(k,l)-多样性数据发布研究

杨高明[1] 李敬兆[1] 杨 静[2] 朱广丽[1]
(安徽理工大学计算机科学与工程学院 淮南 232001)[1]
(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)[2]

摘 要 发布未经处理的数据会导致身份泄露和敏感属性泄露,通过概化准标识符可以达到隐私保护的目的,但信息损失过大。针对该问题提出一种基于聚类的(k,l)-多样性数据发布模型并设计算法予以实现。通过使用概率联合分布度量数据对象的离散属性和连续属性相似性,提高了数据的效用。详细论述了簇的合并、调整和概化策略,结合参数k和l提出隐私保护度概念,指出了基于聚类的最优化(k,l)-多样性算法是NP-难问题,并分析了算法的复杂度。理论分析和实验结果表明,该方法可以有效减少执行时间和信息损失,提高查询精度。

关键词 隐私保护,数据发布,l-多样性,数据效用,聚类,相似性度量
中图法分类号 TP309.2 **文献标识码** A

Achieving(k,l)-Diversity in Privacy Preserving Data Publishing

YANG Gao-ming[1] LI Jing-zhao[1] YANG Jing[2] ZHU Guang-li[1]
(School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China)[1]
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)[2]

Abstract In order to avoid disclosure of individual identity and sensitive attribute, reduce the information loss when data release, a clustering-based algorithm to achieve(k,l)-diversity(CBAD) in data publishing was presented. The discrete attributes and continuous attributes mixed in the data set were fully taken into account while clustering. The probability distribution was used as metrics to measure similarity between the data objects. We solved the confusion of the information loss and the distance between data objects, pointed out that the clustering-based optimization(k,l)-diversity algorithm is NP-hard problem, proposed the concept of privacy protection degree with parameter k and l, and analysed the complexity of the algorithm. Theoretical analysis and experimental results show that the method can effectively reduce the execution time and information loss, improve query precision.

Keywords Privacy preserving, Data publishing, l-Diversity, Data utility, Clustering, Similarity measures

1 前言

隐私保护的数据发布[1]在数据库领域是一个非常活跃的研究课题。为保护个人隐私,仅仅从数据中删除标识符信息(如姓名、身份证号码等)并不能阻止隐私的泄露,把发布的数据与其他数据进行联接会导致个体的身份泄露和敏感属性泄露[1]。k-匿名技术可以有效地防止连接攻击,但是不能防御同质攻击和背景知识攻击[2]。为有效地解决这个问题,l-多样性(l-diversity)[2]、(α,k)-匿名[3]、t-逼近(t-closeness)[4]等模型被陆续提出。这些隐私保护的数据发布模型主要采用概化/隐匿方法实现。概化/隐匿方法的基本思想是同一个簇的准标识符属性值用相同的概化值代替,这种处理方法基本保持了原来的语义信息,但会造成信息损失且降低数据的效用。如Machanavajjhala等提出了l-多样性原则并给出了两个实例:递归(c,l)-多样性和熵l-多样性[2],它们的算法在Incognito的基础上实现多样性原则,而Incognito使用全域概化方法[5],这导致信息损失过大。

隐私保护的数据发布要求公布的数据集中每个簇概化到相同的准标识符,这与数据挖掘中的聚类过程十分相似。于是有学者使用聚类方法实现k-匿名[6,7]和l-多样性[8]。文献[8]提出的聚类实现l-多样性方法要求每个簇包含l个元组且每个元组的敏感属性值不同,这个要求高于熵-多样性和递归(c,l)-多样性[2],Machanavajjhala已经说明了熵-多样性缺少实用性[2]。

综合研究了目前存在的k-匿名[9,10]和l-多样性数据发布文献以后,我们提出基于聚类的(k,l)-多样性模型并设计算法予以实现,主要贡献有:1)针对聚类实现k-匿名或者l-多样性方法使用信息损失度量元组之间的距离,不能反映数据集

到稿日期:2012-10-23 返修日期:2012-12-21 本文受国家自然科学基金(61073043,61170060),安徽高等学校省级自然科学基金(KJ2011Z098)资助。

杨高明(1974—),男,博士,讲师,主要研究方向为隐私保护、数据挖掘,E-mail:ygm868@163.com;李敬兆(1963—),男,博士,教授,博士生导师,主要研究方向为物联网;杨 静(1962—),女,博士,教授,博士生导师,主要研究方向为隐私保护、数据挖掘;朱广丽(1971—),女,副教授,主要研究方向为智能信息处理。

中对象之间的真实差异程度，我们提出把数据集中离散属性和连续属性的联合概率分布作为对象之间相似性的度量标准。2)针对 k-匿名不能抵御背景知识攻击和同质攻击，l-多样性原则缺少实用性，我们把这两种模型结合起来，提出(k,l)-多样性模型。该模型兼顾了 k-匿名和 l-多样性的优点。3)设计基于聚类的算法实现(k,l)-多样性模型(Clustering-Based Algorithm to achieve(k,l)-Diversity,CBAD)并使用 KL_散度、信息损失和执行时间检验了算法的有效性。4)指出了(k,l)-多样性最优化算法是 NP-难问题，对算法的复杂性做了理论分析。

2 基本概念

2.1 (k,l)-多样性

数据库表通常有 3 种属性，分别是标识符属性(Identifier)、准标识符属性(Quasi-Identifier,QI)、敏感属性。为防止隐私信息泄露，标识符属性在数据发布时可以直接删除，而敏感属性是需要保护的属性，隐私保护的数据发布主要是处理准标识符属性，它是数据属性中与外部数据连接可以唯一识别个体的非标识符属性的最小集合。无论采用概化/隐匿方法还是聚类方法，都是对准标识符进行处理。数据库表在属性集上包含相同值的一组元组集合，称为一个簇。例如表 1(a)中属性集合{Zip,Gender,Age}是准标识符。表 1(b)中记录 1 和记录 2 关于准标识符{Zip,Gender,Age}组成一个簇，它们具有相同的属性值。

表 1

(a) 原始数据表

Zip	Gender	Age	Disease
43520	Male	22	Cancer
43522	Male	25	Flu
43518	Male	23	Cancer
43533	Female	21	Obesity
43567	Female	30	Crazy
43562	Female	27	Flu

(b) 2-匿名化表

Zip	Gender	Age	Disease
4352 *	Male	[21,25]	Cancer
4352 *	Male	[21,25]	Flu
4353 *	Person	[21,25]	Cancer
4353 *	Person	[21,25]	Obesity
4356 *	Female	[26,30]	Crazy
4356 *	Female	[26,30]	Flu

定义 1(k-匿名)[11] 设 $PT(tid,A_{C1},A_{C2},\cdots,A_{Gn},A_{N1},A_{N2},\cdots,A_{Nn},s)$ 是数据库表，属性集合 $\{A_{C1},A_{C2},\cdots,A_{Gn},A_{N1},A_{N2},\cdots,A_{Nn}\}$ 为准标识符属性，s 为敏感属性，将 PT 划分为簇 $\{C_1,C_2,\cdots,C_p\}$，$|C_i|\geqslant k$，$1\leqslant i\leqslant p$，C_i 中的元组在各个 A_j 上有相同的值，则数据表 PT 满足 k-匿名。

定义 2(l-多样性) 设 $PT=\sum_{i=1}^{p}C_i$ 是一个给定的数据集，$|C_i|\geqslant l$，$|C_j|\geqslant l$，$C_i\cap C_j=\phi$，$1\leqslant i\neq j\leqslant p$，$s$ 是簇 C_i 中最频繁的敏感值，$Num_{C_i}(s)$ 是簇 C_i 中敏感值等于 s 的元组 t 的个数。若 $\forall C_i$，均有 $Num_{C_i}(s)\geqslant l$，则数据集 PT 满足 l-多样性原则。

k-匿名要求每个簇至少包含 k 个对象，使识别敏感信息变得困难，入侵者即使知道个体在匿名化表中，也不能以小于 $1/k$ 的概率推导出确定个体的隐私信息，这种隐私保护数据发布方法对簇中相异敏感属性个数没有要求，而 l-多样性要求每个簇至少有 l 个不同的敏感值，对簇中对象个数没有要求。我们结合两者的优点提出(k,l)-多样性模型，它既有 k-匿名抵御连接攻击的特性，又有 l-多样性抵御背景知识攻击和同质攻击的特性。

定义 3(k,l)-多样性) 设 $PT=\sum_{i=1}^{p}C_i$ 是一个给定的数据集，$|C_i|\geqslant k$，$|C_j|\geqslant k$，$C_i\cap C_j=\phi$，$1\leqslant i\neq j\leqslant p$，$s$ 是簇 C_i 中最频繁的敏感值，$Num_{C_i}(s)$ 是簇 C_i 中敏感值等于 s 的元组 t 的个数。若 $\forall C_i$，均有 $Num_{C_i}(s)/|C_i|\leqslant 1/l$，则数据集 PT 满足(k,l)-多样性。

例如表 1(a)是医院原始数据表，表 1(b)是运算以后得到的 2-多样性表，也是(2,2)-多样性表。

2.2 信息损失衡量标准

(k,l)-多样性模型先对数据集聚类，然后对每个簇进行概化处理，最后发布概化后的处理结果。数据概化会导致信息损失。目前信息损失的定义很多，比如分类度量标准(Classification Metric,CM)，差异度量标准(Discernibility Metric,DM)等。综合考察了已有的信息损失以后，结合我们的算法特点，本文采用全局确定性损失度量标准(Global Certainty Penalty,GCP)[12]。信息损失定义如下：

定义 4(信息损失) 设 $QI=\{A_{C1},A_{C2},\cdots,A_{Gn},A_{N1},A_{N2},\cdots,A_{Nn}\}$ 是数据表 PT 的准标识符属性，C 是其中的一个簇，$\{A_{C1},A_{C2},\cdots,A_{Gn}\}$ 是离散属性，$\{A_{N1},A_{N2},\cdots,A_{Nn}\}$ 是连续属性。则有：

$$ILD_i=\frac{card(A_G)}{|A_G|} \tag{1}$$

$$ILN_j=\frac{|v_{\max}-v_{\min}|}{|A_{Nj}|} \tag{2}$$

$$IL(C)=|C|(\sum_{i=1}^{m}ILD_i+\sum_{j=1}^{n}ILN_j) \tag{3}$$

式中，ILD_i 是簇 C 离散属性 A_G 的信息损失，ILN_j 是簇 C 连续属性 A_{Nj} 的信息损失，card(A_G)是属性 C_i 在分类系统树(Taxonomy tree)所包含的子集数，$|A_G|$ 是簇 C_i 分类系统树的叶子数；$v_{\max}$ 和 $v_{\min}$ 分别是 A_{Ni} 的最大值和最小值，$|A_{Ni}|$ 是 N_i 域区间，$IL(C)$ 是簇 C 的信息损失。于是发布的数据所有簇的全局确定性损失为：

$$GCP(PT)=\frac{\sum_{C\in PT}IL(C)}{(m+n)N} \tag{4}$$

式中，PT 是簇的集合，N 是全部数据集的个数，$m+n$ 是准标识符的个数。

3 (k,l)-多样性聚类算法及相关定义

如何针对具体的数据集进行聚类模型的设计是实现(k,l)-多样性中最关键的步骤。设计不同的聚类模型就可以得到不同的聚类算法，聚类模型的设计要以描述聚类问题的具体数据类型为基础。而聚类算法则通常又依据聚类模型来建立，其设计一般包括两个方面：一是样本与类间相似度(距离)的定义，二是对聚类具体策略和步骤的设计。通常用一个距离函数来衡量样本之间、样本与类之间的近似程度。距离函数不同，聚类算法也不同，得到的聚类效果也有好有坏，因此需要慎重地选择距离函数。聚类算法具体步骤的设计在整个聚类分析的过程中占有重要的地位，这一步骤是整个聚类分析中的核心过程，也是聚类成败的关键。为使发布的数据

集效用更大，我们采用凝聚的聚类实现(k,l)-多样性。

3.1 对象邻近度

常用的距离有曼哈顿距离、欧几里德距离、契比雪夫距离等，它们主要处理连续属性数据。而随着信息技术的发展，很多数据集包含连续属性数据和离散属性数据。直接把连续属性转化成离散属性或者直接把离散属性转化为连续属性，然后计算它们之间的距离都会导致另外一种属性意义的丧失。为此本文采用文献[13]的方法，使用联合概率分布来度量数据对象之间距相似性(距离)。

一组对象(i,j)比另一组对象(l,m)更加相似，当且仅当i和j的匹配程度大于l和m的匹配程度。换句话说，就是对象之间的相似性以它们特征值的匹配程度度量。设有两组对象(i,j)和(l,m)，对于属性A_k，$(V_i)_k=(V_j)_k$，$(V_l)_k=(V_m)_k$，但是$(V_i)_k\neq(V_l)_k$。$(V_i)_k$在数据集中出现的频率等于或者大于$(V_l)_k$在数据集中出现的频率，即$((p_i)_k=(p_j)_k)\geqslant((p_l)_k=(p_m)_k)$。$(p_i)_k$，$(p_j)_k$，$(p_l)_k$，$(p_m)_k$是数据集中对象在属性值$A_k$上出现的概率。属性$A_k$对组对$(l,m)$的贡献大于对组对$(i,j)$的贡献，这些可以概括为：

$$\left.\begin{array}{l}((V_i)_k=(V_j)_k)\wedge((V_l)_k=(V_m)_k)\\((p_i)_k=(p_j)_k)\geqslant((p_l)_k=(p_m)_k)\end{array}\right\}\Rightarrow(S_{ij})_k\leqslant(S_{lm})_k \tag{5}$$

对连续特征值，相似性度量需要考虑特征值的差异量和一对属性值之间不同值的个数。$((V_i)_k,(V_l)_k)$之间值的差异量越小，随机选择一对值落在$(V_i)_k$和$(V_l)_k$为端点的区间内的几率越小，因此这组对象的相似性越大。比如两组对象(i,j)和(l,m)，对连续属性值A_k有：

$$|(V_i)_k-(V_j)_k|>|(V_l)_k-(V_m)_k|\Rightarrow(S_{ij})_k<(S_{lm})_k \tag{6}$$

当$|(V_i)_k-(V_j)_k|=|(V_l)_k-(V_m)_k|$时，相似性值由端点值所决定的区域段内不同属性值的个数所决定，它由除端点之外所有值的频率和相加得到。例如，成组对象(i,j)相似性值由$\sum_{t=(V_i)_k}^{(V_j)_k}p_t$计算，$(l,m)$相似性值由$\sum_{t=(V_l)_k}^{(V_m)_k}p_t$计算。若两个区域段长度相等端点不同，则包含较少累积频率的区域段对相似性度量的贡献度较大。

$$\left.\begin{array}{l}|(V_i)_k-(V_j)_k|=|(V_l)_k-(V_m)_k|\\\sum_{t=(V_i)_k}^{(V_j)_k}p_t\geqslant\sum_{t=(V_l)_k}^{(V_m)_k}p_t\end{array}\right\}\Rightarrow(S_{ij})_k\leqslant(S_{lm})_k \tag{7}$$

3.1.1 离散属性相似性

对离散属性A_k，其相似性量(the similarity index)使用下式计算。

$$\begin{array}{l}(V_i)_k\neq(V_j)_k\Rightarrow(S_{ij})_k=0\\(V_i)_k=(V_j)_k\Rightarrow(S_{ij})_k=1\end{array} \tag{8}$$

正如前面所讨论的，当$(V_i)_k=(V_j)_k$时，相似性值是数据集内特征值的函数。对两个具有相同特征值A_k的对象i和j，我们首先定义它的更相似性特征值集合(more similar feature value set，$MSFVS(V_i)_k$)，这是所有特征值A_k的成对值集合，这些集合里面的值等于或者更加相似于值对$((V_i)_k,(V_j)_k)$。这些值对按照式(5)所定义的关系选择。需要注意的是，若值对有更小的发生频率，则相似性更大。随机选择一个值对$((V_i)_k,(V_j)_k)\in MSFVS(V_i)_k$的概率为：

$$(p_l)_k^2=\frac{(f_l)_k\cdot((f_l)_k-1)}{n\cdot(n-1)} \tag{9}$$

此处$(f_l)_k$是值$(V_i)_k$在数据集中发生的频率，n是数据集中的对象数。所有这种成对的概率和构成成组对象的差异值$(D_{\bar{u}})_k$，因此目标对$((V_i)_k,(V_j)_k)$的相似性计算等式为：

$$(S_{\bar{u}})_k=1-(D_{\bar{u}})_k=1-\sum_{l\in MSFVS(V_i)_k}(p_l)_k^2 \tag{10}$$

3.1.2 连续属性相似性

连续属性数据相似性的计算类似于离散属性数据。若一组对象(i,j)的第k个特征值是$((V_i)_k,(V_j)_k)$，首先确定其更相似性特征区域集合(more similar feature segment set，$MSFSS((V_i)_k,(V_j)_k)$)，这个集合包含成对的满足式(6)和式(7)的值。从数据集中随机选择第k个特征值为$(V_i)_k$和$(V_j)_k$的两个对象的概率如下式所示，其中$((V_i)_k,(V_j)_k)\in MSFSS((V_i)_k,(V_j)_k)$。

$$\begin{cases}(p_l)_k(p_m)_k=\dfrac{2(f_l)_k\cdot(f_m)_k}{n(n-1)}, & (p_l)_k=(p_m)_k\\(p_l)_k(p_m)_k=\dfrac{(f_l)_k\cdot((f_l)_k-1)}{n(n-1)}, & (p_l)_k\neq(p_m)_k\end{cases} \tag{11}$$

此处f_l和f_m分别是$(V_i)_k$和$(V_j)_k$的特征值出现的频率，n是数据集中对象的总数目。把$MSFSS((V_i)_k,(V_j)_k)$中所有目标对关于特征值k的概率值相加就得到特征值k的差异值$(D_{\bar{u}})_k$。于是$((V_i)_k,(V_j)_k)$的相似性为：

$$(S_{ij})_k=1-(D_{ij})_k=1-\sum_{l,m\in MSFSS((V_i)_k,(V_j)_k)}(p_l)_k(p_m)_k \tag{12}$$

对数连续属性来说，相似性的计算是距离和密度的函数。仅仅坐标轴改变而属性值的相对距离不变，则相似性值不变。

3.1.3 混合属性数据的相似性

混合属性数据包含连续属性和离散属性，对于连续属性数据，我们使用费舍尔χ^2变换计算它的相似性值。

$$(\chi_c)_{ij}^2=-2\sum_{k=1}^{n}\ln((D_{ij})_k)$$

此处n是数据集中连续属性的个数。拥有n个自由度的χ_c分布也满足χ^2分布。对于离散属性的相似性值的计算使用Lancaster均值χ^2变换。

$$(\chi_d)_{ij}^2=2\sum_{k=1}^{m}\left(1-\frac{(D_{ij})_k\ln(D_{ij})_k-(D_{ij})_k'\ln(D_{ij})_k'}{(D_{ij})_k-(D_{ij})_k'}\right) \tag{13}$$

此处m是数据集的离散属性个数，$(D_{ij})_k$是离散属性对$((V_i)_k,(V_j)_k)$的差，$(D_{ij})_k'$是离散属性集中下一个更小的差，χ_d^2是m自由度的χ^2分布。

两个χ^2分布的和仍然是χ^2分布，其自由度等于原来两个自由度的和，也就是说两个不同属性的χ^2分布和仍然是χ^2分布，其自由度为$(m+n)$。这种χ^2分布的显著值可以通过查询标准表得到，或者使用下面的近似公式得到。

$$S_{ij}=1-D_{ij}=1-\exp\left(-\frac{\chi_{ij}^2}{2}\right)\sum_{k=0}^{t_d+t_c+1}\frac{\left(\frac{1}{2}\chi_{ij}^2\right)^k}{k!} \tag{14}$$

此处$\chi_{ij}^2=(\chi_c)_{ij}^2+(\chi_d)_{ij}^2$。

3.2 算法思想

基于聚类实现(k,l)-多样性的基本思想是寻找任意小于k的簇并与相似性最大的簇合并以组成更大的簇，重复迭代这个过程直到每个簇至少包含k($k\leqslant n$)个数据点且至少l个数据对象的敏感值各不相同，同时保证所有簇的总信息损失最小。寻找给定数据集PT的(k,l)-多样性最优解可以转化为下列动态规划问题。设数据集包含的元组数为n，簇的大

小为 k，簇中包含的不同敏感值个数为 l，则动态规划问题可以描述为：

$$\bigcup_{i=1}^{p} C_i = PT$$

$$\forall i \neq j, C_i \cap C_j = \Phi, i, j = 1, 2, \cdots, p$$

$$\forall C_i \in PT, k \leqslant |C_i|, l \leqslant |C_i.S|$$

$$GCP(PT) \text{ is minimized}$$

性质 1 最优化的(k,l)-多样性是 NP-难问题。

证明：(k,l)-多样性问题在 $k=l$ 时，如果不考虑敏感值的多样性问题就简化为 k-匿名问题，Meyerson 等人[14]已经证明了最优化 k-匿名是 NP-难的；如果考虑敏感值的多样性，则其变成 l-多样性问题，Xiao 等已经证明最优化 l-多样性是 NP-难的[15]。$k>l$ 时，最优化(k,l)-多样性约束小于最优化 l-多样性，因此最优化(k,l)-多样性是 NP-难的。

基于聚类的算法合并相似度最高的簇对象，因此信息损失较小。算法 1(CBAD 算法)给出了具体过程，在随后我们将主要讨论簇的合并、调整和概化的具体过程。

算法 1

输入：数据集 PT，参数 l，k，准标识符 QI

输出：满足(k，l)-多样性的数据集 PT′

1. 使用式(14)计算数据集 PT 中对象之间的相似性，并构造相似性矩阵 M；
2. 初始状态每个数据对象为一个簇，$PT' = \{C_1, C_2, \cdots, C_n\}$；
3. 由相似性矩阵中选择相似性最大的两个簇 C_i，C_j；
4. 若 $|C_i| + |C_j| < 2k$，合并簇$\{C_i\}$，$\{C_j\}$为簇$\{C_p\}$，$PT' = PT' - \{C_i\} - \{C_j\} + \{C_p\}$，否则调整簇$\{C_i\}$，$\{C_j\}$，使之满足 $k \leqslant |C_i| < 2k$，$k \leqslant |C_j| < 2k$；
5. 循环执行步骤 3－4，直到所有簇不小于 k；
6. 检查每个满足 k-匿名的簇，调整簇中对象，使之满足(k,l)-多样性；
7. 概化每个簇；
8. 输出 PT′。

3.2.1 簇的合并过程

算法第 4 步簇的合并有两种选择，一是合并成一个较大的簇，二是生成两个簇。若 $|C_i| + |C_j| < 2k$，则需要把两个较小的簇合并为一个较大的簇，在合并簇时删除相似性矩阵 M 中簇 C_i、C_j 所在行和列，添加簇 C_p 与 M 中其他对象的相似性值到矩阵 M 中。簇 C_p 与其他对象的相似性值计算如下：

$$S[C_p, C_k] = \frac{n_{C_i}}{n_{C_i} + n_{C_j}} S[C_i, C_k] + \frac{n_{C_j}}{n_{C_i} + n_{C_j}} S[C_j, C_k]$$

式中，n_{C_i}，n_{C_j} 分别是簇 C_i，C_j 所包含的对象数；若簇 $|C_i| + |C_j| \geqslant 2k$，则合并以后的簇 C_p 必须分裂为两个簇，假设 $|C_i| < k$，则需要从 C_j 调整 $m = k - |C_i|$ 个对象到 C_i 中，这时簇 C_i，C_j 均满足要求。这时相似性值采用下式计算：

$$S[C_i, C_k] = \frac{n_{C_i}}{n_{C_i} + n_{C_m}} S[C_i, C_k] + \frac{n_{C_m}}{n_{C_i} + n_{C_m}} S[C_m, C_k]$$

$$S[C_j, C_k] = \frac{n_{C_j} - n_{C_m}}{n_{C_j}} S[C_j, C_k] - \frac{n_{C_m}}{n_{C_j}} S[C_m, C_k]$$

3.2.2 簇的调整

对于给定的参数 k，l，设敏感保护度 $\lambda = \frac{l}{k}$。当 $l=1$ 时，敏感保护度 $\lambda = \frac{1}{k}$，这时候易受同质攻击。若 $l=k$，敏感保护度 $\lambda=1$，这时保护能力最强，可以抵御连接攻击、同质攻击和背景知识攻击。k 值一定情况下，l 值越大隐私保护度越高，信息损失也越大；l 值越小隐私保护度越低，信息损失越小，极端的情况是数据失去效用。因此 k，l 值的选择要适当。算法 1 的第 6 步调整簇，使之满足(k,l)-多样性，具体过程如下。

设初始生成的簇按照其敏感度非降序排列为 $PT = \{C_1, C_2, \cdots, C_p, C_{p+1}, C_{p+2}, \cdots, C_q\}$，其中$\{C_{p+1}, C_{p+2}, \cdots, C_q\}$为满足$(k,l)$-多样性的簇，对于不满足$(k,l)$-多样性的簇$\{C_1, C_2, \cdots, C_p\}$必须调整。设簇 C_i 的隐私保护度为 λ_i，$p+1 \leqslant i \leqslant q$ 时有 $\lambda_i > \lambda$，$0 \leqslant j \leqslant p$ 时有 $\lambda_j < \lambda$。$\lambda_j < \lambda$ 的簇$\{C_1, C_2, \cdots, C_p\}$必须调整，使之满足 $\lambda_j \geqslant \lambda$。由 $\lambda_j = l_j / k_j$ 知增大 λ_j 可以通过增加 l_j 或者减少 k_j 得到，具体情况可以分为 3 种。

- 若 $k_j = k$，$l_j < l$，则从其他簇中选择与簇 C_j 相似性最大且与簇 C_j 的对象敏感值各不相同的 $l - l_j$ 个对象加入簇 C_j，然后调整计算矩阵 M。
- 若 $k < k_j < 2k$，$l_j < l$，则从其他簇中选择与簇 C_j 相似性最大的 $l - l_j$ 个对象且这 $l - l_j$ 个对象的敏感值各不相同，把它们加入簇 C_j，并选择簇 C_j 中 $k_j + l - l_j - k$ 个敏感属性值较多的对象加入与它们相似性最大的簇。
- 若 $k < k_j < 2k$，$l_j < l$，则选择簇 C_j 中敏感属性值较多的对象放入与它们相似性最大簇，直到 $\lambda_j = \lambda$ 为止。

上面的调整过程会导致已经满足 $\lambda_j \geqslant \lambda$ 的簇不再满足这个条件，这时需要进行下一轮调整，直到全部簇满足(k,l)-多样性。

3.3 算法复杂度分析

算法 CBAD 的复杂度分为两部分，分别是计算数据集的相似性矩阵复杂度和生成满足(k,l)-多样性数据表复杂度。生成相似性矩阵的时间复杂度为：$(n-1)+(n-2)+\cdots+2 = (n^2-1)/2$，所以其复杂度为 $O(n^2)$。簇的生成阶段采用自下而上的合并聚类，其时间复杂度是 $O(n^2)$。簇调整阶段生成的簇数 m 小于原始数据集中对象数 n，调整的时间复杂度为 $O(m^2)$。由上可以看出，基于密度的(k,l)-多样性算法时间复杂度为 $O(n^2)$ $(k<n)$。

4 算法性能分析与实验结果

本实验采用 UCI 机器学习数据库中的 Adult 数据集[1)]，该数据集是隐私保护的数据发布领域标准测试数据集，共有 45222 个记录。本文选择{age, education-num, hours-per-week, capital-gain, race, marital-status, native- country, work class}作为准标识符属性，其中属性{age, education-num, hours-per-week, capital-gain}为连续属性，其余的属性为离散属性，{Occupation}为敏感属性。实验采用的硬件环境为 Intel Pentium IV 3.0GHz CPU、1.5GB RAM，操作系统是 Microsoft Windows XP Professional。编程环境为 Visual C++ 6.0。为论述方便，把文献[16]中的比较对象称为 Hilb，该方法使用 Hilbert 空间映射的方法实现数据的划分。

4.1 查询精度分析

我们首先使用 KL_散度(Kullback-Leibler Divergnece, KLD)度量查询精度，以评价数据效用。此处使用文献[16]的查询精度作为对比，由于文献[16]中的 l-多样性模型并不考虑每个簇的大小，为公平起见，此处设置 $k=l$。KLD 的值越

1) http://www.ipums.org/

小，匿名数据集与原始数据集近似度越大，我们主要以 l 变化和 QI 值变化比较查询精度。聚集查询如下：

SELECT QT_1,QT_2,…,QT_i,COUNT(*)

FROM medical data

WHERE Disease='cancer' AND Age≥40 AND Zip Code In [3000−4000]

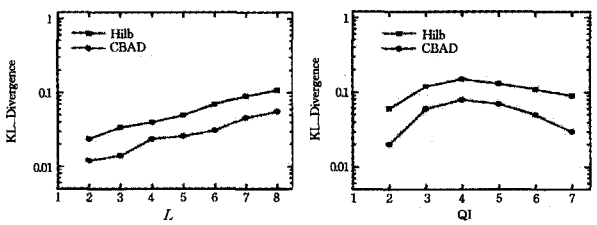
设 P 表示原始数据集，Q 表示聚类并概化以后的数据集，P_c 和 Q_c 分别表示在 P 和 Q 中的查询元组数。则估计误差定义为：

$$KLD(P,Q)=\sum_{\forall c\in IC} P_C \log \frac{P_C}{Q_C}$$

在最好的情况下，$P_c=Q_c$，$KLD=0$。

图 1(a)说明了 l 值变化时查询精度的变化情况。由图可以看出 CBAD 明显好于 Hilb，两个算法都随着 l 值的增加而误差率增大。CBAD 大约好于 Hilb 一个数量级。主要原因是随着 l 的增加，簇的大小也相应地增加，查询时会产生较大的误差。CBAD 采用聚类的方法，而 Hilb 采用的是类似聚类的技术，因此散度都较小。Hilb 算法采用的 Hilbert 空间是欧几里德空间的推广，把离散属性映射到 Hilbert 空间时改变了离散属性的物理意义，所以散度较大，而我们采用概率分布度量数据对象之间的相似性，保持了数据属性之间的联系，因此方法较优。

图 1(b)说明了改变准标识符的维数检测查询精度。低维会产生更紧凑的 l-多样性簇，因此查询精度较高。然而由于每个簇的大小是个常量，低的准标识符维度也会导致小的查询范围，这会降到查询精度。就准标识符属性域来说，这两个情况都很重要。这也是散度先升后降的原因。



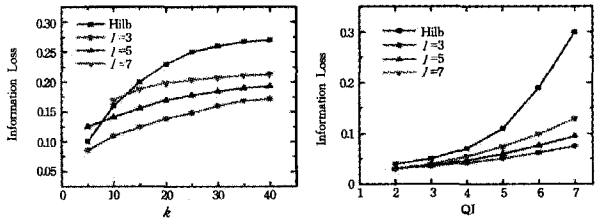
(a) L 值变化时的查询精度　　(b) QI 变化时的查询精度

图 1　查询精度分析

4.2　信息损失分析

图 2 给出了 k 值和 QI 值变化时 Hilb 和 CBAD 的信息损失情况，信息损失的计算采用式(4)。Hilb 算法把准标识符映射到 Hilbert 空间，而我们是直接计算每个对象的概率密度值。文献[16]给出了 k-匿名算法和 l-多样性算法，此处我们比较信息损失较小的 k-匿名算法。由图 2(a)可以看出在相同的 k 值情况下，k 值较小时 Hilb 信息损失较小，随着 k 值的增加 CBAD 算法的优越性越明显。由隐私保护度 $\lambda=\frac{l}{k}$ 可知，相同的 k 值，l 值越大隐私保护能力越强。

图 2(b)给出了 QI 增加时信息损失变化情况，在 QI 较小时，l 的不同取值与 Hilb 相比差别不大，随着 QI 值的增加 Hilb 增加明显，而 CBAD 变化不大。Hilb 增加明显主要是随着 QI 的增加，需要把属性不同的 QI 值映射到一个空间，对于离散属性，映射到 Hilbert 空间就失去了其物理意义，因此信息损失变大，而 CBAD 算法以概率密度度量相似性，QI 值的增大对于数据映射几乎没有影响，信息损失的增加是由于 QI 的增加需要对更多的属性进行概化，而这必然会导致信息损失增加。图 2 的比较结果说明我们的算法性能优越。



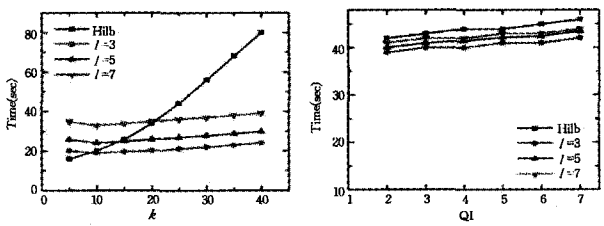
(a) k 值变化时的信息损失　　(b) QI 变化时的信息损失

图 2　K 和 QI 变化时的信息损失

4.3　执行时间分析

图 3 给出了 k 和 QI 变化时执行时间的比较。图 3 与图 2一样采用 k 值或者 QI 值变化，不同的 l 值与 Hilb 相比较。图 3(a)给出了 k 值变化时算法的执行时间比较。在 k 值较小的情况下，CBAD 算法的执行时间明显大于 Hilb 算法，这时由于 CBAD 算法除了需要计算每个属性的概率外，在生成 (k,l)-多样性簇时需要调整元组，而 Hilb 直接生成大小为 k 的簇即可。

图 3(b)给出了 QI 变化时的执行时间情况。在准标识符增加时它们的执行时间变化不大，Hilb 与 CBAD 都是采用映射的方法，时间的增加主要是计算映射产生的，在数据集映射完成以后，聚类与簇的生成阶段执行时间变化较小。相对于全局概化方法实现 l-多样性在 QI 为 7 时将近 5 分钟[2]，我们的算法提高了将近 7 倍。



(a) k 值变化时的算法执行时间　　(b) QI 变化时算法执行时间

图 3　k 和 QI 变化时的算法的执行时间

结束语　本文结合 k-匿名阻止连接攻击、l-多样性阻止同质攻击和背景知识攻击的特性，提出 (k,l)-多样性模型并采用聚类方法实现，详细讨论了簇的合并、调整和概化过程。在综合前人研究的基础上，引入统计学中概率统计方法度量数据对象相似性，避免了连续或者离散数据转换到另一种数据类型而造成数据物理意义丢失。文中指出最优 (k,l)-多样性本质是寻找动态规划模型的最优解，证明了 (k,l)-多样性模型是 NP-难问题。虽然理论分析和实验结果都说明我们方法的优越性，但是还有不完美的地方。将来的工作主要研究如何解决存在的问题，在理论上和应用上解决增量的数据发布和数据流发布。

参 考 文 献

[1]　杨高明；杨静，张健沛. 隐私保护的数据发布研究[J]. 计算机科学，2011，38(9)：11-17

[2]　Machanavajjhala A，Gehrke J，Kifer D，et al. l-Diversity：Privacy beyond k-anonymity[C]//22nd International Conference on Data Engineering：Institute of Electrical and Electronics Engineers Computer Society. Atlanta，G A，United states，2006：24

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

[4] Ninghui L, Tiancheng L, Venkatasubramanian S. t-Closeness: Privacy beyond k-anonymity and l-diversity[C]// Proceedings of the 23rd International Conference on Data Engineering. Inst. of Elec. and Elec. Eng. Computer Society, Istanbul, Turkey, 2007: 106-115

[5] Lefevre K, Dewitt D J, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity [C] // ACM SIGMOD International Conference on Management of Data. United states. Association for Computing Machinery, Baltimore, Maryland, 2005: 49-60

[6] Kabir M E, Wang H, Bertino E. Efficient systematic clustering method for k-anonymization [J]. Acta Informatica, 2011, 48, (1): 51-66

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

[8] 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(04):680-693

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

[10] Ni W, Chong Z. Clustering-oriented privacy-preserving data publishing[J]. Knowledge-Based Systems, 2012, 35: 264-270

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

[12] Xu J, Wang W, Pei J, et al. Utility-based anonymization using local recoding [C] // Philadelphia, PA, USA. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. USA: ACM, 2006: 785-790

[13] 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

[14] Meyerson A, Williams R. On the complexity of optimal k-anonymity [C]// Proceedings of the twenty-third ACM SIGMOD-SIGACT- SIGART symposium on Principles of database systems ACM. 2004: 223-228

[15] Xiao X, Yi K, Tao Y. The hardness and approximation algorithms for L-diversity[C] // 13th International Conference on Extending Database Technology: Advances in Database Technology. Association for Computing Machinery, Lausanne, Switzerland, 2010: 135-146

[16] Ghinita G, Karras P, Kalnis P, et al. A framework for efficient data anonymization under privacy and accuracy constraints[J]. ACM Transactions on Database Systems, 2009, 34(2)

(上接第 128 页)

行步骤如下:

(1)客体资源 o 进行逻辑分割,形成 XML 数据树,树中每个节点为客体中独立的数据单元。

(2)依据 XML 安全标记绑定算法,将安全标记与数据客体进行绑定,形成多级 XML 数据客体。

(3)当某 s 访问数据客体 o 时,将 o 的多级 XML 数据客体通过安全通道发送给 s。安全通道是依据通信双方安全标记协商的具有安全级别的逻辑通道,通过该通道可保证传输数据的机密性、完整性。

(4)当数据客体到达 s 后,依据 s 安全标记对其访问的数据客体进行片段抽取,获得 s 能够访问的数据单元集。

(5)依据客体 o 的类型以及 o 中数据单元之间的关系,对数据单元进行组合,使得 o 相对于主体 s 来说是可读的。同时,s 保存 o 到本地数据库中。至此,完成了主体对客体的安全访问。

结束语 安全标记与数据客体的绑定,是等级保护网络中数据安全共享的关键。本文通过采用 XML 方式,有效地实现了网络数据客体的统一表示,定义了 XML 安全标记,合理巧妙地完成了数据客体与安全标记的绑定。该方法不仅提高了安全标记绑定的灵活性,实现了数据客体与安全标记绑定的统一,而且能够实施更为细粒度的访问控制。同时,数据客体的统一表示与 XML 客体安全标记,也能够解决多级信息系统间异构数据交换访问控制难的问题。当然,安全标记绑定还有许多方面待于研究,比如进程与安全标记、数据流与安全标记的绑定等,今后我们将针对这些方面做进一步的研究。

参 考 文 献

[1] GB/T 22239-2008. 信息安全技术信息系统安全等级保护基本要求[S]. 中国国家标准化管理委员会,2008

[2] Bell P D E, Padula L J L. Secure computer system: unified exposition and multics interpretation [R]. ESD-TR-75-306. MTR 2997 Rev. 1, The MITRE Corporation, 1976

[3] 季庆光,卿斯汉,等. 一个改进的可动态调节的机密性策略模型[J]. 软件学报,2004,15(10):1547-1557

[4] 何建波,卿斯汉,等. 对两个改进的 BLP 模型分析[J]. 软件学报,2007,18(6):1501-1509

[5] Peng P C, Rohatgi P, Keser C. Fuzzy multi-level security: an experiment on quantified risk-adaptive access control[C] // IEEE Symposium on Security and Privacy. Oakland, CA, May 2007: 222-230

[6] Magnani M, Montesi D. A Unified Approach to Structured, Semistructured and Unstructured Data[R]. UBLCS- 2004-9. University of Bologna, 2004

[7] Lee T Y. Formalisms on Semi-structured and Unstructured Data Schema Computations [D]. University of Hong Kong, Hong Kong Special Administrative Region, 2010

[8] 李斓,何永忠,冯登国. 面向 XML 文档的细粒度强制访问控制模型[J]. 软件学报,2004,15(10):1528-1537

[9] Oudkerk S. A Proposal for an XML Confidentiality Label and Related Binding of Metadata to Data Objects [R]. RTO-MP-IST-091 -22. NATO C3 Agency. 2010

[10] Blazic A J, Saljic S. Confidentiality Labeling Using Structured Data Types[C]// 2010 Fourth International Conferences on Digital Society. ST, Maarten, Feb. 2010: 182-187

[11] Pernul G, Winiwarter W, Tjoa A M. The entity-relationship model for multilevel security[C]// Proceedings of the 12th international conference on the entity-relationship approach: entity-relationship approach. Arlington, Texas, USA, December 1994: 166-177