿名方法[10],依据敏感属性值上的差异程度实现聚类分簇.算法要求引入保护属性,数据失真也较大.近 10 余年来,很多种 *k*-匿名方法陆续被提出来.Meyerson 等人的研究则表明,求最优匿名法是个 NP 难问题[11].一般来说,不同匿名方法都有一定的优、劣势和应用特点.

自 Aggrawal 等人在 *k*-匿名中引入聚类方法以来[6],已有不少文献研究了隐私保护数据发布中的聚类匿名算法.Li 等人提出一种 KACA 匿名方案[12],应用了聚类思想,匿名过程为反复合并等价类,即每次随机选取一个个数小于 k 的等价类,依据合并概化信息损失量最小原则,选取某个等价类与其合并概化,直到所有等价类包含 k 个以上元组为止.该方法需对所有准标识属性预定义概化层次结构,这固化了概化模式,容易带来不必要的概化.王智慧等人提出一种 *L*-clustering 方法[13],对准标识属性分类概化,通过考察概化前后属性值的不确定性变化程度,给出了信息损失的度量方法,从而将数据匿名问题转化为带特定约束的聚类问题,但算法优先考虑 *l*-多样性保持,执行中可能无解.郭昆等人针对数据流匿名保护问题,提出一种基于聚类的数据流匿名方法[14],但元组间的距离与元组到簇间的距离定义并不清晰,聚类方法的效果与起始元组 t 的选取密切相关,难以保证整体信息损失最小.Zhang 等人提出一种基于信息熵的 *k*-匿名聚类算法 EBKC[15],首先根据设定的阈值δ,将表数据按准标识属性值上的平均距离最小原则划分为一些大小不一的记录子集,然后分别将其中记录数小于 k 的子集合并,将记录数大于 $2k$ 的子集拆分,使每个子集记录数在 k 与 $2k$ 之间.但是,该算法重复划分与合并操作,因复杂度较高而影响实用性.综上所述,许多匿名方法都不能同时保障信息损失量和时间效率趋于最优.

2 表数据 *k*-匿名与概化信息损失度量

二维表数据是信息发布的主要形式,故本文以关系数据库中的表数据为研究对象,探讨其隐私保护方法.表数据隐私保护的主要方法是数据匿名,而数据概化是实现数据匿名的主要手段.

2.1 表数据*k*-匿名概念

为方便讨论,假定待发布的表数据中,行项包含 n 个元组,通常也称记录;列项由 d 个准标识属性和 1 个敏感属性构成.设 $T=(t_1,t_2,\ldots,t_n)$为待发布的表数据,$t_j(j=1,2,\ldots,n)$为表中第 j 个元组.记 $A^q=(A_1^q,A_2^q,\ldots,A_d^q)$为表中所有准标识属性, $A_i^q(i=1,2,\ldots,d)$为表中第 i 个准标识属性,其属性值是数值型或分类型,例如,表 1 中“Age”和“Zipcode”的属性值为一个数值,即为数值型;“Sex”的属性值为一种类别,即为分类型.记 A^s 为表中敏感属性.记元组 t_j 在属性 A_i^q 上的值为 $t_j[A_i^q]$.

定义 1(等价类). 在表数据中,A^q 取值相近的若干元组作为一个概化单位,概化后具有相同的 A^q 值,称这些元组为一个等价类,记为 C_i.

k-匿名过程需要将表数据分成多个等价类,使任意两个等价类相互没有共同元组,所有等价类的集合构成表数据.即 $\bigcup_{i=1}^{m}C_i=T,C_i\bigcap C_j=\varnothing,i\neq j$.

定义 2(等价元组). 经过概化后的某个等价类 C_i,类中所有元组具有相同的准标识属性值,统称为等价元组,

记为 t_{c_i}.

定义 3(*k*-匿名). 将表数据 T 分成 m 个等价类,每个等价类至少包含 k 个元组,表 T 满足 k-匿名,当且仅当在每个等价类中,各元组在准标识属性上取值相同,不能区分.即 $\forall t_i,t_j \in C_w$,有 $t_i[A^q]=t_j[A^q]$,这里,$i \neq j, w \in [1,m]$.

可见,k-匿名要求对表数据中任一元组,至少存在 $k-1$ 个其他元组,它们具有相同的准标识属性值.因此,即便侵犯者具备背景知识,既知道被侵者的准标识属性值,又知道表中肯定包含入侵者的数据记录,能够顺利获知被侵者数据在某一等价类中,但并不能确定到底是该组中哪一条元组,理论上侵犯者只有 $1/k$ 的概率能够定位出元组序号.因此,k-匿名能够防范基于背景知识的隐私攻击.

2.2 数据概化及其信息损失

在匿名处理前的原始表数据中,一个元组和另一元组的准标识属性值一般不会完全相同,不妨假设不存在准标识属性值相同的元组.为了达到 k-匿名,通过值域概化的方法,使同一等价类中各元组具有相同的准标识属性值.

定义 4(值域概化). 对原始表数据中的某个准标识属性 A_i^q (i=1,2,…,d),对各个元组的 A_i^q 值,分别选取某一范围更广的域值取代其原属性值,这个过程被称为值域概化(domain generalization)或数据概化,简称概化.

概化的本质是通过将属性值进行放大,使多个不同的属性值在某个被放大的范围内取值相同.概化后的属性值不再是具体的某一属性值,而是体现为一个包含各个概化前原值的集合概念,故表数据概化能够阻止根据具体准标识属性值进行的表连接.由于链接攻击[3]需要在不同的表数据之间,根据共同的属性值进行表连接,因此 k-匿名能够有效地防范表链接攻击.

概化方法一般是对需要概化的准标识属性 A_i^q (i=1,2,…,d)建立域概化层次树(domain generalization hierarchy tree),根据概化树对各属性值实施值域概化,例如文献[4,7,9,10,12]中的概化方法.本文从保证最小化数据信息损失角度出发,根据准标识属性的不同类型,对表数据采用不同方式概化.即,对数值型属性,用包含同类中各个属性值的最小值域取代原来各个具体属性值,如,表 1 中不同的“Age”值 45、55,用[45,55]取代.此类概化过程不需要依据预定义的概化树进行,无需预定义概化树.对分类型属性,各个属性值被概化为可以概括各原有属性值的范围更广的最小类型值,如,对表 1 中的不同“Sex”值男、女,可以用“个人”取代.此类概化需要根据预定义的概化层次树进行.

毫无疑问,概化用某个抽象属性值取代原来多个不同的具体属性值,将导致不同程度的数据失真.习惯上用信息损失量来衡量数据失真程度.目前,多数衡量信息损失的办法采用统一的计算公式,如文献[4,9,10,12]等用到的度量方法.为了更为合理地评价数据失真度,本文特别区分属性类型,对数值型和分类型属性分别采用不同方法度量其概化信息损失.

2.2.1 数值型属性概化信息损失度量

假设 d 个准标识属性中,数值型属性个数为 d_1,分类型属性个数为 d_2,$d=d_1+d_2$.记 A_i^{nq} 和 A_i^{cq} 分别表示第 i 个数值型属性和第 i 个分类型属性,记所有数值型和分类型准标识属性分别为 A^{nq} 和 A^{cq} .对某个数值型准标识属性 A_i^{nq}, 记 $\max T(A_i^{nq})$和$\min T(A_i^{nq})$ 分别为该属性在表数据中的最大值和最小值,记 $\max C_r(A_i^{nq})$ 和 $\min C_r(A_i^{nq})$ 分别为该属性在等价类 C_r 中的最大值和最小值.我们使用域值 $[\min C_r(A_i^{nq}),\max C_r(A_i^{nq})]$ 对等价类 C_r 中各个元组的 A_i^{nq} 值进行概化.在等价类 C_r 上,属性 A_i^{nq} 概化后的信息损失可以通过以下公式计算.

$$iloss(C_r,A_i^{nq})=|C_r|\times\frac{\max C_r(A_i^{nq})-\min C_r(A_i^{nq})}{\max T(A_i^{nq})-\min T(A_i^{nq})},\ |C_r| \text{ 为等价类 } C_r \text{ 中元组个数.}$$

在等价类 C_r 上,所有数值型非标识属性概化后的信息损失为

$$iloss(C_r,A^{nq})=\sum_{i=1}^{d_1} iloss(C_r,A_i^{nq}).$$

容易计算出,在表数据 T 上,所有数值型非标识属性概化后的信息损失为

$$iloss(T,A^{nq})=\sum_{r=1}^{m} iloss(C_r,A^{nq}).$$

2.2.2 分类型属性概化信息损失度量

分类型属性需要根据预定义的概化层次树进行概化,每个分类型属性对应建立一棵概化层次树,图 1 即为一棵概化层次树.如图 1 所示,树中叶子节点为表数据 T 中该属性上的各个取值,中间节点为各个层次的概化值,根节点为最终概化值.两个属性值的最小概化值即为对应两个叶子节点的最小上界节点,如,“Influenza”和“Asthma”的最小概化值为“Respiratory diseases”.具有垂直关系的两个属性值的路径长度为它们之间的边数,如,“Asthma”到“Diseases”的路径长度为 2.

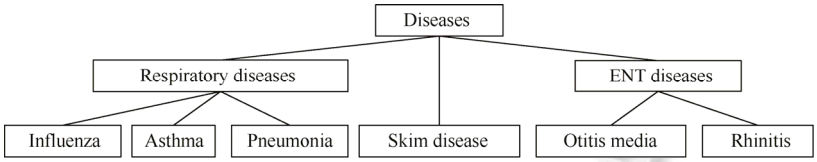


Fig.1 Generalization hierarchy tree for “disease”

图 1 “disease”属性概化层次树

分别记 $root_i$ 为属性 A_i^{cq} 的概化层次树的根节点值,$t_j[A_i^{cq}]$ 为元组 j 的第 i 个分类型准标识属性值,$t_j^*[A_i^{cq}]$ 为 $t_j[A_i^{cq}]$ 的概化值,$h(t_j[A_i^{cq}],t_j^*[A_i^{cq}])$ 为概化层次树上节点 $t_j[A_i^{cq}]$ 和 $t_j^*[A_i^{cq}]$ 间的路径长度,$h(t_j[A_i^{cq}],root_i)$ 为 $t_j[A_i^{cq}]$ 到根节点的路径长度.在等价类 C_r 上,属性 A_i^{cq} 概化后的信息损失定义为

$$iloss(C_r,A_i^{cq})=\sum_{j=1}^{|C_r|}\frac{h(t_j[A_i^{cq}],t_j^*[A_i^{cq}])}{h(t_j[A_i^{cq}],root_i)},\ |C_r|\text{为等价类}C_r\text{中元组个数.}$$

在等价类 C_r 上,所有分类型非标识属性概化后的信息损失为

$$iloss(C_r,A^{cq})=\sum_{i=1}^{d_2}iloss(C_r,A_i^{cq}).$$

在表数据 T 上,所有分类型非标识属性概化后的信息损失为

$$iloss(T,A^{cq})=\sum_{r=1}^{m}iloss(C_r,A^{cq}).$$

综合第 2.2.1 节和第 2.2.2 节的计算,我们得到原始表数据 T 在 k-匿名处理后的总数据信息损失为

$$iloss(T,A^q)=iloss(T,A^{nq})+iloss(T,A^{cq}).$$

3 聚类匿名方法 GAA-CP(greedy k-anonymity algorithm based on clustering partition)

如第 1 节所述,针对数据发布的 k-匿名隐私保护,近年已有多种聚类方法被提出来.聚类(clustering)的基本思想是将一个数据集按相似性程度划分为若干类(cluster,或称为簇),使同一簇中的数据比簇间数据具有更强的相似性.k-匿名一般是将数据表按元组的准标识属性值相近程度分成若干等价类,分别概化,以使相同数目的类中元组比类间元组具有更小的概化信息损失量.可见,聚类划分与 k-匿名中等价类划分,两者在数据划分上有相近的思想.本文在 k-匿名中引入聚类方法,其中关键之一是如何度量准标识属性值的相近性.为此,考虑到最大程度地减少数据信息损失,我们将准标识属性值的相近性与概化信息损失量关联起来,通过定义元组间距离,来反映元组间准标识属性的相近程度和概化信息损失量大小.元组间距离越小,它们之间的准标识属性值就越相近,合并概化到同一等价类所造成的信息损失也就越小.

3.1 元组与等价类距离

定义 5(元组间距离). 表数据 T 中任意两个元组 t_p 和 t_q 在对应各个准标识属性上取值差异程度的平均值称为元组间的距离,记为 $dist(t_p,t_q)$.

由定义 5 可知,元组间距离值反映两者准标识属性值间的相近程度,由彼此在各个准标识属性上的取值差

异度共同决定.为此,可以先按类型不同,对数值型和分类型准标识属性分别引入属性差异度计算方法.

两个数值型属性值 $t_p(A_i^{nq})$ 和 $t_q(A_i^{nq})$ 间的差异度计算公式如下:

$$div(t_p(A_i^{nq}),t_q(A_i^{nq}))=\frac{\left|t_p(A_i^{nq})-t_q(A_i^{nq})\right|}{\max T(A_i^{nq})-\min T(A_i^{nq})},\ \left|t_p(A_i^{nq})-t_q(A_i^{nq})\right|$$ 表示两者间的绝对值.

两个分类型属性值 $t_p(A_i^{cq})$ 和 $t_q(A_i^{cq})$ 间的差异度计算公式如下:

$$div(t_p(A_i^{cq}),t_q(A_i^{cq}))=\frac{1}{2}\times\left(\frac{h(t_p[A_i^{cq}],t_{p\wedge q}^{\ *}[A_i^{cq}])}{h(t_p[A_i^{cq}],root_i)}+\frac{h(t_q[A_i^{cq}],t_{p\wedge q}^{\ *}[A_i^{cq}])}{h(t_q[A_i^{cq}],root_i)}\right).$$

上式中,$t_{p\wedge q}^{\ *}[A_i^{cq}]$ 为概化树上节点 $t_p[A_i^{cq}]$ 和 $t_q[A_i^{cq}]$ 的最小上界节点,即两者的最小概化值.由此,可以定义两个元组间的距离,计算公式如下:

$$dist(t_p,t_q)=\frac{1}{d}\times\left(\sum_{i=1}^{d_1}div(t_p(A_i^{nq}),t_q(A_i^{nq}))+\sum_{i=1}^{d_2}div(t_p(A_i^{cq}),t_q(A_i^{cq}))\right).$$

命题 1. 两个元组间的距离与这两个元组最小合并概化的信息损失量成正比.

证明:首先因为两元组合并概化后,分类型属性值是根据相应的属性概化树概化到两者原值的最小上界,所以称为最小合并概化.以元组 t_p 和 t_q 为例,根据第 2.2 节的内容,经过最小合并概化后,两者数值型属性值将概化为 $[\min(t_p(A_i^{nq}),t_q(A_i^{nq})),\max(t_p(A_i^{nq}),t_q(A_i^{nq}))]$, 信息损失为

$$iloss(t_p\vee t_q,A^{nq})=2\times\sum_{i=1}^{d_1}\frac{\max(t_p(A_i^{nq}),t_q(A_i^{nq}))-\min(t_p(A_i^{nq}),t_q(A_i^{nq}))}{\max T(A_i^{nq})-\min T(A_i^{nq})},$$

分类型属性值将概化为 $t_{p\wedge q}^{\ *}[A_i^{cq}]$,信息损失为

$$iloss(t_p\vee t_q,A^{cq})=\sum_{i=1}^{d_2}\left(\frac{h(t_p[A_i^{cq}],t_{p\wedge q}^{\ *}[A_i^{cq}])}{h(t_p[A_i^{cq}],root_i)}+\frac{h(t_q[A_i^{cq}],t_{p\wedge q}^{\ *}[A_i^{cq}])}{h(t_q[A_i^{cq}],root_i)}\right).$$

故两个元组最小合并概化后的信息损失量计算为

$$iloss(t_p\vee t_q,A^q)=iloss(t_p\vee t_q,A^{nq})+iloss(t_p\vee t_q,A^{cq}),$$

与上面两元组间距离计算公式比较,容易得出 $iloss(t_p\vee t_q,A^q)=2\times d\times dist(t_p,t_q)$,可见命题成立. □

定义 6(元组到等价类的距离). 对于等价类 C_q 和任一元组 t_p $(t_p\notin C_q)$,C_q 对应的等价元组为 t_{c_q}, 定义 t_p 到 C_q 的距离为 $dist(t_p,t_{c_q})$,记为 $dist(t_p,C_q)$.

显然,计算元组到等价类的距离即为求元组间的距离.因此,我们有如下结论.

命题 2. 元组到等价类的距离与该元组和等价类最小合并概化的信息损失量成正比.

证明与命题 1 类似,这里省略.

根据命题 1、命题 2 的结论,由于元组间距离和元组到等价类的距离都直接反映两者间最小合并概化的信息损失量,故我们可以得出下述推论.

推论 1. 根据元组间或元组与等价类间距离最小原则,选择元组构造等价类,能够保证生成的等价类具有最小概化信息损失值.

3.2 GAA-CP匿名算法

本节提出一种基于贪心聚类划分的匿名算法 GAA-CP,其基本思想为:将待匿名发布的表数据 T,依据推论 1 的结论,利用贪心法和聚类思想,划分成一些个数不小于 k 的聚类(即等价类),最后对每个等价类按预定规则进行概化匿名,使匿名后总信息损失量最小.下面我们通过伪码形式给出算法实现步骤.

算法 1. 基于贪心聚类划分的匿名算法 GAA-CP.

输入:待发布的表数据 T,匿名参数 k.

输出:满足约束条件的匿名表数据 T^*.

步骤:

1. $T^*=\varnothing$; $m=0$; /* 初始化匿名表 T^*和变量 m,m 记录等价类个数 */
2. $\forall t_r\in T$, $C=\{t_r\}$; $T=T-\{t_r\}$; /* 从 T 中任取一个元组作为等价类 C 的初始元组 */
3. while $|T|\geq k$ do /* 只要 T 中元组个数不小于 k,即循环划分等价类 */
{3.1. while $|C|<k$ do {从 T 中找出与 C 距离最小的元组(集)加入};
3.2. $T^*=T^*\cup\{C\}$; /* 将等价类 C 添加到 T^*中 */
3.3. 在 T 中找与 C 距离最远的某元组 $t_{\max i}$;
3.4. $C=\{t_{\max i}\}$; $T=T-\{t_{\max i}\}$;}; /* 将 $t_{\max i}$从 T 移出到 C,构造一个新等价类 C */
4. 对 T 中剩余元组,逐个找到距离最近的等价类加入;
5. 依次取 $C_i\in T^*(i\in[1,m])$,对其所有准标识属性 A^{qi} 进行最小概化;

3.3 GAA-CP算法分析

3.3.1 复杂性分析

根据上述伪代码描述的算法具体步骤,可以分析其计算复杂性.其中,主要的步骤 3 是一个双层循环,其外循环次数为划分成的等价类个数 m,$m\leqslant[n/k]$($[n/k]$表示不大于 n/k 的整数).内循环操作主要是实现等价类构造,其循环次数等于元组添加次数(每次添加与 C 距离最小的元组(集),可能不止一个元组),故内循环次数上限为 k,可以认为约等于 k.内循环操作中,每次选取与 C 距离最小的元组(集)(步骤 3.1)和每次选取等价类的初始元组(步骤 3.3),都需要做$|T|$次($|T|$表示 T 中剩余元组个数)距离求值与比较,此操作可以在 $O(n)$时间内完成.以上累计计算,步骤 3 可以在 $m\times k\times O(n)$时间内完成,其时间复杂度为 $O(n^2)$.最后步骤 4 和步骤 5 都可以在 $O(n)$时间内完成,其时间复杂度为 $O(n)$.因此,算法总的时间复杂度为 $O(n^2)$.

在算法运行过程中,内存空间消耗主要是存储两个表数据 T 和 T^*,其大小与原始表数据 T 的元组个数和属性值个数相关,一般不需要作特别考虑.

3.3.2 有效性分析

本算法主要实现将包含 n 个元组的表 T 划分为 m 个等价类,使得:$\forall i\in[1,m]$,$|C_i|\in[k,2k)$,且 $iloss(T,A^q)$有最佳的极小值,其实质是解决一个带约束条件的聚类划分问题.由第 3.2 节所给出的算法可见,类划分过程的每步都是按距离最小原则选取元组,依据第 3.1 节的推论 1,算法能够保障划分后的各等价类在概化匿名时的数据信息损失总和具有最优的极小值.根据算法描述细节中的步骤 3.1 可知,所划分等价类的元组个数均大于等于 k,结合步骤 4 考虑可知,最终各等价类元组个数不会达到 $2k$,所以,对$\forall i\in[1,m]$,$|C_i|\in[k,2k)$.

由于划分后各等价类元组个数不小于 k,经过概化匿名后,同一等价类中各元组在准标识属性上取值相同.因此,在算法输出的匿名表 T^*中任取一个元组 t,至少存在 k–1 个其他元组,它们具有相同的准标识属性值,显然,表 T^*符合 k-匿名模型.根据第 2.1 节和第 2.2 节相关部分的论述,满足 k-匿名要求的输出表 T^*能够有效地抵抗基于背景知识的攻击和链接攻击.

4 实验与结果分析

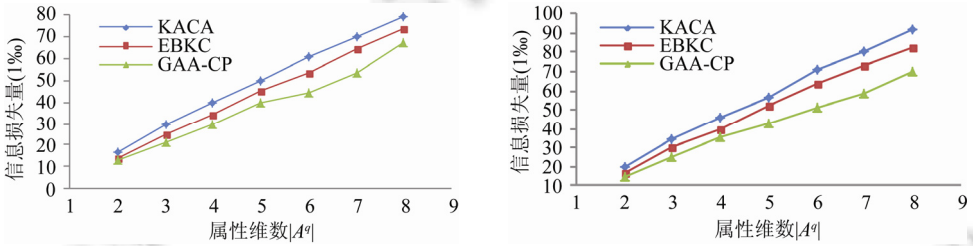
本节通过实验分析验证 GAA-CP 算法的性能,并将其与文献[12]提出的 KACA 算法和文献[15]提出的 EBKC 算法进行比较.在第 1 节,我们按年代顺序例举了近 10 年来有一定典型意义的 4 种聚类匿名方法,它们分别出自文献[12−15].其中,文献[12]中提出的 KACA 算法和文献[15]中提出的 EBKC 算法在距离定义、信息损失度量、隐私保护的对象和目标等方面与本文的 GAA-CP 算法有更近的相似性和可比性,故我们选择这三者进行实验对比和分析.

实验数据集来源于隐私保护研究领域被广泛使用的 UCI 机器学习数据库中的 Adult 数据集(http://archive.ics.uci.edu/ml/datasets/Adult).和文献[12]一样,在删除了那些具有未知属性值的元组以后,得到一个共包含 45 222 个元组的表数据,表中每个元组都保留了 9 个属性,分别为 Age,Gender,Education,Marital Status,Race,Work Class,Native Country,Salary Class 和 Occupation,其中,Occupation 为敏感属性,前面 8 个属性均为准标识属性,在这 8 个准标识属

性中,Age 为数值型,其余 7 个均为分类型.实验环境为:英特尔 Pentium 双核 E2140 @1.60GHz CPU;1GB(DDR)内存;希捷 ST3160815AS(160GB/7200 转/分)主硬盘;Windows XP 专业版 32 位 SP3 操作系统;Microsoft SQL Server2000 数据库系统,算法均采用 Microsoft Visual C++ 7.0 实现.考虑到各算法在构造新等价类时,随机选取起始元组会导致结果有略微差别,每组实验重复进行 10 次,结果取其平均值.

4.1 信息损失量分析

为了分析数据信息损失量随准标识属性维数$|A^q|$以及k-匿名中对应k值改变而变化的规律,我们各进行了一组实验,数据结果按千分之一比例取值.图 2(a)和图 2(b)分别给出了当 k=5 和 k=10 时,KACA,EBKC 和 GAA-CP 中$|A^q|$的变化对信息损失大小的影响.可以发现,当 k 值相同时,随着$|A^q|$的增大,数据信息损失量不断增大,且增大比例不完全成正比.这是因为,$|A^q|$增大,需要概化的数据量也显著增大,导致信息损失量对应显著增大.但是,在 8 个准标识属性中,由于只有第 1 个“Age”属性是数值型属性,显然准标识属性的维数从 2 增加到 8,每次增加的都是一个分类型属性,而一般来说,分类型属性概化导致的信息损失会略大于数值型属性,故随着$|A^q|$的增加,信息损失量不是完全成比例地增大,且略微体现先快后慢的趋势.比较 k 值和$|A^q|$值均相等时的信息损失量,始终有 KACA>EBKC>GAA-CP,这是因为 KACA 对全部准标识属性概化都需要依据预定义的概化树,这容易导致一些属性值过度概化.EBKC 在聚类初始划分和聚类合并时是按信息最小原则进行,而聚类拆分过程却是在最大熵值属性上按熵值最大化原则进行,而 GAA-CP 在划分全部等价类过程中,每一次元组添加都是按照信息损失最小原则来选取.

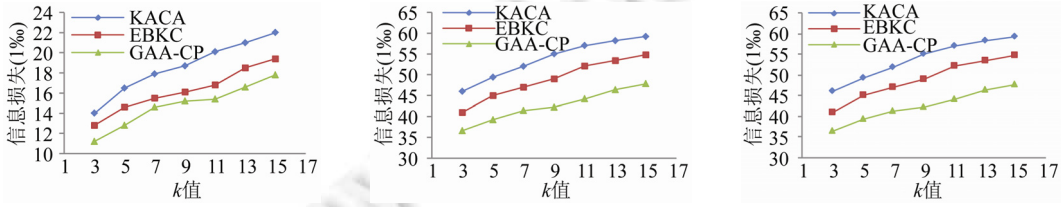


(a) k=5 (b) k=10

Fig.2 How the quantity of information loss changes with $|A^q|$ when k is constant

图 2 当 k 值确定、$|A^q|$值改变时,信息损失量变化情况

图 3(a)~3(c)分别给出了当$|A^q|$=2,5,8 时,数据信息损失随着反映匿名程度的 k 值改变而变化的情况.可以看出,当$|A^q|$值确定时,随着 k 值增大,信息损失量有所增大.这是因为 k 值变大,等价类中元组个数将增多,将这些更多的元组概化到同一属性值时,一般需要增大概化程度,以至于各元组数据信息损失量将变大,相应的整体信息损失量增大.然而,k 值变大,意味着匿名强度高,匿名后将有更多元组具有相同准标识属性值.这也验证了信息损失量与匿名强度是两者不可兼得的矛盾关系.前者反映匿名代价,后者反映隐私保护强度,需要我们根据实际平衡取舍.



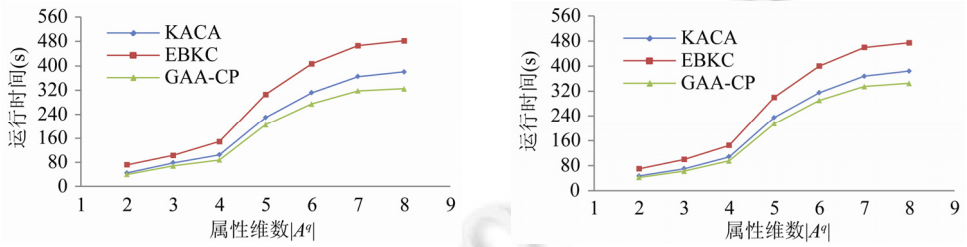
(a) $|A^q|$=2 (b) $|A^q|$=5 (c) $|A^q|$=8

Fig.3 How the quantity of information loss changes with k when $|A^q|$ is constant

图 3 当$|A^q|$值确定、k 值改变时,信息损失量变化情况

4.2 运行时间分析

为了进一步比较分析GAA-CP算法的运行时间特点,我们分别进行了3组实验.第1组实验为保持元组个数不变,分别固定 k=5 和 k=10,考察$|A^q|$值改变时的运行时间变化,图 4(a)和 4(b)分别给出了其实验结果.可以看出,当 k 值一定时,随着$|A^q|$值的增大,3 种算法的运行时间都明显增大,且 EKBC 算法运行时间最大,KACA 其次,GAA-CP 最小.这是因为随着$|A^q|$值的变大,各种算法均需要对更多的准标识属性进行概化,并根据其信息损失来求元组间距离,从而运算量明显增加,故运行时间增大.EBKC 由于要重复等价类合并和划分过程,导致运行时间较长,GAA-CP 每步根据所包含元组个数不超过 k 值逐个构造等价类,均衡了各等价类大小,使等价类个数和等价类之间距离的计算量都趋于最优,因而运行时间较短.

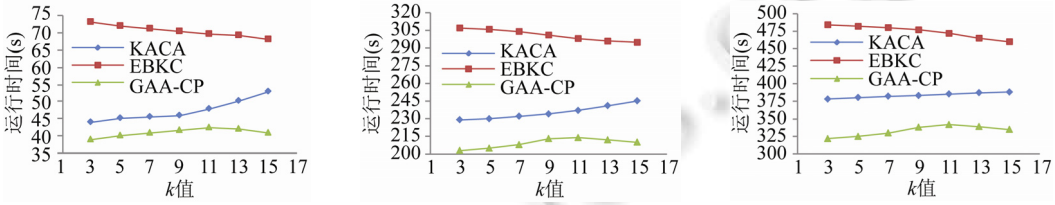


(a) k=5 (b) k=10

Fig.4 How the running time of each algorithm changes with $|A^q|$ when k is constant

图 4 当 k 值确定、$|A^q|$值改变时,算法运行时间变化情况

分析算法运行时间的第2组实验是,分别固定值$|A^q|$=2,5,8,考察 k 值改变时的运行时间变化,图 5(a)~5(c)分别给出其实验结果.可以发现,在运行时间上,随着 k 值变大,3 种算法均只有微小变化,且一直维持 EBKC>KACA>GAA-CP,其中,EBKC有微量缩小,KACA有微量增大,GAA-CP大体呈微小幅度先增后减.因为一般情况是,k值增大,等价类将变大,故构造单个等价类时间将变长,但同时由于总元组个数固定,划分出的等价类个数将减少,因此算法总的运行时间变化很小.然而,综合考察等价类的规模和数量变化可能给运行时间带来的此增彼减,各算法总体运行时间大小应该与其最佳性能发挥大小趋向一致.故运行时间变化之所以是 EBKC 微幅减小,是因为算法中选取的阈值(δ=0.5)比较有利于较小等价类的形成;KACA 微幅增大,是因为算法性能最佳化与等价类规模适量变大趋向基本吻合;GAA-CP 微幅先增后减,是因为算法在划分等价类的规模适中(k=9~11)时,性能趋向最佳.



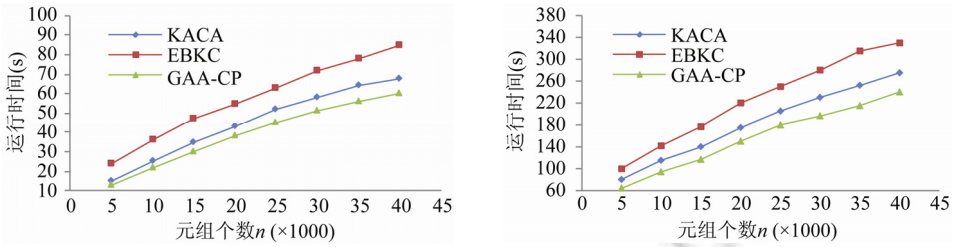
(a) $|A^q|$=2 (b) $|A^q|$=5 (c) $|A^q|$=8

Fig.5 How the running time of each algorithm changes with k when $|A^q|$ is constant

图 5 当$|A^q|$值确定、k 值改变时,运行时间变化情况

分析算法运行时间的第 3 组实验是,固定$|A^q|$值和 k 值,考察匿名表规模,即总体元组个数成倍增减时的运行时间变化.图 6(a)和 6(b)分别给出了当 k=8,$|A^q|$=3 和 k=8,$|A^q|$=6 时的实验结果.可以看出,保持$|A^q|$值和 k 值不变,随着匿名表规模的成倍增加,算法运行时间也在近似成倍地增大.这是因为匿名表规模成倍增加,意味着有成倍的元组数量需要概化,算法总体运算量也对应成倍的增加,但由于各算法如重复执行,在构造起始等价类时都有

一定的随机性,导致运算量可能会有微小差别,所以总运算量不会是严格意义地成倍增加,故运行时间是近似成倍地增大.从图中还可以看出,k 值和$|A^q|$值相同时,如果总元组个数相同,运行时间都有 EBKC>KACA>GAA-CP;但当 k 值和总元组个数相同,而$|A^q|$值取为 3 和 6 时,运行时间相差悬殊,这两种情况与本节上述两个实验结果吻合,原因同以上分析.



(a) k=8, $|A^q|$=3 (b) k=8, $|A^q|$=6

Fig.6 How the running time changes with the size of tables holding $|A^q|$ and k uniform

图 6 当固定$|A^q|$和 k 值、匿名表规模改变时,运行时间变化情况

5 结束语

k-匿名是一种主流的隐私保护模型,聚类是实现 k-匿名的一种重要手段.面向表数据发布隐私保护,本文提出了一种基于贪心聚类思想的匿名方法.该方法在完成全部等价类划分时,每步选取元组都依据距离最小原则,符合信息损失最小要求,从而保证了信息损失总量极小.同时,由于采用贪心聚类法划分等价类,均衡了各等价类大小,使距离计算总量趋于最小,因而节省了运行时间.围绕信息损失小、运行时间短两个目标,通过多组实验结果的比较与分析,验证了算法的有效性.针对隐私属性分布特点对隐私保护的影响,以及表中有多个隐私属性的数据发布隐私保护,我们将在此后的研究中进一步深入考察.

References:

[1] Big data branch of CCF. The development overview of technology and industry for big data in China (2013). 2013 (in Chinese). http://www.ccf.org.cn/sites/ccf/ccfziliao.jsp?contentId=2774793649105

[2] Blum A, Katrina L, Aaron R. A learning theory approach to noninteractive database privacy. Journal of the ACM, 2013,60(2):1–25. [doi: 10.1145/2450142.2450148]

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

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

[5] Wang PS, Wang JD. Survery of research on anonymization privacy-preserving techniques. Journal of Chinese Computer Systems, 2011,32(2):248–252 (in Chinese with English abstract).

[6] Aggarwal G, Feder T, Kenthapadi K, Khuller S, Panigrahy R, Thomas D, Zhu A. Achieving anonymity via clustering. In: Proc. of the 25th ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems. New York: ACM, 2006. 153–162. [doi: 10.1145/1142351.1142374]

[7] LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: Efficient full domain *k*-anonymity. In: Proc. of the ACM SIGMOD Conf. on Management of Data. Baltimore, 2005. 49–60. [doi: 10.1145/1066157.1066164]

[8] LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian multi-dimensional *k*-anonymity. In: Proc. of the 22nd Int'l Conf. on Data Engineering. Atlanta, 2006. 25–35. [doi: 10.1109/ICDE.2006.101]

[9] Han JM, Yu J, Yu HQ, Jia J. Individuation privacy preservation oriented to sensitive values. Acta Electronica Sinica, 2010,38(7):1723–1728 (in Chinese with English abstract).

[10] Wang B, Yang J. A personalized privacy anonymous method based on inverse clustering. Acta Electronica Sinica, 2012,40(5):883−890 (in Chinese with English abstract). [doi: 10.3969/j.issn.0372-2112.2012.05.004]

[11] Meyerson A, Williams R. On the complexity of optimal *k*-anonymity. In: Proc. of the 23rd ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems. New York: ACM Press, 2004. 223−228. [doi: 10.1145/1055558.1055591]

[12] Li JY, Wong RCW, Fu ACW, Pei J. Achieving *k*-Anonymity by clustering in attribute hierarchical structures. In: Proc. of the Data Warehousing and Knowledge Discovery 2006. LNCS 4081, 2006. 405−416. [doi: 10.1007/11823728_39]

[13] Wang ZH, Xu J, Wang W, Shi BL. Clustering-Based approach for data anonymization. Ruan Jian Xue Bao/Journal of Software, 2010,21(4):680−693 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3508.htm [doi: 10.3724/SP.J.1001.2010.03508]

[14] Guo K, Zhang QS. Fast clustering-based anonymization algorithm for data streams. Ruan Jian Xue Bao/Journal of Software, 2013, 24(8):1852−1867 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4330.htm [doi: 10.3724/SP.J.1001.2013.04330]

[15] Zhang JP, Zhao Y, Yang Y, Yang J. A *k*-anonymity clustering algorithm based on the information entropy. In: Proc. of 2014 IEEE the 18th Int'l Conf. on Computer Supported Cooperative Work in Design. 2014. 319−324. [doi: 10.1109/CSCWD.2014.6846862]

附中文参考文献:

[1] 中国计算机学会大数据专委会.中国大数据技术与产业发展白皮书(2013).2013. http://www.ccf.org.cn/sites/ccf/ccfziliao.jsp?contentId=2774793649105

[5] 王平水,王建东.匿名化隐私保护技术研究综述.小型微型计算机系统,2011,32(2):248−252.

[9] 韩建民,于娟,虞慧群,贾洞.面向敏感值的个性化隐私保护.电子学报,2010,38(7):1723−1728.

[10] 王波,杨静.一种基于逆聚类的个性化隐私匿名方法.电子学报,2012,40(5):883−890. [doi: 10.3969/j.issn.0372-2112.2012.05.004]

[13] 王智慧,许俭,汪卫,施伯乐.一种基于聚类的数据匿名方法.软件学报,2010,21(4):680−693. http://www.jos.org.cn/1000-9825/3508.htm [doi: 10.3724/SP.J.1001.2010.03508]

[14] 郭昆,张岐山.基于聚类的快速数据流匿名方法.软件学报,2013,24(8):1852−1867. http://www.jos.org.cn/1000-9825/4330.htm [doi: 10.3724/SP.J.1001.2013.04330]

姜火文(1974－),男,江西南昌人,博士生,副教授,CCF 学生会员,主要研究领域为隐私安全,软件演化,智能计算.

曾国荪(1964－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为并行计算,可信软件,信息安全.

马海英(1977－),女,博士,副教授,主要研究领域为隐私保护,公钥密码学,网络安全.