聚类的(α,k)-匿名数据发布

杨高明,杨 静,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: 为更好的抵御背景知识攻击和同质攻击,保护特定的敏感值或全部敏感值,定义了单敏感值(α,k)-匿名模型和多敏感值(α,k)-匿名模型,并分别设计了两个聚类算法予以实现,同时分析了算法的正确性和复杂性.对于即包含连续属性又包含分类属性的数据集,给出了数据集的详细映射与处理方法,使数据集中点的距离可以方便的计算,彻底避免了把数据点距离和信息损失混淆的情况.详细的理论分析和大量的实验评估表明算法有较小的信息损失和较快的执行时间.

关键词: 数据发布; k-匿名; l-多样性; 隐私保护; 聚类
中图分类号: TP309.2 **文献标识码:** A **文章编号:** 0372-2112 (2011) 08-1941-06

Achieving (α, k)-Anonymity via Clustering in Data Publishing

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

Abstract: To better protect personal privacy against background knowledge attack and homogeneity attack, single sensitive value and multi sensitive values (α, k)-anonymity models were defined respectively. For achieving this purpose, two clustering algorithms were designed. At the same times, we made correctness and complexity analysis for the algorithms. Since the data sets contain continuous attributes and classification attributes, a detailed mapping and processing method was given, that make the distance between data points can calculate easily, and avoid completely the case that confusion data points distance and information loss. Experiment results and detailed theory analysis demonstrate that our methods are effective on both information loss and execution time comparing with existing methods.

Key words: data publishing; k-anonymity; l-diversity; privacy preserving; clustering

1 引言

随着信息技术的发展,各企事业单位在使用信息时不可避免的引起个人隐私信息泄露.为有效保护个人隐私信息,隐私保护的数据发布(PPDP)技术[1,2]应运而生.Sweeney L为达到隐私保护的目的提出 k-匿名的隐私保护模型[3].该模型可以有效的抵御连接(link)攻击,但是不能阻止背景知识攻击和同质攻击[4].为此Machanavajjhala A 提出 l-多样性模型[4],它要求每个簇内的敏感值满足 l-多样性约束,以提高敏感值与其所属个体的连接难度.王智慧等[5]提出使用聚类方法实现 l-多样性隐私保护,他们首先对数据进行聚类,然后对聚类后的簇概化处理.(α,k)-匿名模型[6,7]也是为了弥补 k-匿名的不足而提出的隐私模型,它是通过控制等价类中敏感值的出现频率来实现敏感值的多样性.其中文[6]提出简单(α,k)-匿名模型的概念,文[7]给出使用概化方法的初步完善方案.韩建民等[8]针对简单(α,k)-匿名模型,为每个敏感值设定不同的 α,提出面向敏感值的个性化(α,k)-匿名隐私保护模型,这种方法对敏感值数目较少时可以很好的提高隐私保护度,不适用于敏感数值较多的情况,另外他们的方法仅限于处理分类属性.

不同的单位和个人对隐私的要求不一样,若原始数据表 1(a)中仅需保护敏感值"Lues",我们定义了单敏感值(α,k)-匿名模型实现这种情况隐私保护;若原始数据表中全部敏感值均需要保护,我们定义了多敏感值(α,k)-匿名模型.在综合研究聚类数据类型的映射和处理基础上[9],并结合了数据集的数据类型,本文不使用概化高度作为元组之间的距离,而是使用相异度作为元组之间距离度量标准,使用聚类算法实现,避免了匿名时信息损失过大问题.

收稿日期:2010-12-28;修回日期:2011-03-16
基金项目:国家自然科学基金(No.61073043,No.61073041,No.60873037);黑龙江省自然科学基金(No.F200901)

2 基本概念和度量空间

k-匿名及其演化的各种数据发布方法把数据表属性分为三类:显式标识符属性、准标识符属性 QI(Quasi Identifier)以及敏感属性.显式标识符是唯一标识个体身份的属性,如用户身份证号码,姓名等.数据发布之前应删除或加密这些属性;准标识符是通过这些属性的连接来唯一标识个体身份的一组属性,如表 1(a)属性组{Age,Sex,Country};敏感属性指包含个体隐私信息的属性,如薪水、身体状况等.

定义 1 等价类.数据集 DT 上的等价类 EC 为数据集 DT 中部分元组的集合,它们在准标识符 QI 上具有相同的属性值.

例如表 1(b)中元组 1、2 关于{Age,Sex,Country}构成一个等价类.等价类包含元组数的多少标志着类中个体的身份保护强度.如果一个数据集 DT 的每个等价类关于 QI 包含的元组数大于或者等于 k,则这个数据集是 k-匿名的.

表 1(a) 原始数据表

Age	Sex	Country	Disease
25	M	USA	Lues
30	F	Haiti	Heart disease
28	M	USA	Lues
31	F	Haiti	Cancer

表 1(b) 匿名表

Age	Sex	Country	Disease
[24 – 28]	M	USA	Lues
[24 – 28]	M	USA	Lues
[30 – 33]	F	Haiti	Heart disease
[30 – 33]	F	Haiti	Cancer

表 1(c) (α,k)-匿名表

Age	Sex	Country	Disease
[32 – 49]	Any	America	Lues
[32 – 49]	Any	America	Heart disease
[32 – 49]	Any	America	Lues
[32 – 49]	Any	America	Cancer

2.1 单敏感值(α,k)-匿名模型

定义 2 单敏感值(α,k)-匿名.给定数据表 $DT=\{A_1,A_2,\cdots,A_m,S\}$,其中准标识符 $QI=\{A_1,A_2,\cdots,A_m\}$,敏感属性为 S.设存在映射 $F(DT)\rightarrow DT'$,使得 DT' 满足 k-匿名.对指定的 $s\in S$,设(EC,s)为等价类 EC 中包含敏感值 s 的元组集合,$\alpha(0<\alpha<1)$为用户指定的阈值.如果 s 在每个等价类中的频率都不大于 α,即 $\forall EC$,都有 $|(EC,s)|/|EC|\leqslant\alpha$,则匿名数据表 DT' 关于准标识符 QI 和敏感值 s 满足单敏感值(α,k)-匿名.

单敏感值(α,k)-匿名约束面向一个指定的敏感值,例如表 1(b)是关于准标识符{Age,Sex,Country}和敏感值 Cancer 是单敏感值(0.5,2)-匿名的,它限制了敏感值 Cancer 的频率,但没有限制其他敏感值(如 Lues)的频率,攻击者可能会以较高的概率推导出患有其他疾病个体的敏感信息,因此该模型不安全.

2.2 多敏感值(α,k)-匿名模型

定义 3 多敏感值(α,k)-匿名.给定数据表 $DT=\{A_1,A_2,\cdots,A_m,S\}$,其中准标识符 $QI=\{A_1,A_2,\cdots,A_m\}$,敏感属性为 S.设存在映射 $F(DT)\rightarrow DT'$,使得 DT' 满足 k-匿名.对 $\forall s\rightarrow S$,设(EC,s)为等价类 EC 中包含敏感值 s 的元组集合,$\alpha(0<\alpha<1)$为用户指定的阈值.如果 s 在每个等价类中的频率都不大于 α,即 $\forall s\in S$,$\forall EC,|(EC,s)|/|EC|\leqslant\alpha$,则匿名数据表 DT' 关于准标识符 QI 和敏感属性 S 满足多敏感值(α,k)-匿名.

多敏感值(α,k)-匿名模型将单敏感值(α,k)-匿名约束扩展到敏感属性的所有值,为所有的敏感值设置统一的频率约束,因此数据集中每个属性的每个敏感值都得到保护.表 1(c)是多敏感值(0.5,2)-匿名表.

2.3 距离度量

本文讨论的聚类算法仅仅考虑数值和分类这两种数据类型,其他数据类型可以映射到这两种数据类型之一.与这两种数据类型相关的域分别称为数值域和分类域[9,10].假设数据库对象中每个属性集合 A_i 描述一个域值,以 $DOM(A_i)$表示并与一个预定义的语义或者数据类型相联系.将不同类型的变量组合在单个相异度矩阵(Dissimilarity Matrix)中,把变量转换到共同标度的区间[0.0,1.0].在不引起混淆的情况下以向量 $\boldsymbol{T}$ 表示数据集 DT 中的元组对象集合.

$\boldsymbol{T}=[x_1^r,x_2^r,\cdots,x_p^r,x_{p+1}^c,\cdots,x_m^c]$,此处前面 p 个元素是数值属性值,剩余的是分类属性值.如果 $\boldsymbol{T}$ 仅仅包含一种类型值,他被简化为:$[x_1,x_2,\cdots,x_m]$.如果 $\boldsymbol{T}$ 仅仅由数值构成则称为数值对象,若由分类值构成则称为分类对象,如果既包含数值又包含分类值则称为混合类型对象.假设每个对象有 m 个属性,且不包含空值.设 $\boldsymbol{T}=\{T_1,T_2,\cdots,T_n\}$为 n 个对象的集合,对象 $\boldsymbol{T}_i$ 表示为$[x_{i,1},x_{i,2},\cdots,x_{i,m}]$.如果 $x_{i,j}=x_{k,j},1\leqslant j\leqslant m$,则 $T_i=T_k$.此处 $T_i=T_k$ 并不意味着 T_i 和 T_k 在数据集中是一个对象,仅仅意味着两个对象在属性 $A_1,A_2,\cdots,A_m$ 上有相等的值.

在计算对象的相异度时,分类属性中的序数变量的处理与数值变量(区间标度变量)非常类似.序数变量的值可以映射为秩.假设数据集包含 m 个混合类型的变量,元组对象 i 和 j 之间的相异度的 $d(i,j)$定义为:

$$d(i,j)=\frac{\sum_{f=1}^{m}\delta_{ij}^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{m}\delta_{ij}^{(f)}} \tag{1}$$

其中,如果 x_{if}或x_{jf}缺失(即对象 i 或对象j 没有变量f 的度量值),或者 $x_{if}=x_{jf}=0$,且变量 f 是非对称二元变量,则指示项 $\delta_{ij}^{(f)}=0$;否则,指示项 $\delta_{ij}^{(f)}=1$.变量 f 对 i 和j 之间相异度的贡献 $d_{ij}^{(f)}$根据它的类型计算:

• 如果 f 是区间标度变量: $d_{ij}^{(f)}=|x_{if}-x_{jf}|/(\max_h x_{hf}-\min_h x_{hf})$其中 h 遍取变量f 的所有非缺失对象

• 如果 f 是二元或者分类变量:若 $x_{if}=x_{jf}$, $d_{ij}^{(f)}=0$,否则 $d_{ij}^{(f)}=1$.

• 如果 f 是序数变量:计算秩 r_{if}和 $z_{if}=(r_{if}-1)/(M_f-1)$,并将 z_{if}作为区间标度变量对待.

上面的步骤与我们所见到的各种单一变量类型的处理相同.唯一的不同就是基于区间的变量,其中规格化使得变量值映射到区间[0.0,1.0].这样,即便描述对象的变量具有不同类型,对象之间的相异度也能够计算.

相异度矩阵存储所有成对的 n 个对象的邻近度(Proximity),通常用一个 $n\times n$ 矩阵表示:

$$\boldsymbol{DM}=\begin{bmatrix}0&&&&\\d(2,1)&0&&&\\d(3,1)&d(3,2)&0&&\\\vdots&\vdots&\vdots&&\\d(n,1)&d(n,2)&\cdots&\cdots&0\end{bmatrix} \tag{2}$$

其中 $d(i,j)$是元组对象 i 和j 之间的测量差或相异度.通常 $d(i,j)$是一个非负的数值,元组对象 i 和j 越相似或越"近",其值越接近 0;两个元组对象越不同,其值越大.由于 $d(i,j)=d(j,i)$,并且 $d(i,i)=0$,所以相异度矩阵是个对角下(上)三角矩阵.

2.4 匿名效用度量

传统概化/匿名方法对每个属性单独处理,导致大的信息损失,而聚类方法使用中心点方法从全局考虑问题,信息损失较小[11].

2.4.1 数值属性失真度

设数据表 $DT=\{A_1,A_2,\cdots,A_m,S\}$,其中准标识符 $QI=\{A_1,A_2,\cdots,A_m\}$.元组 $t=(x_1,\cdots,x_n)$概化为 $t'=([y_1,z_1],\cdots,[y_n,z_n])$, $y_i\leqslant x_i\leqslant z_i(1\leqslant i\leqslant n)$,则数值属性 A_i 的信息损失为:

$$InfoLoss_{A_i}(t_i)=z_i-y_i/|A_i| \tag{3}$$

其中$|A_i|=\max_{t\in T}t.A_i-\min_{t\in T}t.A_i$

2.4.2 分类属性失真度

分类属性的概化通常伴随着分类层次系统树,它为属性值指定不同的粒度.设元组 t 在属性 A_i 上的值为 v, v 概化为一系列值$v_1,\cdots,v_m$.它们在层次树上的公共祖先表示为 $ancestor(v_1,\cdots,v_m)$,则分类属性值的失真度为:

$$InfoLoss_{A_i}(t)=\frac{|ancestor(v_1,\cdots,v_m)|}{|A_i|} \tag{4}$$

其中$|A_i|$为分类层次树上叶子结点数.

2.4.3 元组和数据集的失真度

对于数值属性 A_i, $InfoLoss_{A_i}(t)$度量元组 t 关于属性A_i 在概化区间的概化量;对分类属性 A_i, $InfoLoss_{A_i}(t)$度量元组 t 关于属性A_i 概化值覆盖的不同数目的量.对于即包含数值属性又包含敏感属性的元组 t,其信息损失为:

$$InfoLoss(t)=\sum_{i=1}^{m}InfoLoss_{A_i}(t) \tag{5}$$

整个数据表的信息损失为:

$$InfoLossDB(T)=\sum_{t\in T}InfoLoss(t) \tag{6}$$

3 匿名模型的聚类算法

目前基于聚类实现匿名化隐私保护数据发布有两种发布方式,一种是首先聚类数据对象然后发布簇中心和簇内的元组数和半径信息[11],另一种是聚类以后对每个簇进行概化/隐匿操作,并发布概化/隐匿以后的数据[5,8],本文采用概化/隐匿方法,把簇中心和半径信息重新映射到初始度量空间以发布概化的个体记录,即先对数据集匿名处理然后对聚类后的簇概化.下面分别给出单敏感值(α,k)-匿名算法和多敏感值(α,k)-匿名算法.

3.1 单敏感值(α,k)-匿名模型

单敏感值(α,k)-匿名模型仅指定一个敏感属性,其算法基本思想是:首先计算数据集的相异矩阵,求出元组对象之间的距离,然后进行聚类处理,最后调整不满足单敏感值(α,k)-匿名的簇,使给定的敏感值在每个簇内满足单敏感值(α,k)-匿名要求.算法 1 给出了具体过程.

算法 1 对第一个质心的选择是任意的,选择了质心以后从相异矩阵找到距离质心最近的 k-1 个点(元组).步骤 3 到步骤 7 是数据聚类过程,主要把数据集聚成大小为 k 的簇.算法步骤 4 和 7 是关键,每加入一个点都要调整簇质心,以确保找到距离质心最近的点.簇的半径越小,聚类以后概化的信息损失越小.为避免离群点干扰,步骤 6 不能选择最远的点质心.步骤 8 到步骤 10 是调整簇阶段,主要目的是使每个簇内指定的敏感值其在簇内比率不大于 α.数据集的个数一般不会正好等于 k 的整倍数,对于剩余的元组则放到最后处理,即步骤 11.剩余的点若加入簇以后导致簇的半径明显变大,则需把该点做隐匿处理.

算法 1 单敏感值(α,k)-匿名聚类算法

输入:数据表 DT,匿名约束 k,指定的敏感值 s 及其频率约束α

输出:满足单敏感值(α,k)-匿名约束的匿名表 DT′

步骤:

1. $C=\emptyset$;
2. 计算数据表 DT 的相异矩阵 Matrix(DT);
3. 任选一点 t_i 作为簇 C_i 的中心点;
4. 选择距离 t_i 最近的 $k-1$ 个数据点加入 C_i,调整簇 C_i 的质心;
5. DT = DT − C_i; $C=C+C_i$;
6. 在剩余数据集 DT 中选择一点做为簇 C_j 的质心;
7. 若|DT| > k,转到步骤 4,否则放弃步骤 6;
8. 对每个包含敏感值 s 的簇,计算 s 的频度,并把簇按照 s 的频度由大到小排序;
9. 选择 s 频度最大的簇 C_k,寻找簇 C_k 内距离质心最远且使簇 C_k 不满足(α,k)-匿名约束的 $l(l<k)$个元组,把这 l 个元组加入距离它们最近的簇(簇 C_k 除外),从其他簇内选择距离簇 C_k 质心最近的 l 个元组加入簇 C_k;
10. 循环执行步骤 9,直到所有簇内是的频度小于 α;
11. 将未分配的剩余簇加入距离他们较近的簇且使加入后敏感值的频率不大于 α;

3.2 多敏感值(α,k)-匿名模型

多敏感(α,k)-匿名算法基本思想是:计算数据集的相异矩阵,在选择数据点加入簇时考虑是否满足(α,k)-匿名.具体过程见算法 2.

算法 2 多敏感值(α,k)-匿名聚类算法

输入:数据表 DT,匿名约束 k,敏感属性 S 的频率约束α

输出:满足多敏感值(α,k)-匿名约束的匿名表 DT′

步骤:

1. $C=\emptyset$,标记 DT 中的全部点为 False;
2. 计算数据表 DT 的相异矩阵 Matrix(DT);
3. 任选一点 t_i 作为簇 C_i 的质心,标记 t_i 为 True;
4. 选择标记为 False 且距离簇 C_i 质心最近的数据点,设其敏感值为 s,若该点加入簇 C_i 之后导致该簇敏感值为 s 的数据点数目大于 $\alpha\cdot k$,则放弃该点,否则把该点加入簇 C_i,并标记该点为 True,调整簇 C_i 的质心;
5. 循环执行步骤 4,直到簇 C_i 元组数达到 k 为止;
6. DT = DT − C_i; $C=C+C_i$;
7. 在剩余数据集 DT 中一点做为簇 C_j 的质心;
8. 转到步骤 4;
9. 将未分配的剩余簇加入距离他们较近的簇,且保证满足(α,k)-匿名;

算法 2 的关键是步骤 4,这一步保证每个簇内敏感值满足多敏感值(α,k)-匿名.由于簇有 k 个元素,每个敏感值的个数与 k 的比率不超过α,因此每个簇内每个敏感值的个数不多于 $\alpha\cdot k$.加入数据点时调整质心的原因是质心的位置决定了簇对应的元组属性,准确的质心保证概化信息损失最小,若代表质心的点严重偏离簇的实际质心将不可避免的带来大的信息损失,降低数据效用.而选择距离已经存在的簇最远的点作为下一个簇的质心是为了避免已经存在的簇影响将要生成的簇.第 9 步在执行时要考虑不能仅仅把剩余的这些点加入其最近的点,若某个点加入其最近的簇导致簇违背的敏感值个数大于 $\alpha\cdot k$,则只能寻找次优的簇加入.

3.3 算法的正确性与复杂性分析

3.3.1 正确性分析

算法 1 在第 1 步到第 8 步保证每个簇至少包含 k 个元组,而第 10 步和第 11 步保证每个簇对于指定的敏感值其频率不大于 α.最后处理剩余的元组时依然保证了指定的敏感值频率不大于 α,因此算法 1 是正确的.算法 2 步骤 4 到步骤 8 保证了每个簇生成时满足多敏感值(α,k)-匿名,步骤 9 加入最后剩余元组时,依然考虑每个簇满足多敏感值(α,k)-匿名,因此算法 2 也是正确的.

3.3.2 复杂性分析

对于算法 1,计算相异矩阵时间代价是 $O(n^2)$,步骤 4 到 7 是循环聚类过程,其时间开销是:$O(n+(n-k)+(n-2k)+\cdots+(n-(\lfloor n/k\rfloor-1)k))=O(n^2/k)$.步骤 8 到 10 是簇的调整过程,目的是使每个簇满足 α 约束,时间开销最坏情况下是:$O(\alpha k\lfloor n/k\rfloor)=O(\alpha n)=O(n)$.所以总的时间开销为 $O(n^2)+O(n^2/k)+O(n)=O(n^2)$.对于算法 2,计算相异矩阵时间代价是 $O(n^2)$,与算法 1 不同,算法 2 直接生成满足条件的簇.步骤 3 到步骤 8 是簇的生成过程,其时间代价为:$O(O(n)+O(n-k)+O(n-2k)\cdots+O(n-(\lfloor n/k\rfloor-1)k))=O(n^2/k)$.其总的时间复杂度为 $O(n^2)$.

4 实验数据及结果分析

本实验主要验证信息损失和算法的执行时间,并与 Wong R 提出的算法[7]相比较,此处称之为 SimAnony 算法.我们提出的单敏感值(α,k)-匿名称之为 SingAnony,多敏感值(α,k)-匿名算法称之为 Multi-Anony.

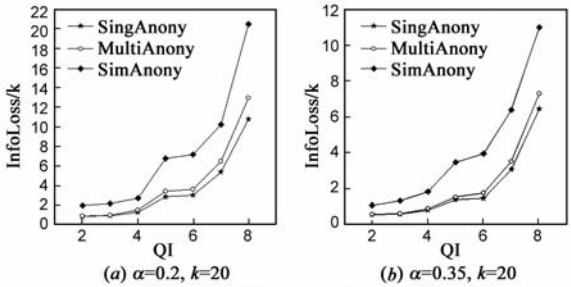
4.1 实验数据及环境

实验使用 Adult 标准数据集*,该数据集是目前数据发布实验使用的事实标准.该数据集有 48842 个记录,去除包含空值的数据集之后还有 45222 个记录.准标识符属性取 3 个数值属性,5 个分类属性,occupation 做为敏感属性.实验的硬件环境为 Intel Pentium IV 3.0GHz CPU,1024MB RAM,操作系统为 Microsoft Windows XP.编译环境是 C ++,数据库是 Microsoft SQL Server 2005.

* http://archive.ics.uci.edu/ml/datasets/Adult

4.2 信息损失分析

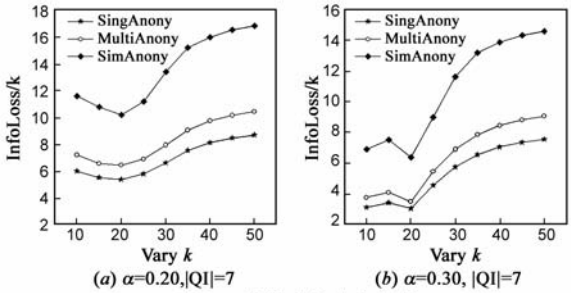
图 1 分别给出了当 $k=20$,$\alpha=0.2,0.35$ 时,准标识符维数|QI|变化时对 SimAnony、SingAnony 和 MultiAnony 信息损失大小的影响.当准标识符维数|QI|增加时,三个算法的信息损失均随之而增加.这是由于当 k 和 α 确定以后,每个等价类包含的敏感值个数就确定了.SimAnony 和 SingAnony 由于仅指定一个敏感值,且该敏感值大于其在整个数据集的分布,所以其信息损失主要是随着|QI|的增加,需要概化更多的元组属性导致信息损失增大.由于每个属性值域大小不同,所以其信息损失虽然总体呈指数增长但幅度不同.在同等条件下,SingAnony 和 MultiAnony 产生的信息损失要远小于 SimAnony.这是因为 SimAnony 使用 Apriori 剪枝概化策略,属性值选择的顺序不同产生的信息损失差别很大;而 SingAnony 和 MultiAnony 采取聚类策略,不受概念层次结构限制,能够找到具有较小信息损失的簇,并予以概化来满足(α,k)模型的匿名保护要求.



(a) α=0.2, k=20　(b) α=0.35, k=20

图1　准标识符维数|QI|变化时的信息损失

图 2 为准标识符属性个数为 7,α 值取 0.20、0.30,k 值变化时,3 种匿名模型的信息损失量的比较.由图 2 知它们的信息损失量都会随 k 值的增加而增加,因为 k 的增加要求每个等价类中的元组数增多,需要对元组进行更高层次的概化,所以会产生更多的信息损失.在 α 值增加时信息损失均有所增加.这是因为随着 α 值的增加,(α,k)-匿名模型允许数据集中每个等价群包含更多的敏感元组(至多 $\alpha\cdot k$ 个元组).因此信息损失也相应增大.另外,随着 k 值的增加,SingAnony 和 MultiAnony 的信息损失只是缓慢增长,而 SimAnony 的信息损失则呈跳跃式增长.



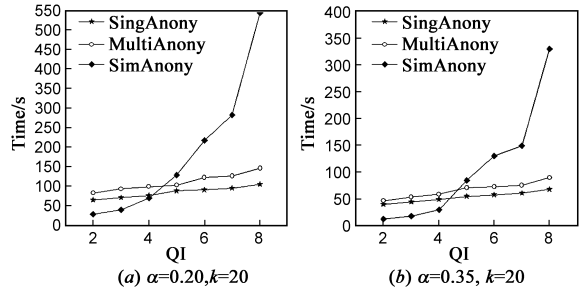
(a) α=0.20,|QI|=7　(b) α=0.30, |QI|=7

图2　k变化时的信息损失

由图 1、图 2 可以看出,在相同情况下单敏感值(α,k)-匿名模型的信息损失量最小,多敏感值(α,k)-匿名模型和简单(α,k)-匿名模型的信息损失量其次.单敏感值(α,k)-匿名和多敏感值(α,k)-匿名虽然都是指定一个敏感值,但由于单敏感值(α,k)-匿名模型采用聚类方式,把距离相近的点聚到一个簇中,使信息损失明显减少.多敏感值(α,k)-匿名模型信息损失量与单敏感值(α,k)-匿名模型差别很小,但多敏感值(α,k)-匿名模型保护能力更强.

4.3 执行时间分析

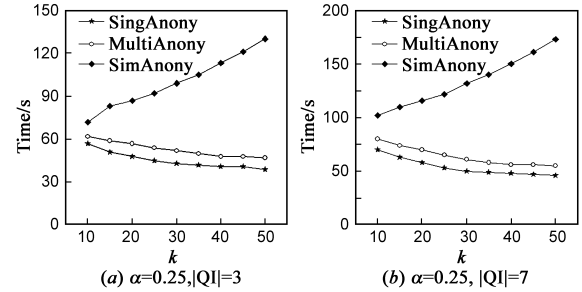
图 3 给出了当 k 和 α 值固定,准标识符维数|QI|变化时对三个算法执行时间的影响.随着|QI|的增加,它们的执行时间都有所增加.但是 SimAnony 的执行时间增长呈明显加速趋势.由于 SimAnony 通过递增地考察准标识符子属性集上的概化属性值组合来寻找可实现匿名保护的全值域概化方案,在最坏情况下,SimAnony 的执行时间随着准标识符维数增加将呈指数式增长.而 SingAnony 和 MultiAnony 通过考察元组与类之间以及类与类之间的距离寻找合适的概化方案,以较小的信息损失来满足匿名保护的需求,其执行时间随着准标识符维数的增加而呈线性增长趋势.



(a) α=0.20,k=20　(b) α=0.35, k=20

图3　准标识符维数|QI|变化时执行时间

图 4 给出了当|QI|和 α 固定,k 值变化时对三个算法执行时间的影响.随着 k 值的增加,执行时间呈线性增长趋势,这是因为 SimAnony 在 QI 的每个子属性集上采取 Apriori 的概化策略.随着 k 值的增加,它需要作更多的概化尝试,直到结果满足(α,k)-匿名模型的需求.所以 k 值增加使其执行时间有增加的趋势.而 SingAnony 和 MultiAnony 在初始化相异矩阵时花费时间



(a) α=0.25,|QI|=3　(b) α=0.25, |QI|=7

图4　k值变化下的执行时间

是 $O(n^2)$,而在后来的聚类阶段花费时间是 $O(n^2/k)$,所以它们都会随着 k 值的增大执行时间减少.在不同的准标识符下 SimAnony 增大较明显,因为准标识符越多,SimAnony 需要执行的剪枝操作越多.而对于 SingAnony 和 MultiAnony 来说,它们在聚类之前已经做了映射,准标识符的增加,并不能明显增加其执行时间.

5 结束语

为更好的保护个人隐私免遭同质攻击和背景知识攻击,我们针对敏感属性的特定敏感值保护定义了单敏感值(α,k)-匿名模型,针对敏感属性的全部敏感值保护定义了多敏感值(α,k)-匿名模型,并分别设计了两个聚类算法予以实现.由于数据包含连续属性和分类属性,给出了详细的映射与处理方法,提出以相异度作为元组之间距离度量标准,使数据集中点的距离可以方便的计算,彻底避免了把数据点距离和信息损失混淆的情况.给出了算法并分析了算法的正确性和复杂性,同时以实验验证我们算法的性能和有效性,并详细分析了实验结果.

参考文献

[1] Fung B C M, Wang K, Chen R, et al. Privacy-preserving data publishing: A survey of recent developments[J]. ACM Comput Surv, 2010, 42(4): 1 – 53.

[2] 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(10):2023 – 2029.
Han Jianmin, Cen Tengteng, Yu Huiqun. Research in microaggregation algorithms for k-Anonymization[J]. Acta Electronica Sinica, 2008, 36(10): 2023 – 2029. (in Chinese)

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

[4] Machanavajjhala A, Kifer D, Gehrke J, et al. l-diversity: Privacy beyond k-anonymity [J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 1 – 52.

[5] 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. (in Chinese)

[6] Wong R, Li J, Fu A, et al. (α,k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing[A]. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. ACM, 2006. 754 – 759.

[7] Wong R, Li J, Fu A, et al. (α,k)-Anonymous data publishing [J]. Journal of Intelligent Information Systems, 2009, 33(2): 209 – 234.

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

[9] Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values[J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283 – 304.

[10] Li C, Biswas G. Unsupervised learning with mixed numeric and nominal data[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(4): 673 – 690.

[11] Aggarwal G, Panigrahy R, Tom, et al. Achieving anonymity via clustering[J]. ACM Trans Aalgorithms, 2010, 6(3): 1 – 19.

作者简介

杨高明 男,1974 年生于安徽临泉.哈尔滨工程大学计算机学院博士研究生.研究方向为隐私保护、机器学习.
E-mail: yanggaoming@hrbeu.edu.cn

杨 静 女,1962 年生于黑龙江哈尔滨.哈尔滨工程大学计算机学院博士生导师,教授.研究方向为隐私保护、机器学习.
E-mail: yangjing@hrbeu.edu.cn