基于敏感属性熵的微聚集算法

杨 静,王 超,张健沛
(哈尔滨工程大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 在聚类过程中,不合适的距离度量会导致匿名过程中不必要的信息损失,因此对于不同类型的属性定义一个适当的距离度量一直是个难以解决的问题.本文提出语义属性的概念,并提出编码层次树来表示语义属性,有效地降低了匿名过程中的信息损失.在 p-敏感 k-匿名模型中,敏感属性值在聚类结果中分布不均匀会导致敏感信息泄露,因此本文提出一种基于敏感属性熵的微聚集算法,并提出匿名保护指数来描述隐私保护程度,在聚类过程中通过保证匿名保护指数最大,来提高敏感属性在聚类结果中分布的均匀程度,以应对背景知识攻击,降低隐私泄漏的风险.最后,通过实验验证了算法的合理性和有效性.

关键词: 隐私保护; 编码层次树; 微聚集; p-敏感 k-匿名; 敏感属性熵
中图分类号: TP309.2 **文献标识码:** A **文章编号:** 0372-2112 (2014)07-1327-11
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.07.013

Micro-Aggregation Algorithm Based on Sensitive Attribute Entropy

YANG Jing, WANG Chao, ZHANG Jian-pei
(*College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*)

Abstract: In the process of clustering, inappropriate distance measure leads to unnecessary loss of information during the anonymous process, so it is a difficult problem to define a proper distance measurement for different types of variables. We put forward the concept of semantic attribute, and propose a coding hierarchy tree to represent semantic attribute and to reduce the information loss in the anonymous process. In the p-sensitive k-anonymity model, the uneven distribution of the sensitive attribute values in the clustering results may cause sensitive information disclosure, so we propose a micro-aggregation algorithm based on sensitive attribute entropy. Moreover, we propose the concept of anonymous protection factor to describe the degree of privacy protection. During the process of clustering, in order to improve the uniformity of the distribution of sensitive attribute values in the clustering results, the algorithm ensures the maximum of anonymous protection factor, so it can deal with the background knowledge attack and reduce the risk of privacy leaking. Finally, the rationality and validity of the algorithm is verified by experiment.

Key words: privacy preserving; code hierarchy tree; micro-aggregation; p-sensitive k-anonymity; sensitive attribute entropy

1 引言

随着信息技术的快速发展,大量的个人数据能够使用数据挖掘的方法进行分析,虽然知识发现和数据挖掘等数据分析技术充分地挖掘了信息资源,但是也极有可能造成个人的隐私信息泄露.因此,为了保护个人的隐私信息,在数据共享和发布之前,必须对数据进行处理,对其中的敏感属性进行保护.

目前,学术界对隐私保护的技术做了比较深入地研究[1~3],大致可以分为 3 类:基于数据失真的技术[4,5]、基于数据加密的技术[6,7]和基于限制发布的技术[8~11].基于限制发布的技术不仅能保证较高的隐私保护度,还能保证较低的数据依赖性,数据损失以及计算开销,因此得到了广泛的应用.但是它在实现过程中,大多采用泛化、隐匿的技术实现,而属性值泛化域的确定一直是一个难以解决的问题,而且对连续型数据的泛化会丢失较多的数值语义.针对泛化、隐匿技术的不足,Laszlo[12],Solanas[13],Domingo-Ferrer[14]等采用了微聚集的算法,很好地解决了数值语义丢失较多的问题.尽管如此,攻击者仍然能够在得到的数据基础上,结合其他背景知识来

收稿日期:2013-06-25;修回日期:2013-10-17;责任编辑:覃怀银
基金项目:国家自然科学基金(No.61370083,No.61073043,No.61073041);高等学校博士学科点专项科研基金(No.20112304110011,No.20122304110012);哈尔滨市科技创新人才研究专项资金(优秀学科带头人)(No.2011RFXXG015)

推断实体的敏感属性值.

针对 p-敏感 k-匿名模型[15]中,可能存在的敏感属性值在聚类结果中分布不均匀,而导致的敏感信息泄露问题,本文提出了一种基于敏感属性熵的微聚集算法(Micro Aggregation Algorithm based on Sensitive Attribute Entropy,MAA-SAE),在实现 p-敏感 k-匿名模型的基础上,通过微聚集算法来降低数据概化过程中的信息损失,并提出匿名保护指数(Anonymous Protection Factor,APF)的概念,用来描述隐私保护的程度,在聚类过程中通过保证匿名保护指数最大,来提高敏感属性在聚类结果中分布的均匀程度,以此来应对"背景知识攻击".同时,在数据概化过程中,本文创造性地提出语义属性的概念,并提出一种新的数据结构-编码层次树来表示语义属性,定义了其距离度量,有效地降低了匿名过程中的信息损失.最后,分析和实验验证了该方法的有效性和合理性.

2 数据匿名的基本概念

定义 1 等价类 数据表中的一个等价类为数据表中若干元组的集合,其中每个元组在准标识符上具有相同的属性值.

定义 2 k-匿名模型 如果一个等价类至少包含 k 个元组,那么称该等价类是 k-匿名的;如果数据表中的任何一个等价类都是 k-匿名的,则称该数据表满足 k-匿名模型[16].

满足 k-匿名的数据表使攻击者不能唯一的确定某个个体的敏感属性值,但是这并没有破坏个体与敏感属性的关联关系,还可能遭遇同质攻击和背景知识攻击.例如,表 1 是原始数据表,表 2 是满足 2 – 匿名模型的发布数据表,其中的元组集合$\{t_1,t_2\}$,$\{t_3,t_4,t_5\}$,$\{t_6,t_7\}$分别是 3 个等价类.但是,如果能够从其他途径确定某病人的年龄为 24,邮编是 115000,性别是男,从表 2 中仍然可以唯一的确定该病人患有艾滋病.

表 1 原始个人医疗信息记录表

姓 名	年 龄	性 别	邮 编	疾 病
艾 强	24	男	115000	艾滋病
巴 干	25	男	115001	艾滋病
蔡 华	27	男	117000	流 感
杜 康	29	男	117001	癌 症
鄂 林	30	男	117002	肥胖症
方 芳	42	女	156000	肺 炎
古 月	45	女	156010	糖尿病

针对 k-匿名模型的不足,Truta 等人提出了 p-敏感 k-匿名(p-sensitive k-anonymity)模型,使等价类中的敏感属性值足够多样化,来避免一致性攻击.

表 2 满足 2-匿名的发布数据表

序号	年 龄	性 别	邮 编	疾 病
t_1	[24 – 25]	男	11500 *	艾滋病
t_2	[24 – 25]	男	11500 *	艾滋病
t_3	[27 – 30]	男	11700 *	流 感
t_4	[27 – 30]	男	11700 *	癌 症
t_5	[27 – 30]	男	11700 *	肥胖症
t_6	[42 – 45]	女	1560 * *	肺 炎
t_7	[42 – 45]	女	1560 * *	糖尿病

定义 3 p-敏感 k-匿名模型 如果一个等价类至少包含 k 个元组,并且在敏感属性上至少有 p 个不同取值,则称该等价类是 p-敏感 k-匿名的;如果数据表中的任何一个等价类都是 p-敏感 k-匿名的,则称该等价类满足 p-敏感 k-匿名模型.

表 3 是满足 2-敏感 2-匿名的发布数据表,它可以有效的抵御一致性攻击.尽管如此,p-敏感 k-匿名还有可能遭到背景知识攻击.例如,一个满足 3-敏感,10 匿名的数据集,如果某个等价类中 80% 以上的元组的敏感属性都是艾滋病,并且攻击者可以把某个患者关联到该等价类中,那么攻击者可以有 80% 的概率确定该患者患有艾滋病.所以,本文希望在一个等价类中,所有元组的敏感属性值能够呈现出相对均匀地分布特征,而不体现某种聚集特征.因此,本文主要研究如何在降低数据概化过程中信息损失的基础上,提高敏感属性在聚类结果中分布的均匀程度.

表 3 满足 2-敏感 2-匿名的发布数据表

序号	年 龄	性 别	邮 编	疾 病
t_1	[24 – 30]	男	11 * * * *	艾滋病
t_2	[24 – 30]	男	11 * * * *	艾滋病
t_3	[24 – 30]	男	11 * * * *	流 感
t_4	[24 – 30]	男	11 * * * *	癌 症
t_5	[24 – 30]	男	11 * * * *	肥胖症
t_6	[42 – 45]	女	1560 * *	肺 炎
t_7	[42 – 45]	女	1560 * *	糖尿病

本文将准标识符属性分为连续属性和离散属性,为了降低离散属性在数据概化过程中的信息损失,本文根据离散属性值间的极差关系和语义层次关系,创造性地将离散属性分为有序属性,标称属性和语义属性,并提出一种新的数据结构-编码层次树来表示语义属性,分别给出了不同类型属性的距离度量,以及信息损失的量化定义;然后提出一个基于敏感属性熵的微聚集算法,在实现 p-敏感 k-匿名模型的基础上,通过微

聚集算法来降低数据概化过程中的信息损失,并提出匿名保护指数的概念,在聚类过程中通过保证匿名保护指数最大,来提高敏感属性在聚类结果中分布的均匀程度,应对背景知识攻击.

3　度量空间

3.1　连续属性的距离度量

由于不同的连续属性取值的值域区间大小及单位不同,在计算不同的元组间距离时,这种差异会产生很大的影响.为了平抑属性值间的差异,需要把连续属性的取值规范化到[0,1]之间,转换的公式为

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

其中,$x_{\max}$和 $x_{\min}$分别是属性 X 取值的最大值和最小值.x_i 是元组 i 在属性 X 上的取值,x_i' 是 x_i 规范化后的值.

假设元组 i 和元组 j 在属性 X 上的值经过规范化之后分别为 x_i' 和 x_j',则元组 i 和 j 在属性 X 上的距离定义为

$$d(i,j) = \left|(x_i' - x_j')\right| \tag{2}$$

假设 G_c 是数据表的一个聚类,则 G_c 中所有元组在连续属性 X 上的质心 C 可以表示为

$$C = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

其中,n 表示 G_c 中元组的个数,x_i 表示元组 i 在属性 X 上的取值.

3.2　离散属性的距离度量

为了计算离散属性值之间的距离,必须将离散属性值数值化.文献[17]对离散属性进行了介绍,并根据属性值之间的极差关系,将离散属性分为了有序属性和分类属性,但是,该划分忽略了离散属性值间的语义层次关系,造成了离散属性划分的不准确.针对该问题,本文同时考虑离散属性值间的极差关系和语义层次关系,创造性地将离散属性分为有序属性,标称属性和语义属性.

定义 4　有序属性(Ordinal Attribute,OA)　如果属性的取值间有一定的级差关系或者大小关系,则称该属性为有序属性.

例如,"排名奖励"的取值"金牌"、"银牌"、"铜牌"之间是有级差的,"金牌"和"银牌"之间的距离应该比"金牌"和"铜牌"之间的距离要小.因此,"排名奖励"属性为一个有序属性.为了定量地衡量有序属性的取值,我们用有序属性的秩来表示[17].参照 Domingo Ferrer 给出的距离度量方法[18],本文对有序属性的距离度量进行了改进.属性秩间的权重记为 $W_{1,2}, W_{2,3}, \cdots, W_{n-1,n}$ $(W_{i-1,i} > 0, i = 2,3,\cdots,n)$,权重表示属性值间的相异程度.对于有序属性 A_i 的不同值 $a, b(a \leqslant b)$,它们的距离定义为

$$\mathrm{dis}_{\mathrm{OA}}(a,b) = \frac{\sum_{j=a+1}^{b} W_{j-1,j}}{\sum_{j=2}^{|D(A_i)|} W_{j-1,j}} \tag{4}$$

其中,$D(A_i)$为 A_i的值域.

为了更多地保留有序属性的语义信息,本文采用凸中位数[18]定义有序属性的质心.设有序属性 A_i 的值域 $D(A_i) = \{c_1, c_2, \cdots, c_m\}$,$n$ 个数 $\{a_1, a_2, \cdots, a_n\}$ $(a_j \in D(A_i))$组成一个类,设 $f(c_j)$为类中 c_j 出现的频率,则频率函数 f'为

$$f'(c_j) = \min(\max(f_{c_j \leq c_i}(c_i)), \max(f_{c_j \geq c_i}(c_i))) \tag{5}$$

定义 5　标称属性(Nominal Attribute,NA)　如果属性的取值间既没有级差关系,也没有大小关系,则称该属性为标称属性.

例如"职位"的取值可以有"教师","工人","记者"等,它们之间是没有级差关系和大小关系的.因此,"职位"属性是一个标称属性.本文使用文献[19]提出的定类离散属性的数值化转换方法对标称属性进行预处理,并使用一个 $n \times n$ 的方阵来表示处理过的具有 n 个不同取值的标称属性."职位"属性表示见表 4.

表 4　标称属性变量"职位"的数值转换表

职位	S_1	S_2	S_3
教师	$\sqrt{0.5}$	0	0
工人	0	$\sqrt{0.5}$	0
记者	0	0	$\sqrt{0.5}$

转换后 2 个元组 X 和 Y 在该属性上的距离为

$$\mathrm{dis}_{\mathrm{NA}}(X,Y) = (S_{1x} - S_{1y})^2 + (S_{2x} - S_{2y})^2 + (S_{3x} - S_{3y})^2 \tag{6}$$

假设 G_n 是数据表的一个聚类,则 G_n 中所有元组在标称属性 X 上的质心 Z 可以表示为

$$Z = \sum_{j=1}^{m} x_j \tag{7}$$

其中,m 表示聚类 G_n 中元组的个数,x_j 表示第 j 个向量在属性 X 上的取值.

定义 6　语义属性(Semantic Attribute,SA)　如果属性的取值之间没有明显的级差关系和大小关系,但是属性的取值之间具有一定的语义层次关系,并且可以形式化表示为:$\forall e_i, e_j \in S, e_i = (r_{i1}, r_{i2}, \cdots, r_{im}), e_j = (r_{j1}, r_{j2}, \cdots, r_{jm}), (r_{i1}, r_{i2}, \cdots, r_{im} \in R, r_{j1}, r_{j2}, \cdots, r_{jm} \in R), \forall 0 < i, j, k \leqslant m, (r_{ik}, r_{jk} \in N)$,其中,$S$ 表示语义属性,R 表示有序属性,N 表示标称属性,即任何一个语义属性值都可以看做是 m 个互不相交的有序属性值的集合,任何两个语义属性值划分成 m 个集合后,对应的集合元素值可以看做是标称属性,则称该属性为语义属性.

语义属性具有以下特点：

(1)属性值可以划分为多个部分,这些部分之间具有一定的层次关系；

(2)相同前缀越多的属性值之间彼此相似程度越高,反之属性值之间相似程度越低.

例如,“邮政编码”属性的取值{150001}与{150002}间的距离明显比{150001}与{061110}之间的距离要近的多.因此,“邮政编码”属性是一个语义属性.

鉴于语义属性的以上特点,不能简单地使用有序属性和标称属性的距离度量来衡量两个语义属性值间的距离.

文献[20]提出了泛化层次树的概念,可以用来计算离散属性的距离,但是泛化层次树只能计算层次间的节点距离,对同一层次节点距离的计算无能为力;文献[21]提出了编码层次平衡树的概念,但是编码层次平衡树只能用于计算叶子节点间的距离,对层间节点距离的计算无能为力.

考虑到语义属性值之间的语义层次关系,既要计算层次间的节点距离,也要计算叶子节点间的距离的要求,针对编码层次平衡树和泛化层次树的不足,本文提出了一种新的数据结构-编码层次树(Code Hierarchy Tree,CHT),用来表示语义属性.它综合了编码层次平衡树和泛化层次树的优点,能够较准确地衡量编码层次树中任意两个节点间的距离.

本文将语义属性值集合转换为一棵编码层次树 T,任意一个语义属性 P 的取值可以看作是从树 T 的根到叶子节点所经过的所有节点的有序排列(根节点除外),记作 $T(P)$.编码层次树 T 上每层节点间的权重记为 $W_{1,2}, W_{2,3}, \cdots, W_{n-1,n}$ ($W_{i-1,i}>0, i=2,3,\cdots,n$),权重表示节点间的相异程度,越小表示节点越相似.图1为“邮政编码”属性的编码层次树.

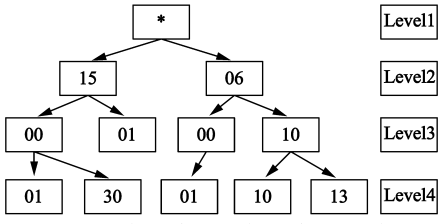


图1 “邮政编码”属性的编码层次树

本文使用如下步骤来计算编码层次树中任意两个节点 P_1 和 P_2 间的距离：

(1)求两个待比较节点 P_1和 P_2的最近公共祖先节点 C,两个节点 P_1 和 P_2 相等当且仅当 $\mathrm{dis}_{\mathrm{NA}}(P_1,P_2)=0$

(2)使用加权层次距离分别求两个待比较节点 P_1 和 P_2 与最近公共祖先节点 C 的距离 $\mathrm{dis}(P_1,C)$和 $\mathrm{dis}(P_2,C)$；

P_1 和 P_2 节点间的距离 $\mathrm{dis}_{\mathrm{SA}}(P_1,P_2)=\mathrm{dis}(P_1,C)$和 $\mathrm{dis}(P_2,C)$；

节点 P 与 C 之间的距离可表示为两个有序属性之间的距离：

$$\mathrm{dis}_{\mathrm{OA}}(P,C)=\frac{\sum_{j=\mathrm{level}(C)+1}^{\mathrm{level}(P)} W_{j-1,j}}{\sum_{j=2}^{h} W_{j-1,j}} \tag{8}$$

其中,level(P)和 level(C)分别表示节点 P 和 C 在编码层次树中的层数,h 表示树的高度.

文献[21]给出了权重 $W_{j-1,j}$定义的两种形式：

(1) $W_{j-1,j}=1,(2\leqslant j\leqslant h)$；

(2) $W_{j-1,j}=1/(j-1)^{\beta},(2\leqslant j\leqslant h)$；$\beta$ 由用户指定

考虑到形式(1)是形式(2)的一种特例($\beta=0$),本文在实验阶段采用权重 $W_{j-1,j}$定义的第二种形式,并取值 $\beta=1$.

为了将 P_1 和 P_2 之间的距离规范化到[0,1]之间,本文把 P_1 和 P_2 之间的距离作与连续属性一样的规范化处理.设编码层次树上两个节点最大距离和最小距离分别为 $L_{\max}$和 $L_{\min}$,则规范化后的距离为：

$$\mathrm{dis}_{\mathrm{SA\text{-}Normal}}(P_1,P_2)=\frac{\mathrm{dis}_{\mathrm{SA}}(P_1,P_2)-L_{\min}}{L_{\max}-L_{\min}} \tag{9}$$

注意:在编码层次树中,根节点的层数为 1,高为 h 的树的叶子节点 i 的层数为 h_i;根节点始终为“*”,只是一个形式上的表示,没有任何意义.在实验中,权重 $W_{1,2}$通常设为 0.

采用编码层次树来表示语义属性,语义属性的属性值间的语义层次关系能够较完整的通过编码层次树的层次关系表示出来,从而能够更准确地度量语义属性值间的相似性,有效地降低匿名过程中的信息损失.

为了降低聚类过程中离群点的干扰,本文使用语义属性值集合的中心点作为集合的质心.语义属性值集合的中心点实质上就是到集合内其他语义属性值距离之和最小的属性值.

定义 7 语义属性值集合的中心点 假设 S 为一个语义属性,如果满足

$$S=\mathrm{Argmin}_{e_i\in G_s}\{D_i \mid D_i=\sum_{e_j\in G_s, i\neq j}|\mathrm{dis}_{\mathrm{SA\text{-}Normal}}(e_i,e_j)|\} \tag{10}$$

则称 S 为集合 G_s的中心点.其中 G_s表示语义属性值的集合,i,j 为整数,且 $0\leqslant i,j<|G_s|, i\neq j, e_i, e_j$ 为 G_s 中任意两个语义属性值,D_i 表示语义属性 e_i 到集合 G_s 中其他语义属性距离之和.

3.3 信息损失度量

本文使用微聚集的方法,对数据表中的元组进行匿名化操作,当用类质心取代元组值的时候,必然会产生信息损失.本文希望在匿名化的过程中,产生的信息

损失最小,以提高数据的可用性.下面在距离度量的基础上,定义信息损失度量.

定义 8　元组间的距离　设 $t_1, t_2 \in T$,则元组 t_1 和 t_2 间的距离为其在所有准标识符上的距离之和:

$$\mathrm{dis}(t_1, t_2) = \sum_{i=1}^{|\mathrm{QI}|} d(t_{1i}, t_{2i}) \tag{11}$$

定义 9　概化元组的信息损失　设 $t, c \in G$, c 为 G 中的质心,则元组 t 被概化产生的信息损失为元组 t 和质心 c 之间的距离 $\mathrm{dis}(t, c)$.

定义 10　概化类的信息损失　设元组 t 被概化产生的信息损失为 $\mathrm{dis}(t, c)$,则 G 中的所有元组被概化产生的信息损失记为 $\mathrm{IL}(G)$,表示如下:

$$\mathrm{IL}(G) = \sum_{t \in G} \mathrm{dis}(t, c) \tag{12}$$

其中,t 为类 G 中任意元组.

定义 11　概化类的增量信息损失　设等价类 G 中的所有元组被概化产生的信息损失记为 $\mathrm{IL}(G)$,将任意集合 X 加入到类 G 中,形成一个新的类 G',则概化类的增量信息损失 $\mathrm{ILA}(G, G')$可表示如下:

$$\mathrm{ILA}(G, G') = \mathrm{IL}(G') - \mathrm{IL}(G) \tag{13}$$

3.4　敏感属性熵增量

为了更好地保护隐私数据,本文希望等价类中元组的敏感属性值尽可能均匀地分布.为了衡量敏感属性值在等价类中的分布情况,本文引入了熵的概念.

定义 12　熵(entropy)　假设数据表的等价类 G 中包含 n 个不同的敏感属性值 $s_1, s_2, \cdots, s_n$, p_i 表示敏感属性值 s_i 的概率,则定义等价类 G 的熵 $\mathrm{Ent}(G)$为:

$$\mathrm{Ent}(G) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{14}$$

定义 13　敏感属性熵增量　假设 G 是数据表的一个类,将元组 t 加入到类 G 中,形成新的类 G',产生的敏感属性熵的增量 $\mathrm{EA}(G, G')$定义为:

$$\mathrm{EA}(G, G') = \mathrm{Ent}(G') - \mathrm{Ent}(G) \tag{15}$$

3.5　匿名保护指数

为了准确地描述隐私保护的程度,本文提出了匿名保护指数(Anonymous Protection Factor, APF)的概念.

定义 14　匿名保护指数　设 G 为数据表的一个类,G' 为类 G 加入元组 t 后形成的新的类,则类 G 的匿名保护指数可以表示为:

$$\mathrm{APF}(G, G') = \frac{\mathrm{EA}(G, G')}{\mathrm{ILA}(G, G')} \tag{16}$$

显然敏感属性熵增量越大,信息损失越小,匿名保护指数越大,隐私保护的效果越好.

4　数据匿名方法

4.1　算法实现

本文针对 p-敏感 k-匿名模型中,可能存在的敏感属性值在聚类结果中分布不均匀,而导致的敏感信息泄露问题,采用 L-clustering[22]算法的匿名过程,提出了一种基于敏感属性熵的微聚集算法(Micro Aggregation Algorithm based on Sensitive Attribute Entropy, MAA-SAE),其主要思想为:将匿名保护指数作为衡量两个元组相似性的标准,这样既能保证较小的信息损失,又能保证聚类结果中敏感属性值较均匀分布.在聚类过程中,保证类内元组最大程度的相似,同时保证同一个聚类内至少有 p 个不同的敏感属性值以及 k 个元组,保证 p-敏感 k-匿名模型的实现.算法的伪码描述如下:

算法 1　MAA-SAE

输入:原始数据表 T,准标识符 QI,敏感属性 S, p-敏感 k-匿名模型的参数 p 和 k($p > 1, k > 1$)

输出:匿名数据表 T^*

```
BEGIN
1. Q = Φ; // Q 表示等价类的集合
2. IF((T 中元组的数量小于 k)OR(T 中不同敏感属性值的个数小于 p)OR(p 的值大于 k 的值))
3. RETURN;
4. WHILE((T 中元组个数大于等于 k)AND(T 中不同敏感属性值的个数不小于 p)):
5. 随机的从 T 中选取元组 t, T = T - {t},生成类 G = {t};
6.   WHILE(G 中元组个数小于 k)
7.     WHILE(G 中元组个数小于 p)
8.       t' = Argmax_{(t∈T)∧(∀t_j∈G, t.s≠t_j.s)}(APF(G, G∪{t}))
9.       G' = Argmax_{G_k∈Q}(APF(G, G∪G_k));
10.      IF(APF(G, G∪{t'}) > APF(G, G∪G')),
11.        THEN T = T - {t'}; G = G∪{t'};
12.      ELSE Q = Q - {G'}, G = G∪G';
13.    END WHILE;
14.    WHILE(G 中元组个数不小于 p,且元组个数小于 k):
15.      t' = Argmax_{t∈T}(APF(G, G∪{t}));
16.      G' = Argmax_{G_k∈Q}(APF(G, G∪G_k));
17.      IF(APF(G, G∪{t'}) > APF(G, G∪G')),
18.          THEN T = T - {t'}; G = G∪{t'};
19.      ELSE Q = Q - {G'}, G = G∪G';
20.    END WHILE;
21.  END WHILE;
22.  Q = Q∪G;
23. END WHILE;
24. WHILE(T 中仍有剩余元组)
25.   随机的从 T 中选取元组 t, T = T - {t};
26.   G = Argmax_{G_k∈Q}(APF({t}, {t}∪G_k));
27.   Q = Q - {G}, G = G∪{t};
28.   Q = Q∪G;
29. END WHILE;
30. 依次处理 Q 中的每个类,将类中的每个元组在准标识符上的属性用该类的质心属性所代替,得到匿名数据表 T*.
END
```

算法 MAA-SAE 首先判断输入数据表 T 中的元组数量是否小于 k,T 中不同敏感属性取值的个数是否小于 p(步骤 2),如果有一个条件满足,那么算法肯定不能产生一个满足 p-敏感 k-匿名的匿名数据表,算法返回(步骤 3).

步骤 4 ~ 23 是本算法的关键,当 T 中的元组数量不小于 k 且 T 中不同敏感属性取值的个数不小于 p 时,算法重复执行步骤 5 ~ 23,每次得到一个类 G,使类 G 中至少有 k 个元组,并且至少含有 p 个不同的敏感属性值.具体的方法是:首先随机的从 T 中选择元组 t,并初始化类 G;然后在步骤 8 ~ 12 中贪婪的选择 T 中匿名保护指数最大的元组 t',并且保证 t' 的敏感属性值与 G 中元组的敏感属性值均不相同.同时,从已生成的类集合 Q 中选取距 G 最近的类 G'(注意:Q 中的任意一个类都满足 p-敏感 k-匿名模型).如果 $\mathrm{APF}(G, G\cup\{t'\}) > \mathrm{APF}(G, G\cup G')$,则将 t' 从 T 中删除并加入到 G 中;否则从 Q 中删除 G',并将 G' 合并到 G 中.这一过程重复执行,直到 G 中有不少于 p 个元组为止;然后,步骤 14 ~ 20,首先判断 G 中的元组个数是否不小于 p 并且小于 k,如果是的话进入循环,在步骤 15 ~ 19,进行与步骤 8 ~ 12 中相同策略的贪婪选择,直到 G 中有不少于 k 个元组为止;最后,将 G 加入到 Q 中.

在步骤 23 结束后,如果 T 中仍有剩余元组,算法在步骤 24 ~ 29 将剩余的元组逐一加入到匿名保护指数最大的已有类中.最后,算法在步骤 30 概化每个元组在准标识符上的属性值,得到满足 p-敏感 k-匿名模型的数据表 T^*.

4.2 算法的合理性和复杂性分析

4.2.1 算法的合理性分析

通过算法实现和分析,不难看出,如果 T 中元组个数不小于 k,并且 T 中不同敏感属性值的个数不小于 p,则 MAA-SAE 算法的匿名数据表 T^* 一定满足 p-敏感 k-匿名模型.实际上,当步骤 23 完成时,算法得到类的集合 Q,使 Q 中的每个类至少有 k 个元组,并且至少包括 p 个不同的敏感属性值,随后的步骤 24 ~ 29 继续保持了这一结果.步骤 30 将每个类概化成为一个等价类,从而得到匿名数据表 T^*,满足 p-敏感 k-匿名模型的匿名要求.

4.2.2 算法的复杂性分析

设原始数据表 T 中元组数为 n,准标识符维数为 d,算法在步骤 23 完成后得到 m 个类.

算法在步骤 2,3 检测 T 中元组个数以及不同的敏感属性值个数,只需要扫描一次数据表 T,执行时间为 $O(n)$.

算法在步骤 4 ~ 23 中,每得到一个新类 G,最多扫描 k 遍 T 和 Q,并且计算在准标识符 QI 上的距离.因为在算法执行过程中,$|T| + |Q| \leqslant n$,所以每生成一个新类用时不超过 $O(dkn)$.步骤 4 ~ 23 共生成 m 个类,执行时间为 $O(dkmn)$.

步骤 23 结束后得到 m 个类,每个类至少 k 个元组,所以至多剩余 n-mk 个元组,每次循环需要扫描 Q 一遍,并计算在准标识符 QI 上的距离,执行时间为 $O(dm)$.所以,步骤 24 ~ 29 的执行时间为 $O(dm(n-mk))$.

算法在步骤 30 生成结果数据集,需要扫描所有元组一遍,同时对类中的每个元组进行概化操作,所以执行时间为 $O(dn)$.

因此,MAA-SAE 算法总体执行时间为 $O(n) + O(dkmn) + O(dm(n-mk)) + O(dn)$,因为 $km < n$,所以在最坏情况下,MAA-SAE 算法的时间复杂度为 $O(dn^2)$.

5 实验及结果分析

5.1 实验数据及参数

本文使用 UCI Machine Learning Repository 中的 Adult 数据集作为实验数据集,该数据集由美国人口普查数据构成,合并训练集和测试集,并去除缺省值记录后,共有 45222 条记录,包含 15 个属性值.为了方便研究,本文保留数据集的 8 个属性:Age, Gender, Race, Education, Native Country, Work Class, Fnlwgt, Occupation.其中,Age 作为连续型属性,Gender, Race, Education, Native Country, Work Class 作为离散型变量的标称属性,Fnlwgt 作为离散型变量的语义属性,并且只保留长度为 6 的记录,Occupation 作为敏感属性.实验的硬件环境为:Intel(R) Core(TM)2 Quad CPU Q8400 @2.66GHz 2.67GHz, 2.00GB内存,操作系统为 Microsoft Windows 7,算法均在 Visual Studio 2005 下实现.

为了对数据匿名结果进行综合评估,本文提出了匿名结果的平均信息损失和平均敏感属性熵的概念,定义如下:

定义 15 **平均信息损失** 设 $Q = \{G_1, G_2, \cdots, G_n\}$ 是聚类形成的等价类集合,$G_i(i = 1,2,\cdots,n)$ 是 Q 中任一等价类,根据式(12)可以得出 G_i 的信息损失 $\mathrm{IL}(G_i)$,则 G_i 的平均信息损失表示为:

$$\mathrm{AVG\_IL}(G_i) = \frac{\mathrm{IL}(G_i)}{|G_i| \cdot |\mathrm{QI}|} \tag{17}$$

其中,$|G_i|$ 表示 G_i 中的元组个数,$|\mathrm{QI}|$ 表示数据集中准标识符的数量.

在此基础上,匿名结果的平均信息损失表示为:

$$\mathrm{AVG\_IL}(Q) = \frac{\sum_{G_i \in Q} \mathrm{AVG\_IL}(G_i)}{|Q|} \tag{18}$$

其中,$|Q|$ 表示集合 Q 中的聚类个数.

定义 16　平均敏感属性熵　设 $Q=\{G_1,G_2,\cdots,G_n\}$是聚类形成的等价类集合,$G_i(i=1,2,\cdots,n)$是 Q 中任一等价类,根据式(14)可以得到 G_i 的敏感属性熵 $\mathrm{Ent}(G_i)$,则匿名结果的平均敏感属性熵表示为:

$$\mathrm{AVG\_Ent}(Q)=\frac{\sum_{G_i\in Q}\mathrm{Ent}(G_i)}{|Q|} \tag{19}$$

其中,$|Q|$表示集合 Q 中的聚类个数.

显然,平均敏感属性熵越大,聚类结果中敏感属性的分布越均匀,隐私泄露风险越小.

在数据发布的隐私保护中,数据可用性和隐私保护程度是两个对立的概念,想要获得较高的隐私保护程度就势必会降低数据的可用性,反之亦然.为了更好的度量发布数据的可用性,本文使用 CAVG 作为数据可用性度量标准.CAVG 是 LeFevre 等人[23]提出的,表示为:

$$\mathrm{CAVG}=\left(\frac{\text{total records}}{|Q|}\right)/k \tag{20}$$

其中 total records 表示数据集中包含的全部元组个数,$|Q|$表示集合 Q 中的聚类个数.CAVG 值越小,数据可用性越高.

5.2　信息损失和隐私保护程度分析

为了验证本文提出方法的合理性,本文分别对 k 和 p 取若干组值,并对每一组 k,p 取值进行了 10 次重复实验,取其平均值.由于具体的 k 和 p 的取值对本实验的影响不大,本文仅取 $k=8,10,12$;$p=5,6,7$ 条件下的结果作为代表.实验结果见表 5.

表 5　Adult 数据集实验结果

k	p	AVG_IL	AVG_Ent
8	5	0.14831	3.01804
8	6	0.15413	3.03963
8	7	0.16455	3.05008
10	5	0.19341	3.23690
10	6	0.19403	3.28224
10	7	0.19521	3.31970
12	5	0.21557	3.34189
12	6	0.21878	3.40647
12	7	0.21946	3.47562

从表 5 的实验结果可以看出,当 k 一定时,随着 p 的增大,平均信息损失逐渐增加,平均敏感属性熵也在增加,系统的隐私保护程度在增加;当 p 一定时,随着 k 的增加,平均信息损失在增加,平均敏感属性熵在增加,系统的隐私保护程度在增加.这说明,增加系统的隐私保护程度与降低信息损失是彼此矛盾的.

为了便于理解匿名保护指数和信息损失的关系,本文修改 MAA-SAE 算法,不要求其在聚类时匿名保护指数最大,改为要求信息损失最小(该方法记为 MAA-MINIL 算法),并比较以上 2 个算法的平均信息损失和平均敏感属性熵,实验结果如图 2 所示.

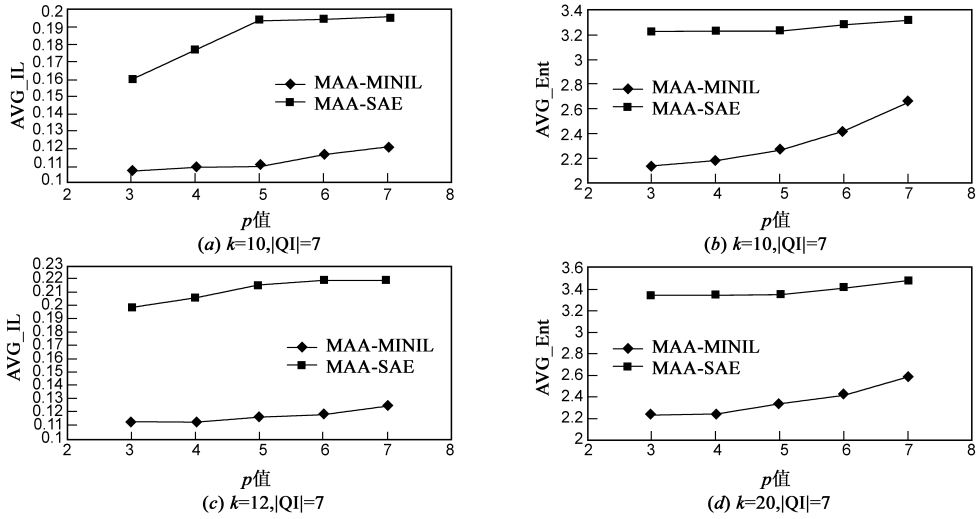
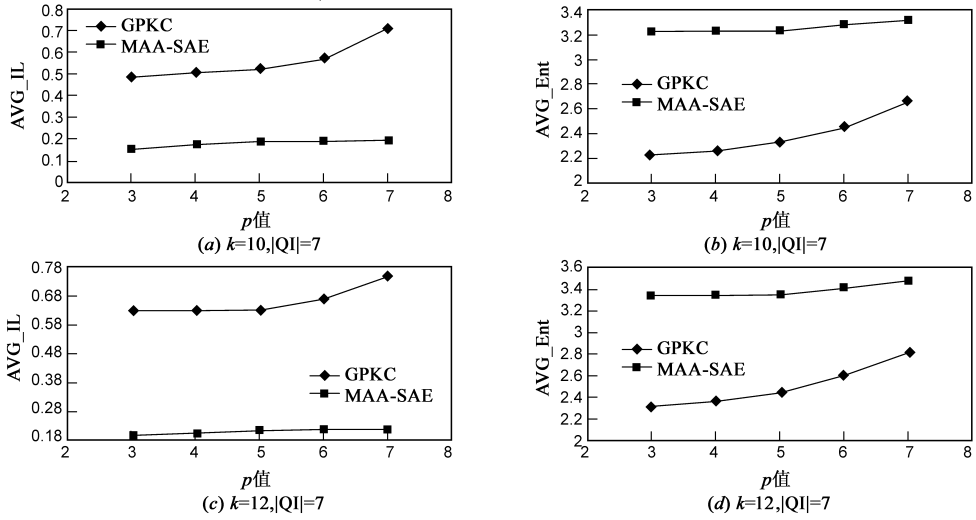


图2　p值变化下的平均信息损失和平均敏感属性熵

由图 2 可以看出,当 k 值一定时,随着 p 值的增加,AVG_IL 和 AVG_Ent 都在增加,但是 MAA-SAE 算法比 MAA-MINIL 算法的 AVG_IL 和 AVG_Ent 都高.这是由于 MAA-MINIL 算法要求信息损失的增量最小,所以降低了信息损失;而 MAA-SAE 算法,由于要保证匿名保护指数最大的同时敏感属性熵尽量大(即要求聚类结果中敏感属性的分布尽量均匀,隐私保护程度高),因此信息损失稍大.相对于 MAA-MINIL 算法,MAA-SAE 算法有稍大的平均信息损失,同时有更大的敏感属性熵,更高的隐私保护程度;这也同样表明,增加系统的隐私保护程度与降低信息损失是彼此矛盾的,增加隐私保护程度会相应的增加信息损失,反之亦然.

为了充分验证 MAA-SAE 算法的性能,本文实现了文献[24]的算法,记作 GPKC,在|QI| = 7,k 值分别取 10 和 12 的情况下,关于 AVG_ IL 和 AVG_ Ent 分别作了 4 组对比实验,实验结果如图 3 所示.



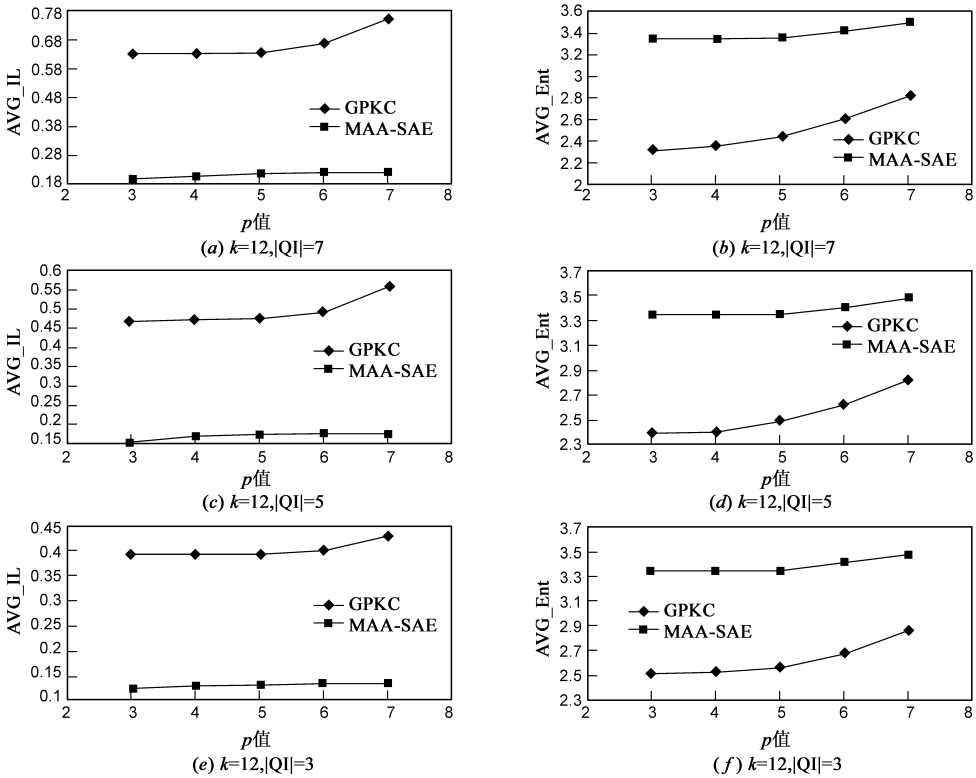
(a) k=10,|QI|=7 (b) k=10,|QI|=7

(c) k=12,|QI|=7 (d) k=12,|QI|=7

图3 p值变化下的平均信息损失和平均敏感属性熵

从图 3 中可以看到,在同等条件下,MAA-SAE 算法比 GPKC 算法的信息损失要小的多,而且随着 p 的增长,MAA-SAE 算法的平均信息损失增长很缓慢,GPKC 算法的平均信息损失增长很快,这是因为在数据概化的过程中,编码层次树能够更好地衡量语义属性之间的距离,减小概化过程中的信息损失;另一方面,MAA-SAE 算法比 GPKC 算法的平均敏感属性熵要大的多,说明 MAA-SAE 算法比 GPKC 算法具有更均匀的敏感属性分布,更好的隐私保护效果.

为了验证|QI|的取值对实验结果的影响程度,本文做了以下对比试验,实验结果如图 4 所示.



(a) k=12,|QI|=7 (b) k=12,|QI|=7

(c) k=12,|QI|=5 (d) k=12,|QI|=5

(e) k=12,|QI|=3 (f) k=12,|QI|=3

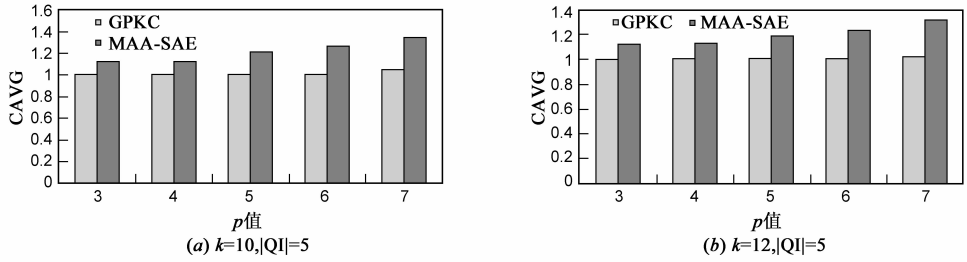
图4 p值变化下的平均信息损失和平均敏感属性熵

图 4 中的各个图给出了当 $k=12$,|QI|分别为 3,5,7 时,两种方法的平均信息损失和平均敏感属性熵,从图 4 中可以看到,在同等条件下,MAA-SAE 算法比 GPKC 算法的信息损失要小的多,随着|QI|的增加,信息损失也在增加,这是因为准标识符中包含的属性越多,微聚集操作的时候损失的信息越大;另一方面,MAA-SAE 算法比 GPKC 算法的平均敏感属性熵要大的多,而且随着|QI|的增加,敏感属性熵的最大值变化很小,说明当 p 值增加到一定程度的时候,增加|QI|也很难提高隐私保护的程度.经实验验证,$k=12$,$p=7$ 时可以取得较好的实验效果.

综合图 3 和图 4 可以发现,当|QI|取相同值时,随着 k 值的增加,MAA-SAE 和 GPKC 算法的平均信息损失都在增加,这是因为随着 k 值的增加,在数据概化过程中,需要考虑更多的元组取值,因此增加了平均信息损失.

5.3　数据可用性分析

图 5 给出了当准标识符维数|QI| = 5 和 k 分别取 10 和 12 时,p 值变化对 MAA-SAE 和 GPKC 数据可用性的影响.



(a) k=10,|QI|=5　　(b) k=12,|QI|=5

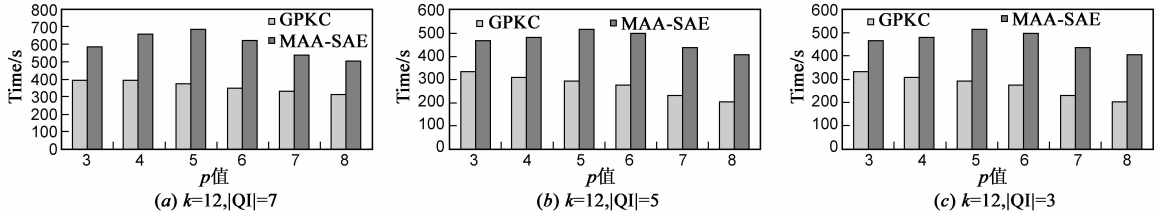
图5　p值变化下的CAVG

从图 5 中可以看到,随着 p 值的增长,MAA-SAE 算法和 GPKC 算法的 CAVG 在增加,数据可用性都在降低,这是因为随着 p 值的增加,在聚类过程中需要考虑的不同敏感属性值增加,增加了聚类的复杂程度,导致数据可用性降低;而且,在同等条件下,MAA-SAE 算法比 GPKC 算法的数据可用性要稍小一些,这是因为 MAA-SAE 算法为了提高匿名数据的隐私保护效果,在聚类过程中,不仅需要判断新聚类是否匿名保护指数最高,还要判断新的元组加入已生成的聚类是否匿名保护指数最高,因此降低了数据的可用性.

通过图 5 还可以发现,当|QI|和 p 取相同值时,随着 k 值的增加,MAA-SAE 和 GPKC 算法的 CAVG 都在降低,数据可用性增加,这是因为随着 k 值的增加,聚类数量减少的不是很多,所以 CAVG 值减少,数据可用性增加.

5.4　执行时间分析

图 6 给出了当准标识符维数|QI|和 k 值固定时,p 值变化对 MAA-SAE 和 GPKC 执行时间的影响.



(a) k=12,|QI|=7　　(b) k=12,|QI|=5　　(c) k=12,|QI|=3

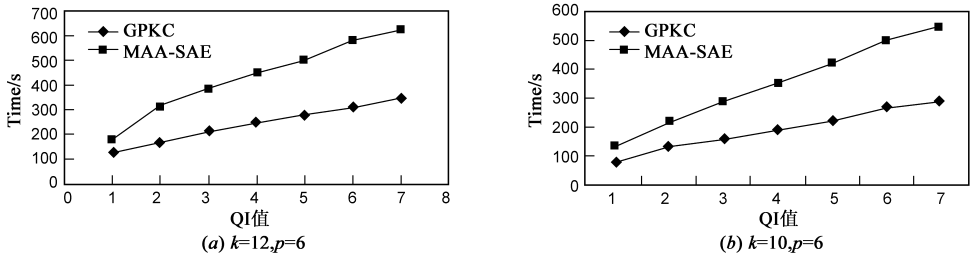
图6　p值变化下的执行时间

对于 GPKC 算法,其执行时间随着 p 值的增加表现为减少的情况.这是因为在 GPKC 的每个类中,第一个元组是随机选取的,而随后加入元组到类中,则需要多次进行距离计算来找到距离最小的元组或类,因此花费时间较长.并且由于 GPKC 要求类中元组在敏感属性上至少有 p 个不同值.随着 p 的增加,满足匿名模型的元组数目减少,通过敏感属性值是否相同就可以排除大多数元组,减少了计算时间.

对于 MAA-SAE 算法,其执行时间随着 p 值的增加表现为先增加而后减少的情况,而且比 GPKC 算法花费更多的时间.MAA-SAE 算法的执行时间更长这很好理解:为了提高匿名数据的隐私保护效果,MAA-SAE 算法在聚类过程中,不仅需要判断新聚类是否匿名保护指数最高,还要判断新的元组加入已生成的聚类是否匿名保护指数最高,增加了算法的执行时间;当 p 值增加时,MAA-SAE 算法产生的类的数目将减少,特别是当 p 较小时,类的数目将随 p 的增加而显著减少.这意味着在聚类时,更多的元组需要进行多次距离计算.因此,当 p 较小时,MAA-SAE 算法的总体执行时间随着 p 值的增加而增加.与此同时,因为 MAA-SAE 算法要求类中元组在敏感属性上至少有 p 个不同值.随着 p 的增加,满足匿名模型的元组组合数目减少.通过敏感属性值

是否相同就可以排除大多数元组,减少了计算时间.因此,随着 p 的增加,MAA-SAE 算法的执行时间先增加而后减少.

图 7 给出了当 k 和 p 值固定时,准标识符维数|QI|变化对 MAA-SAE 和 GPKC 执行时间的影响.随着|QI|的增加,两者的执行时间都有所增加,但是 MAA-SAE 所用的时间更多,时间增长的速度更快.这是因为,MAA-SAE 通过考察元组与类之间以及类与类之间的距离寻找合适的概化方案,以较小的信息损失来满足匿名保护的需求,花费的时间更多.



(*a*) k=12,p=6 (*b*) k=10,p=6

图7 准标识符维数|QI|变化下的执行时间

6 结论

针对 p-敏感 k-匿名模型中,可能存在的敏感属性值在聚类结果中分布不均匀,而导致的敏感信息泄露问题,本文提出一种基于敏感属性熵的微聚集算法,并提出匿名保护指数来描述隐私保护程度,在聚类过程中通过保证匿名保护指数最大,来提高敏感属性在聚类结果中分布的均匀程度;同时,在数据概化过程中,本文创造性的提出语义属性的概念,并提出一种新的数据结构-编码层次树来表示语义属性,并定义其距离度量,有效地降低了匿名过程中的信息损失.分析和实验验证了该方法的有效性和合理性.

参考文献

[1] 韩建民,岑婷婷,虞慧群.数据表 k-匿名化的微聚集算法研究[J].电子学报,2008,36(11):2021 – 2029.
HAN Jian-min, CEN Ting-ting, YU Hui-qun. Research in microaggregation algorithms for k-anonymization[J]. Acta Electronica Sinica, 2008, 36(11): 2021 – 2029. (in Chinese)

[2] 周水庚,李丰,陶宇飞,等.面向数据库应用的隐私保护研究综述[J].计算机学报,2009,32(5):847 – 861.
ZHOU Shui-geng, LI Feng, TAO Yu-fei, et al. Privacy preservation in database applications: a survey[J]. Chinese Journal of Computers, 2009, 32(5): 847 – 861. (in Chinese)

[3] 朱青,赵桐,王珊.面向查询服务的数据隐私保护算法[J].计算机学报,2010,33(8):1315 – 1323.
ZHU Qing, ZHAO Tong, WANG Shan. Privacy preservation algorithm for service-oriented information[J]. Chinese Journal of Computers, 2010, 33(8): 1315 – 1323. (in Chinese)

[4] Saygin Y, Verykios V S, Elmagarmid A K. Privacy preserving association rule mining[A]. Proceedings of the 12th International Workshop on Research Issues in Data Engineering (RIDE)[C]. San Jose, USA: IEEE Computer Society, 2002. 151 – 158.

[5] Aggarwal C C, Yu P S. A condensation approach to privacy preserving data mining[A]. Proceedings of the 9th International Conference on Extending Database Technology (EDBT)[C]. Heraklion, Greece: Springer, 2004. 183 – 199.

[6] Yao A C. How to generate and exchange secrets[A]. Proceedings of the 27th IEEE Symposium on Foundations of Computer Science (FOCS)[C]. Toronto, Canada: IEEE Press, 1986. 162 – 167.

[7] Clifton C, Kantarcioglou M, Lin X, Zhu M Y. Tools for privacy preserving distributed data mining[J]. ACM SIGKDD Explorations, 2002, 4(2): 28 – 34.

[8] 韩建民,于娟,虞慧群.面向敏感值的个性化隐私保护[J].电子学报,2010,38(7):1723 – 1728.
Han Jian-min, Yu Juan, Yu Hui-qun. Individuation privacy preservation oriented to sensitive values[J]. Acta Electronica Sinica, 2010, 38(7): 1723 – 1728. (in Chinese)

[9] 杨静,王波.一种基于最小选择度优先的多敏感属性个性化 l-多样性算法[J].计算机研究与发展,2012,49(9):2603 – 2610.
YANG Jing, WANG Bo. Personalizedl-diversity algorithm for multiple sensitive attributes based on minimum selected degree first[J]. Journal of Computer Research and Development, 2012, 49(9): 2603 – 2610. (in Chinese)

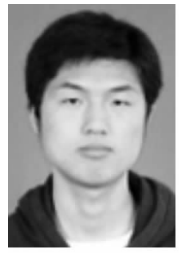
[10] 韩建民,于娟,虞慧群.面向数值型敏感属性的分级 l-多样性模型[J].计算机研究与发展,2011,48(1):147 – 158.
Han Jian-min, Yu Juan, Yu Hui-qun. A multi-levell-diversity model for numerical sensitive attributes[J]. Journal of Computer Research and Development, 2011, 48(1): 147 – 158. (in Chinese)

[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 Sini-

ca,2012,40(5):883 – 890.(in Chinese)
[12] Laszlo M, Mukherjee S. Minimum spanning tree partitioning algorithm for microaggregation[J]. IEEE Transactions on Knowledge and Data Engineering,2005,17(7):902 – 911.
[13] Solanas A, Martinez-Balleste A, Domingo-Ferrer J, et al. A 2^d-tree-based blocking method for microaggregating very large data sets[A]. Proc of the First International Conference on Availability, Reliability and Security[C]. Vienna, Australia: IEEE Press,2006.922 – 928.
[14] 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.
[15] Truta T, Vinay B. Privacy protection: *p*-sensitive *k*-anonymity property[A]. Proc of the 22nd International Conference on Data Engineering Work Shops[C]. Washington DC, USA: IEEE Computer Society,2006.94 – 103.
[16] Sweeney L. *k*-anonymity: A model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems,2002,10(5):557 – 570.
[17] 杨高明,杨静,张健沛.半监督聚类的匿名数据发布[J].电子学报,2011,32(11):1489 – 1494.
YANG Gao-ming, YANG Jing, ZHANG Jian-pei. Semi-supervised clustering-based anonymous data publishing[J]. Acta Electronica Sinica,2011,32(11):1489 – 1494.(in Chinese)
[18] 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 – 212.
[19] 熊平,朱天清,等.基于杂度增益与层次聚类的数据匿名方法[J].计算机研究与发展,2012,49(7):1545 – 1552.
XIONG Ping, ZHU Tian-qing, et al. A data anonymization approach based on impurity gain and hierarchical clustering[J]. Journal of Computer Research and Development, 2012, 49(7):1545 – 1552.(in Chinese)
[20] Li J, Wong R, Fu A, Pei J. Achieving *k*-anonymity by clustering in attribute hierarchical structure[A]. Proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery (DaWaK)[C]. Heidelberg, Berlin: Springer-Verlag,2006.405 – 416.
[21] 彭京,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 Jing, TANG Chang-jie, CHENG Wen-quan, et al. A hierarchy distance computing based clustering algorithm[J]. Chinese Journal of Computers, 2007, 30(5):786 – 795.(in Chinese)
[22] 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 Zhi-hui, Xu Jian, Wang Wei. Clustering-based approach for data anonymization[J]. Journal of Software,2010,21(4):680 – 693.(in Chinese)
[23] LeFevre, K, DeWitt, D, Ramakrishnan, R. Mondrian multidimensional *k*-anonymity[A]. Proceedings of the 22nd International Conference on Data Engineering[C]. Atlanta, Georgia, USA: IEEE,2006.25 – 36.
[24] Campan, A, Truta T M, Miller J, Sinca R. A: clustering approach for achieving data privacy[A]. Proceedings of the 2007 International Data Mining[C]. Las Vegas, Nevada, USA: CSREA Press,2007.321 – 327.

作者简介

杨　静　女,1962 年生于黑龙江哈尔滨.哈尔滨工程大学计算机科学与技术学院教授、博士生导师.主要研究方向为数据库与知识工程、数据挖掘、隐私保护、软件理论等.
E-mail:yangjing@hrbeu.edu.cn



王　超　男,1988 年生于河北省沧州市.哈尔滨工程大学计算机科学与技术学院博士研究生.主要研究方向为数据库与知识工程、数据挖掘、隐私保护.
E-mail:wangchao0605@hrbeu.edu.cn

张健沛　男,1956 年 11 月出生于黑龙江哈尔滨.哈尔滨工程大学计算机科学与技术学院教授、博士生导师.主要研究方向为企业智能计算、数据库与知识工程、数据挖掘、社会网络、软件理论等.
E-mail:zhangjianpei@hrbeu.edu.cn