doi:10.11959/j.issn.1000-436x.2015093

面向近邻泄露的数值型敏感属性隐私保护方法

谢静，张健沛，杨静，张冰
(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：提出一种面向近邻泄露的数值型敏感属性隐私保护方法，该方法首先在保护准标识符属性和数值型敏感属性内在关系的前提下，将数值型敏感属性进行离散化划分；然后，提出一种面向近邻泄露的隐私保护原则——(k, ε)- proximity；最后，设计了最大邻域优先算法 MNF(maximal neighborhood first)来实现该原则。实验结果表明，提出的方法能在有效保护数值型敏感信息不泄露的同时保持较高的数据效用，并且保护了数据间的关系。
关键词：隐私保护；数值型敏感属性；近邻泄露；离散化
中图分类号：TP309.2 **文献标识码**：A

Privacy preserving approach based on proximity privacy for numerical sensitive attributes

XIE Jing, ZHANG Jian-pei, YANG Jing, ZHANG Bing
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

Abstract: A model based on proximity breach for numerical sensitive attributes is proposed. At first, it divides numerical sensitive value into several intervals on the premise of protecting the internal relations between quasi-identifier attributes and numerical sensitive attributes. Secondly, it proposes a (k, ε)- proximity privacy preserving principle to defense proximity privacy. In the end, a maximal neighborhood first algorithm (MNF) is designed to realize the (k, ε)-proximity. The experiment results show that the proposed model can preserve privacy of sensitive data well meanwhile it can also keep a high data utility and protect the internal relations.
Key words: privacy preserving; numerical sensitive attributes; proximity privacy; discretization

1 引言

信息化时代中，许多研究机构和学术组织需要对外发布数据，这些数据是理论以及商业研究的宝贵来源，然而在进行数据研究的同时往往会带来安全隐患，造成个人信息的泄露。因此，如何在数据被使用的同时保护个人的隐私是许多机构面临的一个关键问题，特别是统计部门、医疗部门、网络公司等，而数据发布中的隐私保护的目的就是将原始数据进行处理以新的形式发布以避免隐私泄露和抵御多种攻击[1]。本文侧重关系型数据，发布的数据表中主要涉及 3 种属性：1) 身份标识符属性(identifier)，如姓名、身份证号等；2) 准标识符属性(QI, quasi-identifiers)，通过组合可以确定出个体身份的属性，如生日、年龄、性别等；3)包含个人敏感信息的敏感属性(SA, sensitive attributes)，如疾病、工资等。

近几年，数据发布中的隐私保护技术受到越来越多的关注，相关研究成果也日趋增多。目前，常见隐私保护模型多数都是 k-匿名[2,3]和 l-多样性[4,5]模型的扩展，旨在控制等价类中元组的个数以及敏感值的多样性。此外一些新的隐私保护方法也开始出现，如文献[6]采用集成化抽样和泛化的方法构建了一种新的隐私保护模型。文献[7]运用粗糙集理论

收稿日期：2014-01-13；**修回日期**：2014-02-21
基金项目：国家自然科学基金资助项目(61073041, 61073043, 61370083, 61402126)；教育部高等学校博士学科点专项科研基金资助项目(20112304110011, 20122304110012)
Foundation Items: The National Natural Science Foundation of China (61073041, 61073043, 61370083, 61402126); The National Research Foundation for the Doctoral Program of Higher Education of China (20112304110011, 20122304110012)

来控制数据的匿名操作以保护隐私信息。然而，上述研究大部分都是针对敏感属性为分类型属性的情况，由于数值的特性使分类型属性隐私保护方法无法直接适用于数值型属性。 因此，本文旨在研究敏感属性为数值型情况下的隐私保护问题，提出面向近邻泄露的数值型敏感属性隐私保护方法，首先对数值型属性离散化划分，然后提出一种抵御近邻攻击的隐私原则，最后提出最大邻域优先算法实现该原则，本文贡献如下。

1) 提出一种数值型敏感属性离散化划分方法，根据 QI 属性和 SA 属性间的属性相关度，计算出与 SA 属性相关度最大的两维 QI 属性(数值型和分类型属性各一维)，然后依据 SA 属性和它们间的一致度作为标准来进行离散化划分，划分后的结果保护了 SA 属性和 QI 属性间的内在关系，减少了信息的丢失。

2) 对划分后的 SA 属性值引入敏感值泄露风险和近邻泄露风险概念，提出一种近邻泄露保护的隐私原则——(k, ε)-proximity，并证明满足该隐私原则的等价类中元组的近邻泄露风险小于 1/4，最后采用最大邻域优先策略设计算法实现该原则。

2　相关工作

基于分类型敏感属性的隐私保护模型之所以不能直接适用于数值型敏感属性，是由于数值型属性的特性使即使等价类的敏感值满足多样的要求，但是其在数值上却是相近的，攻击者可以推断出敏感值所在的区间，如区间范围较小，那么即使攻击者不能得出确切的敏感值也可以较高的置信度推断出敏感值的范围，从而产生隐私泄露。

为解决数值型敏感属性的隐私保护问题，Zhang 等[8]提出了(k, e)-匿名，要求等价类中元组个数至少为 k 个，并且敏感值的范围要不小于阈值 e。其目的是控制等价类中敏感值不处于过小的区间范围内，然而，该模型中没有考虑敏感属性值在每个等价类取值区间上的分布问题，当某些敏感属性值在其隶属等价类取值区间的出现频率过高时，攻击者仍然能以高概率推导出某一个体的隐私信息。针对此不足之处，Li 等[9]提出了(ε, m)-匿名，要求等价类 E 中具有敏感属性值 v 的元组在该等价类中与它具有相似敏感值的元组最多存在$|E|/m$ 条，其中与 v 相似的敏感值区间定义为$[v-\varepsilon, v+\varepsilon]$。该模型虽然使等价类中相似元组的个数不会太多，避免了某个敏感值出现频率过高的情况，但是没有控制等价类中敏感值的多样性。为了有效地实现等价类中敏感值的多样性，Han 等[10]提出了面向数值型敏感属性的分级 l-多样性模型。该模型首先将数值型敏感属性域分级，再基于分级信息实现数值型敏感属性的 l-多样性。模型在生成等价类的过程中依据分级相异多样度来控制敏感属性的多样性，虽然满足了分级的多样性，但是并不能保证敏感值的多样性，因为其等级划分方法是等间隔划分，如果间隔较小，即使等价类中每个敏感值等级各不相同，但是敏感值会分布在较小的区间内，攻击者仍然会以高概率推导出敏感值的范围。

上述的数值型敏感属性隐私保护方法虽然在一定程度上可以保护数值型属性的隐私信息，然而仍存在隐私泄露风险，其根本原因是由于数值型属性的特性使发布的数据表易产生近邻泄露[11,12]。本文针对数值型敏感属性的近邻泄露问题，提出了(k, ε)- proximity 隐私保护原则进行隐私信息保护，并且提出最大邻域优先算法来实现该原则，最后通过实验证明该方法的有效性。

3　数值型属性离散化

由于数值型属性没有明确的分类层次树，其每一个数值都是单独的一个属性值，即使等价类满足多样性，等价类中数值型属性的取值也可能非常相似，因此需要对数值进行离散化划分。对于数值型属性的划分方法，文献[10]直接采用等间隔来对数值进行划分，文献[13]对原始的数值进行最小间隔合并来进行操作，即选取差值最小的 2 个值合并。文献[10,13]虽然对数值型属性进行了处理，但是没有考虑属性间的关系。本文采用的方法在尽量保护数据信息的同时，对数值型属性进行离散划分。

现实世界的数据中往往存在内部的关系结构，对数值型数据进行离散化会导致隐藏的关联结构被破坏，甚至会产生无意义的区间，因此，离散化之后结果的质量好坏会受到数据间内在联系的影响。为了能够获得更好的离散化结果，本文采用属性相关度来反映 QI 属性和 SA 属性之间的关系。下面给出属性相关度的定义。

设用户要发布的数据表 $T=\{QI_1,\cdots, QI_i,\cdots, QI_n, SA\}$，其中，$QI_i(1\leqslant i\leqslant n)$为准标识符属性，$SA$ 为数值型敏感属性。对于 T 中的任一条元组 t，$t[A]$表示元组 t 中属性 A 的取值。

定义 1 属性相关度。给定数据表 T，$QI_i(1\leqslant i\leqslant n)$为 T 中任意一维 QI 属性，SA 为数值型敏感属性，QI_i 相对于 SA 的属性相关度 $Rel(QI_i)$定义如下。

1) 若 QI_i 为分类型属性，其值域为$\{v_1,v_2,\cdots,v_d\}$，属性相关度为

$$Rel_{\text{cat}}(QI_i)=d\Big/\sum_{j=1}^{d}D(A_j)$$

其中，$A_j=\{t\mid t[QI_i]=v_j,t\in T,1\leqslant j\leqslant d\}$，$D(A_j)$表示 A_j 中元组敏感值的方差。

2) 若 QI_i 为数值型属性，属性相关度为

$$Rel_{\text{num}}(QI_i)=\left|\frac{\sum_{k=1}^{|T|}\left(v_k^{QI_i}-\overline{v^{QI_i}}\right)\left(v_k^{SA}-\overline{v^{SA}}\right)}{\sqrt{\sum_{k=1}^{|T|}\left(v_k^{QI_i}-\overline{v^{QI_i}}\right)^2}\sqrt{\sum_{k=1}^{|T|}\left(v_k^{SA}-\overline{v^{SA}}\right)^2}}\right|$$

其中，$v_k^{QI_i}$ 和 v_k^{SA} 分别表示数据表 T 中第 k 个元组在属性 QI_i 和 SA 上的取值，$\overline{v^{QI_i}}$ 和 $\overline{v^{SA}}$ 分别表示属性 QI_i 和 SA 上取值的平均值。

为了能够保护属性间的内在关系，选择与敏感属性相关度最高的二维 QI 属性(最优的分类型和数值型属性)，在合并的过程中计算信息转换量来选择最优的区间进行合并。对于分类和数值型属性，由于其数据类型的区别，其信息转换量的计算方法不同，下面分别进行详细介绍。

3.1 信息转换量计算方法

表 1 为分类 QI 属性与 SA 属性离散区间的二维信息表，表中每一列对应一个分类属性值，每一行对应于一个数据区间。δ_{ij} 表示区间 I_i 的所有元组中敏感值为 v_j 的元组个数，δ_j 表示敏感值为 v_j 的元组数，$\delta_{i\cdot}$表示区间 I_i 中元组数，M 表示相邻区间的元组总数，$M=\delta_i+\delta_{(i+1)}$。

表 1 二维信息

相邻区间	v_1	v_2	…	v_m	行和
$I_i:(d_{i-1},d_i]$	δ_{i1}	δ_{i2}	…	δ_{im}	$\delta_{i\cdot}$
$I_{i+1}:(d_i,d_{i+1}]$	$\delta_{i+1,1}$	$\delta_{i+1,2}$	…	$\delta_{i+1,m}$	$\delta_{(i+1)\cdot}$
列和	δ_1	δ_2	…	$\delta_{\cdot m}$	M

定义 2 分类型属性信息转换量。给定数据表 T，A 为 QI 中的分类型属性，I_i，I_{i+1} 为 SA 中 2 个相邻区间，$I=I_i\cup I_{i+1}$，则区间 I_i，I_{i+1} 合并后分类型属性的信息转换量定义为

$$Info_{\text{cat}}(I_i,I_{i+1})=H(I)-H(I_i,I_{i+1})$$

其中，合并后区间 I 信息熵 $H(I)=-\sum_{j=1}^{m}\frac{\delta_{ij}+\delta_{i+1,j}}{M}\cdot\log\frac{\delta_{ij}+\delta_{i+1,j}}{M}$，相邻区间 I_i，I_{i+1} 加权信息熵 $H(I_i,I_{i+1})=\frac{\delta_{i\cdot}}{M}H(I_i)+\frac{\delta_{(i+1)\cdot}}{M}H(I_{i+1})$。

根据熵函数的上凸性，可知 $H(I_i,I_{i+1})\leqslant H(I)$，因此，$Info_{\text{cat}}(I_i,I_{i+1})\geqslant 0$

对于数值型属性，由于其没有明显的类别关系，所以在这里采用方差代替信息熵来度量区间合并前后的信息转换量。

定义 3 数值型属性信息转换量。给定数据表 T，B 为 QI 中的数值型属性，I_i，I_{i+1} 为 SA 中 2 个相邻区间，$I=I_i\cup I_{i+1}$，则区间 I_i，I_{i+1} 合并后数值型属性的信息转换量定义为

$$Info_{\text{num}}(I_i,\ I_{i+1})=D(I)-D(I_i,\ I_{i+1})$$

其中，$D(I)$表示区间 I 中所有元组在属性 B 上取值的方差，相邻区间 I_i，I_{i+1} 的加权方差 $D(I_i,I_{i+1})=\frac{n_i}{n_i+n_{i+1}}D(I_i)+\frac{n_{i+1}}{n_i+n_{i+1}}D(I_{i+1})$。

定理 1 给定数据表 T，I_i，I_{i+1} 为属性 SA 中 2 个相邻区间，$I=I_i\cup I_{i+1}$，区间 I_i，I_{i+1} 的加权方差为 $D(I_i,\ I_{i+1})$，合并区间 I 的方差为 $D(I)$，则 $D(I)-D(I_i,I_{i+1})\geqslant 0$。

证明 由方差的定义可知

$$\begin{aligned}&D(I)-D(I_i,I_{i+1})\\&=\frac{\sum_{j=1}^{n_i}\left(\bar{v}-v_j^i\right)^2+\sum_{k=1}^{n_{i+1}}\left(\bar{v}-v_k^{i+1}\right)^2}{n_i+n_{i+1}}-\frac{n_i}{n_i+n_{i+1}}\frac{\sum_{j=1}^{n_i}\left(\overline{v^i}-v_j^i\right)^2}{n_i}-\\&\quad\frac{n_{i+1}}{n_i+n_{i+1}}\frac{\sum_{k=1}^{n_{i+1}}\left(\overline{v^{i+1}}-v_k^{i+1}\right)^2}{n_{i+1}}\\&=\frac{1}{n_i+n_{i+1}}\left[\sum_{j=1}^{n_i}\left(\bar{v}-v_j^i\right)^2-\sum_{j=1}^{n_i}\left(\overline{v^i}-v_j^i\right)^2+\right.\\&\quad\left.\sum_{k=1}^{n_{i+1}}\left(\bar{v}-v_k^{i+1}\right)^2-\sum_{k=1}^{n_{i+1}}\left(\overline{v^{i+1}}-v_k^{i+1}\right)^2\right]\\&=\frac{1}{n_i+n_{i+1}}\left[\underbrace{\sum_{j=1}^{n_i}\left(\bar{v}+\overline{v^i}-2v_j^i\right)\left(\bar{v}-\overline{v^i}\right)}_{①}+\right.\\&\quad\left.\underbrace{\sum_{k=1}^{n_{i+1}}\left(\bar{v}+\overline{v^{i+1}}-2v_k^{i+1}\right)\left(\bar{v}-\overline{v^{i+1}}\right)}_{②}\right]\end{aligned}$$

对上式中①通过推导可得

$$\sum_{j=1}^{n_i}\left(\bar{v}+\overline{v^i}-2v_j^i\right)\left(\bar{v}-\overline{v^i}\right)$$
$$=\left(\bar{v}-\overline{v^i}\right)\sum_{j=1}^{n_i}\left(\bar{v}+\overline{v^i}-2v_j^i\right)$$
$$=\left(\bar{v}-\overline{v^i}\right)\left[n_i\left(\bar{v}+\overline{v^i}\right)-2\sum_{j=1}^{n_i}v_j^i\right]$$
$$=n_i\left(\bar{v}-\overline{v^i}\right)^2$$

同理，对上式中②进行化简，最后得出

$$D(I)-D(I_i,I_{i+1})=\frac{n_i\left(\bar{v}-\overline{v^i}\right)^2+n_{i+1}\left(\bar{v}-\overline{v^{i+1}}\right)^2}{n_i+n_{i+1}}\geqslant 0$$

证毕。

由定理 1 可知，数值型属性信息转换量 $Info_{\text{num}}(I_i,\ I_{i+1})\geqslant 0$。

3.2　敏感属性离散化算法

在 3.1 节中分别介绍了分类型和数值型属性的信息转换量计算方法，由定义 2 和定义 3 可以给出下面的定义。

定义 4　合并信息转换量。给定数据表 T，令 A、B 分别为 T 的 QI 属性中与 SA 属性相关度最高的分类型和数值型属性，I_i，I_{i+1} 为 SA 中 2 个相邻区间，$I=I_i\cup I_{i+1}$，区间 I_i，I_{i+1} 合并的信息转换量定义为

$$f\left(I_i,I_{i+1}\right)=Info_{\text{cat}}\left(I_i,I_{i+1}\right)+Info_{\text{num}}\left(I_i,I_{i+1}\right)$$

由 3.1 节中的介绍可知 $f\left(I_i,I_{i+1}\right)\geqslant 0$。

在区间合并前后，会出现信息转换，当信息转换量小时，说明 2 个相邻的区间的分布越相似，应该优先进行合并。在合并之后必然会产生信息损失，为了避免产生较大的信息损失，给出一致度概念来控制合并的过程。

定义 5　一致度。给定数据表 T，A、B 为 T 中任意两维属性，属性 A 对于属性 B 的一致度定义为

$$con(A,B)=\frac{\sum_{j=1}^{m}|s(A,B_j)|}{|T|}$$

其中，$|T|$为数据表的基数。属性 A 的值域为 $\{v_1,\cdots,v_i,\cdots,v_d\}$，$A_i$表示属性 A 中取值为 v_i的元组集合；同样地，属性 B 的值域为 $\{v_1,\cdots,v_j,\cdots,v_m\}$，$B_j$ 表示属性 B 中取值为 v_j 的元组集合，$s(A,B_j)=\{x\,|\,x\in A_i,A_i\subseteq B_j,1\leqslant i\leqslant d,1\leqslant j\leqslant m\}$。

离散化首先将 SA 的属性值从小到大排序，然后初始化相邻区间，接着采用自底向上的思想，根据合并信息转换量评价相邻区间，选择最优值合并相邻区间，如此迭代进行，直到信息损失达到一定的阈值，即合并后数据的一致度小于原始数据的一致度的 λ 倍 $(0<\lambda<1)$，则合并结束，将数值型值域转换成有限的区间。具体算法如下。

算法 1　数值型属性离散化算法。

输入：数据表 T，分类型属性 A，数值型属性 B，敏感属性 SA；

输出：敏感属性离散化后的数据表 T^*。

1) 计算一致度 $c=con(A,SA)+con(B,SA)$；
2) 初始化 $c'=c$；
3) 将属性 SA 的值域升序排列 $\{s_1,s_2,\cdots,s_n\}$；
4) 初始化区间集合 $I=\{[s_1,s_2],(s_2,s_3],\cdots,(s_{n-1},s_n]\}$；
5) while $c'\geqslant\lambda c$ do
6)　　for 集合 I 中的每个区间 I_i
7)　　计算信息转换量 $f\left(I_i,I_{i+1}\right)$；
8)　　end for
9)　　合并 $f\left(I_i,I_{i+1}\right)$ 值最小的 2 个相邻区间;
10)　　更新区间集合 I;
11)　　计算离散后的一致度 c';
12) end while
13) return 敏感属性离散化后的数据表 T^*。

4　(*k*, *ε*)- proximity 原则

离散后的区间保持了 SA 属性和 QI 属性之间存在的内在关系，但是由于数值型属性的特点，各个区间包含的信息量不同，比如攻击者得知某个人的薪资的某个区间为[1 000, 1 050]，虽然不能知道该个体具体的薪资，但是个体的薪资在一个很小的区间内，使个体的隐私泄露。对于薪资区间[2 000, 4 000]和[10 000, 12 000]，虽然 2 个区间的间隔都为 2 000，但是由于区间[10 000, 12 000]本身的取值较大，所以其近邻泄露的风险远大于区间[2 000, 4 000]的风险。因此，需要衡量离散后区间的泄露风险，本文中对于离散后的敏感值区间简称为敏感值。

定义 6　敏感值泄露风险。给定数据表 T^*，SA 为敏感属性，$\boldsymbol{B}$ 为 SA 离散化后区间的集合并按升序排列，$\boldsymbol{B}=\{r_i\,|\,r_i=(s_i,s_{i+1}],1\leqslant i\leqslant h\}$，对于 $\forall r_i\in\boldsymbol{B}$，其泄露风险 $\eta_i=s_i/s_{i+1}$。

定义 7 元组的 ε 邻域集。给定数据表 T^*，t 是等价类 E 中任意元组，$t[SA]=r_i$，$r_i \in \boldsymbol{B}$，t 的邻域集 $N_\varepsilon(t)=\{t' \mid t' \neq t, t'[SA]=r_j, r_j \subset [s_i-\varepsilon, s_{i+1}+\varepsilon]\}$。

定义 8 (k, ε)- proximity 原则。给定数据表 T^*，E 为 T^*的等价类，E 满足(k, ε)- proximity 原则如果等价类 E 满足下列条件：

1) 等价类 E 中至少包含 k 个元组；

2) 对于 E 中的任意元组 t，其敏感属性值的泄露风险为 η，在等价类 E 中，t 的 ε 邻域集至多包含 $(1-\eta)(|E|-1)$ 个元组。

定义 9 元组 t 的近邻泄露风险。给定数据表 T^*，t 是等价类 E 中任意元组，则元组 t 的近邻泄露风险

$$risk(t,\varepsilon)=\frac{\eta|N_\varepsilon(t)|}{|E|} \tag{1}$$

其中，η 是 t 中敏感值的泄露风险，$|N_\varepsilon(t)|$ 表示 t 在等价类 E 中的 ε 邻域集中元组个数，$|E|$ 表示等价类中元组个数。

文献[9]中的敏感值没有进行离散划分，因此每个敏感值的泄露风险是 1，当η 的取值为 1 时，泄露风险为 $\frac{|N_\varepsilon(t)|}{|E|}$，即文献[9]中给出的近邻泄露风险公式。

由定义 9，可以进一步定义等价类 E 的近邻泄露风险为 $risk(E,\varepsilon)=\max\limits_{\forall t \in E} risk(t,\varepsilon)$，即为等价类 E 中所有元组中近邻泄露风险的最大值。数据表 T^* 的近邻泄露风险为 $risk(T^*,\varepsilon)=\max\limits_{\forall E \in T^*} risk(E,\varepsilon)$。

定理 2 若数据表 T^*中的等价类 E 满足(k, ε)- proximity 原则，则 E 中每个元组的近邻泄露风险都将小于 1/4。

证明 对于等价类 E 中的任意元组 t，由于其满足(k, ε)- proximity 原则，那么在 E 中 t 的邻域集至多包含 $(1-\eta)(|E|-1)$ 个元组，可以得出元组 t 的近邻泄露风险

$$risk(t,\varepsilon)=\frac{\eta|N_\varepsilon(t)|}{|E|} \leqslant \frac{\eta(1-\eta)(|E|-1)}{|E|}$$

$\eta \in (0,1)$，函数 $\eta(1-\eta)$ 在 $\left(0,\frac{1}{2}\right)$ 上单调递增，在 $\left(\frac{1}{2},1\right)$ 上单调递减，因此 $\eta=\frac{1}{2}$ 时，$\eta(1-\eta)$ 取最大值 $\frac{1}{4}$，那么 $risk(t,\varepsilon) \leqslant \frac{|E|-1}{4|E|}<\frac{1}{4}$。证毕。

定理 3 对于包含 k 个元组的等价类 E，如果其任意元组的 ε 邻域集都不包含等价类中的其他元组，那么该等价类满足(k, ε)- proximity 原则。

证明 等价类 E 中包含 k 个元组，满足定义 8 中的条件 1)。因为 E 中任意元组的邻域集都不包含等价类中的其他元组，即元组的邻域集为空，定义 8 中条件 2)要求元组的邻域集最多为 $(1-\eta)(k-1)$，显而易见，等价类满足条件 2)，因为 E 中每个元组的邻域集都为空，因此该等价类(k, ε)- proximity 原则。证毕。

5 数值型敏感属性发布方法

本节给出一种最大邻域优先算法来实现(k, ε)- proximity 原则，其采用最大邻域优先策略，即在选取元组的过程中，优先选取邻域集最大的元组。之所以优先选取邻域集大的元组，是因为邻域集越大说明其 ε 邻域内的元组越多，优先处理该类元组避免到后期所剩元组存在于一个较小的邻域内，无法再生成满足要求的等价类，带来较大的信息损失。

算法基本思想为：首先根据最大邻域优先策略将数据表中的元组划分成多个包含 k 个元组的分组，并且该 k 个元组的邻域集各不相交，由定理 3 可知生成的分组满足(k, ε)- proximity 原则；其次，对于剩余元组，检查是否能添加到存在等价类中而不破坏(k, ε)- proximity 原则；最后，将无法添加的元组隐匿。

算法 2 最大邻域优先算法(MNF)

输入：敏感属性划分后的数据表 T^*，阈值 k，ε;

输出：泛化后数据表 T'。

//分组阶段

1) 计算 T^*中每个元组的邻域集 $N_\varepsilon(t_i)$；

2) Q={将元组按照 $N_\varepsilon(t_i)$ 的基数降序排列};

3) $E=\varnothing$，$R=\varnothing$，$G=\varnothing$；

4) while $|Q| \geqslant k$ do

5) for j=1:k

6) if Q 中存在未被标记的元组 then

7) {选择第一个未被标记的元组 t，$E=E \cup \{t\}$，$Q=Q \setminus \{t\}$；

8) 标记邻域集中包含 t 的元组 t_i，且更新元组 t_i 的邻域集 $N_\varepsilon(t_i) \setminus \{t\}$；

9) 更新 Q;//将元组按照更新后邻域集的基数重新进行降序排列}

10) else

11) break;

12) end if

13)　end for
14)　if $|E|<k$ then
15)　　$R=R\cup E$;
16)　else
17)　　$G=G\cup E$;
18)　end if
19)　$E=\varnothing$;
20)　取消 Q 中所有元组的标记;
21) end while
22) $R=R\cup Q$;//将 Q 中剩余的元组加入集合 R 中
//处理剩余元组阶段
23) for　R 中每一个剩余元组 t'
24)　if　存在集合 $g\in G$，添加 t' 后仍满足 (k, ε)- proximity　then
25)　　$g=g\cup\{t'\}$;
26)　else
27)　　将 t' 隐匿;
28)　end if
29) end for
30) 将集合 G 以 T' 形式发布;

6　实验结果及分析

本节中通过实验分析(k, ε)- proximity 的性能，并将其与文献[9]提出的(ε, m)-anonymity、文献[10]提出的 multi-level distinct l-diversity 进行比较。实验所采用的数据集为实际数据集 SAL，数据集来自 http: //ipums.org/。该数据集包含部分美国人口信息，总共包含 5×10^5 条元组。本文选取数据集中的 7 个属性作为研究对象，假设其中的 6 个属性{*Age*，*Sex*，*Race*，*Country*，*Birthplace*，*Occupation*}为准标识符属性，*Income* 为敏感属性。

6.1　信息损失分析

本节采用文献[14]介绍的信息损失度量方法作为信息损失的衡量标准，该度量的取值范围为[0,1]，其值越小，说明信息损失越少。

图 1~图 3 分别给出了准标识符属性 QI 维数、ε 值和数据集大小变化对(k, ε)- proximity、(ε, m)-anonymity 以及 multi-level distinct l-diversity 信息损失的影响，其中，对于(ε, m)-anonymity 分别取 m=2，m=5 这 2 种情况。由图 1 可知，当准标识符属性 QI 维数增大时，4 种算法的信息损失也随之增大，这是由于 QI 维数的增大使在每个元组上进行泛化的属性数目增加，在泛化过程中的信息损失也将增大。图 2 给出了不同 ε 值下，3 种算法信息损失的比较结果，由于 multi-level distinct l-diversity 算法中没有参数 ε，因此图 2 中给出了剩余 3 种算法的实验结果。由图 2 可以看出，当 ε 值增大时，3 种算法的信息损失也将增大，这是因为 ε 值增大，对于等价类中敏感值之间的差值要求更大，隐私程度的要求也更加严格，因此信息损失将增加。图 3 给出了在不同数据量|T|下 4 种算法信息损失的比较结果，随着数据集的增大，算法的信息损失都将增大，这是由于数据量增大，将导致泛化的元组增多，因此信息损失也将增大。

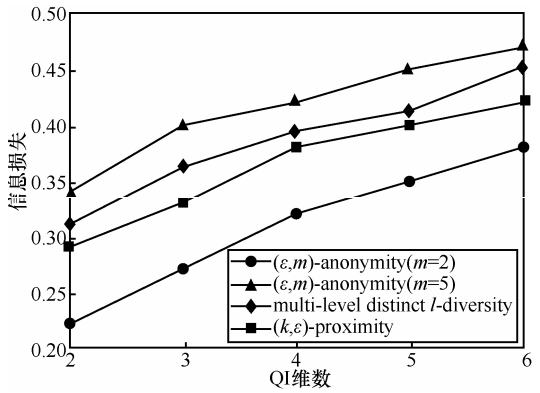


图 1　不同 QI 维数下信息损失的比较

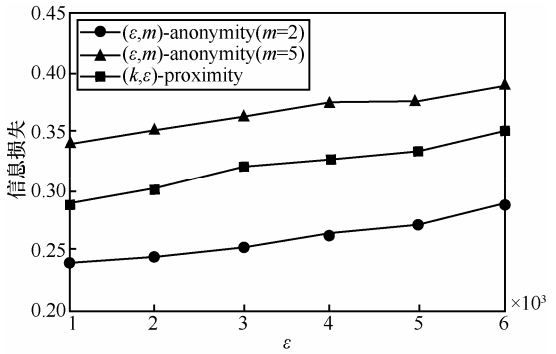


图 2　不同 ε 值下信息损失的比较

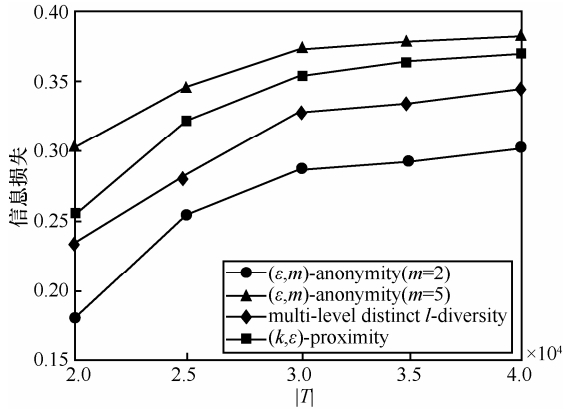


图 3　不同数据集下信息损失的比较

由图 1~图 3 可知，1) 对于(ε, m)-anonymity，m=5 时的信息损失比 m=2 的信息损失要大。这是因为当 m=2 时，其要求对于每个敏感值等价类中至多有 $\frac{1}{2}|E|$ 的元组与其在敏感值上相似，隐私程度要求较松。而当 m=5 时，其要求对于每个敏感值等价类中至多有 $\frac{1}{5}|E|$ 的元组与其在敏感值上相似，隐私程度要求较严格，因此 m=5 时的信息损失比 m=2 的信息损失要大；2)(k, ε)- proximity 的信息损失比(ε, m)-anonymity (m=2) 的信息损失大，比(ε, m)-anonymity (m=5)的信息损失小，对于(k, ε)-proximity，其要求等价类中与每个敏感值相似的元组数最多为 $(1-\eta)(|E|-1)$，在参数 λ 的控制下，多数离散后敏感值的泄露风险 η 会处于 $\frac{1}{2}$ 到 $\frac{4}{5}$ 之间，即等价类中与每个敏感值相似的元组数最多为 $\frac{1}{5}(|E|-1)$ 与 $\frac{1}{2}(|E|-1)$，它的隐私程度比 m=2 时要严格，比 m=5 时宽松，因此，(k, ε)- proximity 的信息损失比(ε, m)-anonymity(m=2)的信息损失大，比(ε, m)-anonymity(m=5)的信息损失小；3) (k, ε)- proximity 的信息损失比 multi-level distinct l-diversity 的信息损失小，这是由于 multi-level distinct l-diversity 在分级过程中没有考虑属性间的关系，在泛化过程中带来的信息损失将更大，因此，multi-level distinct l-diversity 的信息损失较大。

此外，由图 1~图 3 可以看出，虽然(ε, m)-anonymity (m=2) 的信息损失比(k, ε)- proximity 的小，但是由于 m=2 的阈值要求，使 (ε, m)-anonymity 中的近邻泄露风险可以达到 0.5，产生隐私泄露的概率过高，而由定理 2 可知，(k, ε)- proximity 中每个元组的近邻泄露风险都小于 0.25。因此，虽然(ε, m)-anonymity (m=2)信息损失小，但是其不能更好地保护隐私信息不泄露。这也正说明了数据的可用性和隐私性之间的权衡问题。

6.2 隐匿率分析

本节采用文献[15]介绍的隐匿率作为度量标准，它的取值范围是[0,1]，其值越小，说明删除的元组数越少，信息损失也就越小。由于(k, ε)- proximity 实以及 multi-level distinct l-diversity 算法中没有隐匿元组，因此，图 4~图 6 分别给出了准标识符属性 QI 维数、ε 值和数据集大小变化对(k, ε)- proximity 隐匿率的影响，分别取 λ=0.3，λ=0.6，λ=0.8 这 3 种情况。

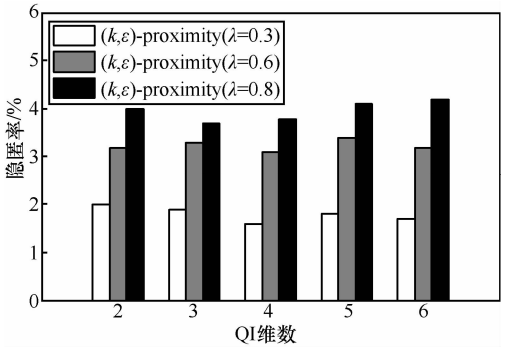


图 4 不同 QI 维数下隐匿率的比较

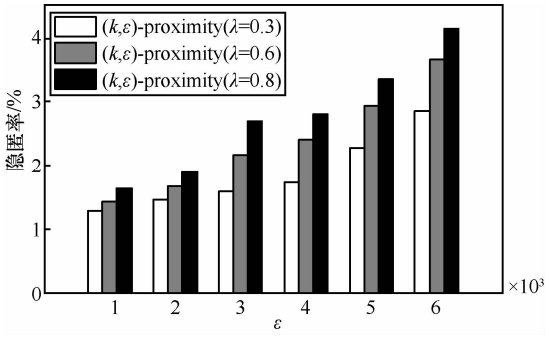


图 5 不同 ε 值下隐匿率的比较

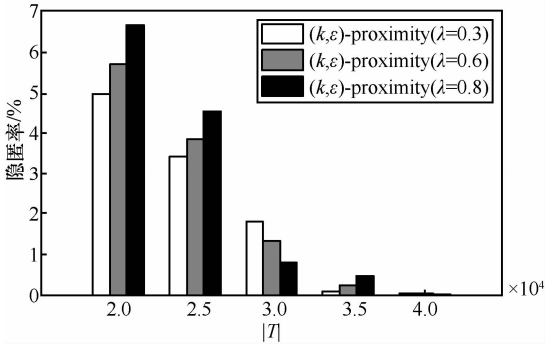


图 6 不同数据集下隐匿率的比较

由图 4 可知，当准标识符属性 QI 维数增大时，3 种算法的隐匿率波动很小，由算法 2 可知，QI 维数的变化并不影响隐匿元组的个数，产生的细微波动是由选取元组的随机性而带来的。在图 5 中，当 ε 值增大时，3 种算法的隐匿率都将增加，这是由于 ε 值增大，使隐私要求更加严格，在生成等价类中的过程中，会产生更多的不满足要求的元组，这些元组都将被隐匿，因此，隐匿率都将增大。由图 6 可知，当数据集增大时，3 种算法的隐匿率都减小，特别地，当$|T|=4\times10^4$ 时，隐匿率都近乎为 0。这是由于随着数据集中元组增多，在分组阶段可以选择的候选元组增多，而且，数据集的增多使分组的数量增多，在处理剩余元组阶段中，剩余元组可供选择的分组增多，隐匿的元组将减少，从而引起隐匿

率降低。

由图 4~图 6 可知，在同等条件下，λ=0.3 时，(k, ε)- proximity 的隐匿率最低，λ=0.8 时，(k, ε)-proximity 的隐匿率最高。这是由于 λ 控制离散化过程中信息损失的程度，其值越大，则允许的信息损失越小，合并次数将减少，使离散化后的敏感值的泄露风险 η 变大，导致 $(1-\eta)(|E|-1)$ 变小，则等价类中每个元组的 ε 邻域内元组个数将变少，隐私要求更加严格，因此，当 λ 增大时，(k, ε)- proximity 的隐匿率会增加。

6.3　执行时间分析

图 7~图 9 分别给出了 QI 维数、ε 值和数据集大小变化对(k, ε)- proximity、(ε,m)- anonymity 以及 multi- level distinct l-diversity 执行时间的影响，其中，对于(ε, m)-anonymity 分别取 m=2，m=5 这 2 种情况。

由图 7 可知，当 QI 维数增大时，(k, ε)- proximity 和(ε, m)-anonymity 的执行时间会增大但是增加幅度很小，这是由于 QI 维数的增大使得需要泛化的属性增多，产生了一些时间的消耗，但是 QI 维数的增大并不影响算法的分组阶段和处理剩余元组阶段，因此，QI 维数的增大造成的时间损失并不是很大，执行时间只是产生了细小的波动。然而，随着 QI 维数的增大，multi-level distinct l-diversity 的执行时间增长较快，因为 QI 维数增大，每次搜索所要考虑的因素会增多，相应的计算量会变大，因此时间开销会变大。图 8 给出了不同 ε 值下，3 种算法执行时间的比较结果(由于 multi-level distinct l-diversity 算法中没有参数 ε，因此图 8 中给出了剩余 3 种算法的实验结果)。在图 8 中，当 ε 值增大时，(k, ε)- proximity 的执行时间将增大，这是由于 ε 值增大，使隐私要求更加严格，为了能够生成满足要求的等价类，将会增加迭代次数，此外，由于隐私要求严格，会产生更多的剩余元组，在处理剩余元组的过程中也将消耗更多的时间，因此，(k, ε)- proximity 的执行时间将增大。此外，由图 8 可知，(ε, m)-anonymity 在 m=2，m=5 这 2 种情况下执行时间都是先增大后减小，这是由于随着 ε 值的增大，算法在划分阶段将带来更多的时间消耗，而当 ε 值增大到一定程度时，使能满足隐私要求的分组越来越少，减少了泛化的时间，因此，执行时间先增大后减小。由图 9 可知，当数据集增大时，4 种算法的执行时间都增大，这是由于随着数据集增大，在算法执行过程中，需要处理的元组数增多，因此导致执行时间增加。

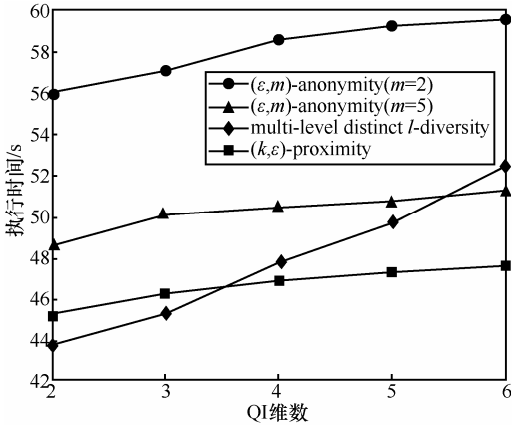


图 7　不同 QI 维数下执行时间的比较

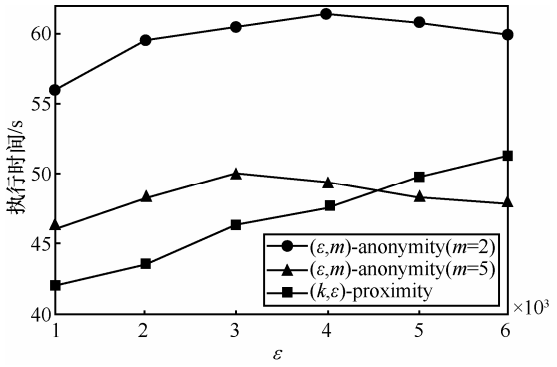


图 8　不同 ε 值下执行时间的比较

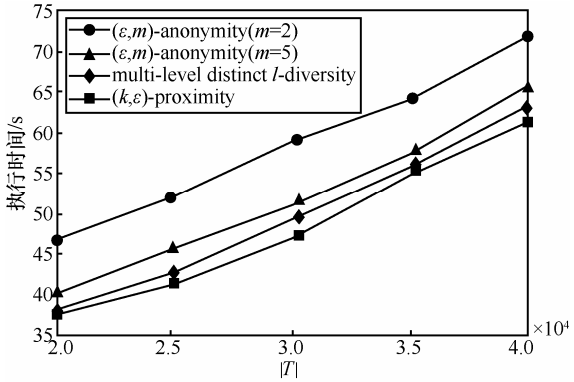


图 9　不同数据集下执行时间的比较

由图 7~图 9 可知，1)对于(ε, m)-anonymity，m=2 时的执行时间比 m= 5 要大，因为随着 m 值的增大，隐私要求程度越严格，满足要求的分组减少，迭代会提前终止，因此，(ε, m)-anonymity(m=5)比(ε, m)-anonymity(m=2)的执行时间短；2)(k, ε)- proximity 的执行时间比(ε, m)-anonymity(m=2)和(ε, m)- anonymity(m=5)的执行时间短。这是由于(ε, m)- anonymity 首先采用 Mondrian[16]算法进行划分，然后再依次检查划分后的分组是否满足要求，对于不满足要求的分组则继续划分。(k, ε)- proximity 中每次只

选择 k 个元组来形成分组，减少了分组过程中的时间消耗，因此，(k, ε)- proximity 的执行时间较短。3) (k, ε)- proximity 的执行时间比 multi-level distinct l-diversity 的短，因为 multi-level distinct l-diversity 每次搜索都与 QI 维数有关，只有当 QI 维数较小时(图 7 中 QI 维数为 2 和 3 时)，才能具有较短的执行时间，其他情况下则执行时间较长。

综上所述，与(ε, m)-anonymity 和 multi- level distinct l-diversity 相比，提出的(k, ε)- proximity 在保证执行效率的同时，能更好地保护隐私信息不泄露。此外，用户可自定义 λ 值来更加灵活地控制数据的效用。

7 结束语

针对数值型敏感属性的近邻泄露问题，提出一种面向近邻泄露的隐私保护方法，首先提出一种数值型敏感属性离散化划分方法，保护了 SA 属性和 QI 属性间的内在关系，减少了信息的丢失；其次，提出一种近邻泄露保护的(k, ε)- proximity 隐私原则，并证明满足该隐私原则的等价类中元组的近邻泄露风险小于 1/4；最后采用最大邻域优先策略设计算法 MMF 实现该原则。实验表明本文的方法在有效地保护敏感信息不泄露的同时保持了较高的数据效用。

随着大数据时代的来到，大数据处理过程中所产生的隐私泄露为个人带来严重困扰，大数据的隐私保护也成为一个热点的研究方向，引起了许多学者的广泛关注。本文所提出的方法只针对单一数据集进行处理，而当面临大数据的情况下，数据来源会更加丰富，即使满足隐私保护的需求，攻击者也可能根据不同来源的数据进行推理攻击来获取隐私信息，因此，大数据下的隐私保护问题将更为复杂，下一步研究重点着力于如何在数据量更大，数据来源更丰富的情况下对用户的隐私信息进行保护。

参考文献：

[1] XU Y, MA T, TANG M, *et al*. A survey of privacy preserving data publishing using generalization and suppression[J]. Appl Math, 2014, 8(3): 1103-1116.

[2] SWEENEY L. *k*-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge based Systems, 2002, 10(5):557-570.

[3] TASSA T, MAZZA A, GIONIS A. *k*-concealment: an alternative model of *k*-type anonymity[J]. Transactions on Data Privacy, 2012, 5(1): 189-222.

[4] MACHANAVAJJHALA A, GEHRKE J, KIFER D. *l*-diversity: privacy beyond *k*-anonymity [J]. ACM Transactions on Knowledge Discovery from Data,2007,1(1):1-52.

[5] ABDALAAL A, NERGIZ M E, SAYGIN Y. Privacy-preserving publishing of opinion polls[J]. Computers & Security, 2013, 37(9): 143-154.

[6] SAROWAR S A H M, LI J, DING X, *et al*. A general framework for privacy preserving data publishing[J]. Knowledge-Based Systems, 2013, 54(12): 276-287.

[7] YE M, WU X, HU X, *et al*. Anonymizing classification data using rough set theory[J]. Knowledge-Based Systems, 2013, 43(5): 82-94.

[8] ZHANG Q, KOUDAS N, SRIVASTAVA D, *et al*. Aggregate query answering on anonymized tables[A].Data Engineering, 2007, ICDE 2007, IEEE 23rd International Conference[C]. 2007. 116-125.

[9] LI J X, TAO Y F, XIAO X K. Preservation of proximity privacy in publishing numerical sensitive data [A]. Proceedings of ACM Conference on Management of Data (SIGMOD) [C]. 2008. 473-486.

[10] 韩建民, 于娟, 虞慧群等. 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 J M, YU J, YU H Q, *et al*. A multi-level *l*-diversity model for numerical sensitive attributes[J]. Journal of Computer Research and Development, 2011, 48(1):147-158.

[11] WANG T, MENG S, BAMBA B, *et al*. A general proximity privacy principle[A].Data Engineering, 2009, ICDE'09, IEEE 25th International Conference[C]. 2009. 1279-1282.

[12] WANG T, LIU L. XColor: protecting general proximity privacy[A].Data Engineering (ICDE), 2010 IEEE 26th International Conference[C]. 2010. 960-963.

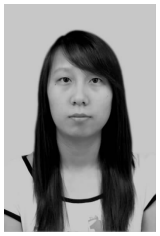
[13] CAMPAN A, COOPER N, TRUTA T M. On-the-fly Generalization Hierarchies for Numerical Attributes Revisited[M].Secure Data Management, Springer Berlin Heidelberg, 2011.18-32.

[14] LEFEVRE K, DEWITT D J, RAMAKRISHNAN R.Incognito: Efficient full-domain *k* –anonymity[A].Proc of the 2005 ACM SIGMOD Int Conf on Management of data[C]. New York, 2005.49-60.

[15] 杨晓春, 王雅哲, 王斌等. 数据发布中面向多敏感属性的隐私保护方法[J].计算机学报, 2008, 31(4):574-587.
YANG X C, WANG Y Z, WANG B, *et al*. Privacy preserving approaches for multiple sensitive attributes in data publishing [J].Chinese Journal of Computers, 2008,31(4):574-587.

[16] LEFEVRE K, DEWITT D J, RAMAKRISHNAN R. Mondrian multidimensional *k*-anonymity[A].Proc of the 22nd Int Conf on Data Engineering[C]. New York, 2006.1-11.

作者简介：



谢静（1986-），女，湖北随州人，哈尔滨工程大学博士生，主要研究方向为数据挖掘、隐私保护。

张健沛（1956-），男，黑龙江哈尔滨人，哈尔滨工程大学教授、博士生导师，主要研究方向为数据挖掘、隐私保护、社会网络等。

杨静（1962-），女，黑龙江哈尔滨人，哈尔滨工程大学教授、博士生导师，主要研究方向为数据挖掘、隐私保护、机器学习等。

张冰（1986-），女，黑龙江哈尔滨人，哈尔滨工程大学博士生，主要研究方向为数据挖掘、隐私保护。