一种面向聚类的对数螺线数据扰动方法

黄茂峰 倪巍伟 王佳俊 孙福林 崇志宏
(东南大学计算机科学与工程学院 南京 211189)

摘 要 面向挖掘应用的隐私保护数据发布要求对数据集进行隐藏的同时维持数据的挖掘可用性,数据扰动是解决该问题的有效方法.现有的面向聚类的数据扰动方法难以兼顾原始数据个体隐私和维持数据聚类可用性,对此提出了一种基于对数螺线的隐私保护数据干扰方法.通过构建面向聚类的隐私保护数据扰动模型,利用对数螺线对原始数据进行扰动隐藏,维持原始数据的 k 邻域关系稳定,实现数据集聚类可用性的有效维护;进一步提出多重对数螺线扰动的策略,提高隐私保护强度.理论分析和实验结果表明:文中方法能够有效地避免数据隐私泄露,同时维持数据的聚类可用性.

关键词 隐私保护;数据挖掘;聚类分析;对数螺线;数据干扰
中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2012.02275

A Logarithmic Spiral Based Data Perturbation Method for Clustering

HUANG Mao-Feng NI Wei-Wei WANG Jia-Jun SUN Fu-Lin CHONG Zhi-Hong
(*School of Computer Science & Engineering, Southeast University, Nanjing* 211189)

Abstract Privacy-Preserving Data Publication requires seeking tradeoffs among maintaining data utility and preserving privacy of the dataset, data perturbation is an effective method to meet this requirement. The existing data perturbation algorithms for clustering usually fall short in accommodating maintaining individual privacy and data availability simultaneously. In this paper, a novel logarithmic spiral based data obfuscation method is proposed. A clustering-oriented data perturbation model is built to regulate the obfuscation process. The original dataset can be perturbed leveraging the logarithmic spiral, which can maintain k neighborhood relationship of the data set as well as its clustering utility effectively. Furthermore, for improving the strength of privacy protection, the paper proposes a multiple logarithmic spiral perturbation strategy. Theoretical analysis and experimental results demonstrate that this method can avoid leaking the data privacy meanwhile maintaining better clustering utility.

Keywords privacy-preserving; data mining; cluster analysis; logarithmic spiral; data perturbation

1 引 言

随着人们对数据发布中的隐私安全日益重视,数据挖掘中的隐私保护问题得到了人们的持续关注[1].如何在保护数据隐私的同时不影响数据的可用性已经成为信息安全与数据挖掘领域的一个重要研究方向[2-4].

基于数据失真的扰动是目前常用的一种隐私保护技术,其主要思想是通过对原始数据的修改实现

收稿日期:2012-06-05;最终修改稿收到日期:2012-08-27.本课题得到国家自然科学基金(61003057,60973023)资助.**黄茂峰**,男,1987 年生,硕士研究生,主要研究方向为数据隐私安全保护. E-mail: huangmaofeng@126.com. **倪巍伟**,男,1979 年生,博士,副教授,硕士生导师,主要研究领域为数据挖掘、数据隐私安全保护. **王佳俊**,男,1987 年生,硕士研究生,主要研究方向为数据隐私安全保护. **孙福林**,男,1988 年生,硕士研究生,主要研究方向为数据隐私安全保护. **崇志宏**,男,1969 年生,博士,副教授,主要研究领域为数据流和 WebDB.

对微数据(个体数据,区别于统计数据)隐私的保护,这种扰动容易造成数据个体差异的改变.聚类挖掘通过对个体数据的相似性和相异性的分析,将具有较低相异性和较高相异性的数据对象分别划分为同一聚簇和不同聚簇,聚类过程严重依赖于个体数据间的相异性[5-6].数据扰动与聚类挖掘在原理上存在弱化数据个体差异与依赖数据个体差异的冲突,导致面向聚类的数据隐藏变得尤为困难.

在面向聚类应用的数据扰动研究方面,Oliveira等人[7-8]提出了通过平移、缩放和旋转的数据转换方法;文献[9]提出的基于数据交换的扰动方法NeNDS通过交换空间距离最近的数据点实现对原始数据的保护;文献[10]提出的CAMP_CREST方法采用最小生成树方法实现扰动.CAMP-CREST将原始数据表中的每个数据看作树中的一个节点,并根据各个节点的欧氏距离生成最小生成树,在不改变敏感属性的条件下将一个节点 b 条邻近记录的准标识符用均值来替换.已有的面向聚类扰动方法多数存在难以兼顾隐私保护强度和聚类可用性的不足,例如RBT方法中任意两条原始数据记录与其发布后数据值的泄露将导致所有原始数据的泄露[11];NeNDS方法[8]也存在破坏数据原有的聚类特征使得扰动后的数据聚类结果发生较大偏差的问题;CAMP_CREST方法[9]中敏感属性泄露的概率则受到记录中的敏感属性值的种类数的约束.如何协调好数据隐私保护强度和聚类可用性成为面向聚类隐私保护数据发布研究的难点[12-15].

在聚类分析中,邻域关系是构成聚簇的基础,数据点间邻域关系可以用来衡量数据点的相似性.本文针对面向聚类应用的隐私保护数据发布问题,提出一种基于对数螺线的隐私保护数据扰动方法LSDP(Logarithmic Spiral based Data Perturbation),在保持数据点邻域关系基本不变的情况下,通过设置合适的对数螺线扰动参数,对原始数据进行扰动,实现对保护数据隐私和维持聚类可用性的兼顾.

本文主要贡献如下:

(1)引入对数螺线扰动的概念,构建了面向聚类的隐私保护数据扰动模型,使扰动后数据具有较好的聚类可用性和较高的隐私保护强度.

(2)为了提高隐私保护的强度,提出了多重扰动策略.该策略以线性增加额外的计算为代价,使得数据隐私保护的安全性得到指数级别的提高,同时不影响聚类的可用性.

本文第2节介绍相关概念及多维、多重对数螺线数据数据扰动方法;第3节介绍一种面向聚类的对数螺线数据扰动算法LSDP,重点对LSDP算法的聚类可用性、隐私保护安全性以及算法效率进行理论分析;第4节给出实验和数据分析;第5节总结全文并展望下一步的工作.

2 相关概念

对数螺线是一根无尽的螺线,其数学表达式为 $r=\alpha e^{\varepsilon\theta}$,其中,$\theta$ 是极角;r 是极径;e 是自然对数的底;α 和 ε 为常数,且 $\alpha>0$,$\varepsilon\neq 0$.对数螺线如图1所示.

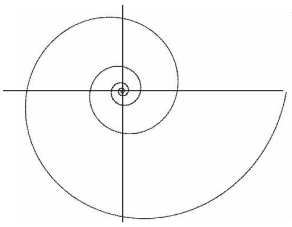

图1 对数螺线

考虑将对数螺线的几何性质应用于微数据隐藏发布,借助对数螺线对数据点进行扰动,隐藏原始数据.具体思路如下:通过对数螺线的旋转和缩放使数据点落于对数螺线上,再使数据点沿螺线方向在螺线上移动,从而对原数据进行扰动保护,将这种扰动方法称之为对数螺线扰动.

设原始数据点为 A,对数螺线扰动函数为 F,扰动后数据点为 A',$F\times A$ 表示为运用函数 F 对数据点 A 进行扰动,则对数螺线扰动可以表示为

$$A' \leftarrow F\times A.$$

在二维平面上,给定一条对数螺线,对于平面上任意一点,若该点落在对数螺线上,则使该点顺着螺线的方向在螺线上移动;若该点不在螺线上,使螺线绕其螺心旋转直至使该点落于螺线上,再使该点在螺线上沿螺线方向移动,将这种扰动方法称之为二维对数螺线扰动,见图2.

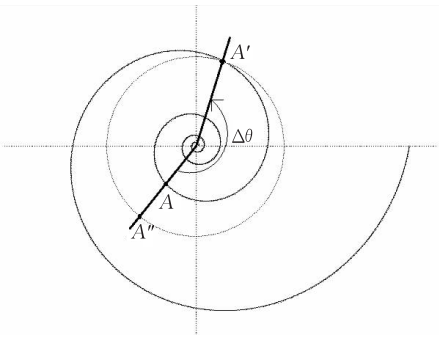

图2 二维对数螺线扰动

设点 $A(A_i,A_j)$ 为二维上的任意一点，设对数螺心为 $O(x,y)$，对数螺线方程为 $r=\alpha e^{\varepsilon\theta}$，旋转扰动角度为 $\Delta\theta$. 如图 2 所示，二维对数螺线扰动可分解为旋转和缩放两部分，点 A 先缩放至点 A''，再围绕螺心旋转至点 A'.

设 θ_A 为点 A 相对于对数螺线的极角，缩放参数 k 为扰动后的点 A' 与螺心的距离 $|OA'|$ 与扰动前的点 A 与螺心的距离 $|OA|$ 的比，则

$$k=\frac{|OA'|}{|OA|}=\frac{r_{A'}}{r_A}=\frac{\alpha e^{\varepsilon(\theta_A+\Delta\theta)}}{\alpha e^{\varepsilon\theta_A}}=e^{\varepsilon\Delta\theta}.$$

其中 r_A、$r_{A'}$ 分别为点 A、A' 的极径.

则扰动后点 $A'(A'_i,A'_j)$ 的坐标为

$$A'_i=k\cdot r_A\cdot\cos(\theta_A+\Delta\theta)+x,$$
$$A'_j=k\cdot r_A\cdot\sin(\theta_A+\Delta\theta)+y.$$

根据给定的参数，设二维数螺线扰动函数为 $F(k,\Delta\theta,O(x,y))$，则二维对数螺线扰动可以表示为

$$A'(A'_i,A'_j)\leftarrow F(k,\Delta\theta,O(x,y))\times A(A_i,A_j).$$

在三维空间中，给定螺心、螺轴向量和对数螺线方程，对于空间中任意一点，先根据螺心和螺轴向量确定该点所在的锥面，再根据对数螺线方程在锥面上确定该点所在的空间对数螺线，最后使该点在螺线上沿螺线方向移动，将这种扰动方法称之为三维对数螺线扰动.

数据点的扰动轨迹在任意一个以螺轴向量作为法向量的平面上的垂直投影都是一条对数螺线，且该对数螺线的螺心即是空间对数螺线的螺心在平面上的垂直投影，见图 3.

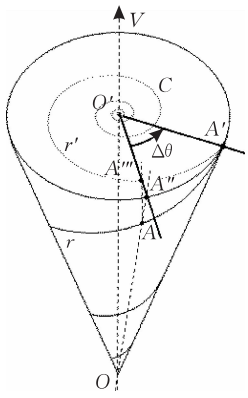

图 3　三维对数螺线扰动

设点 $A(A_i,A_j,A_r)$ 为三维空间中任意一点，螺心为 $O(x,y,z)$，螺轴向量为 $V(a,b,c)$，对数螺线方程为 $r=\alpha e^{\varepsilon\theta}$，旋转扰动角度为 $\Delta\theta$，根据所给参数确定点 A 所在的空间对数螺线为 r. 如图 3 所示，三维对数螺线扰动也可分解为缩放和旋转两部分，点 A 先缩放至点 A''，再围绕螺轴向量旋转扰动至点 A'. 设点 A' 在平面 C 上，且平面 C 垂直于螺轴向量 $\boldsymbol{V}$，其中点 O'、A''' 和对数螺线 r' 是点 O、A 和空间对数螺线 r 在平面 C 上的垂直投影.

设缩放参数 k 为扰动后的点 A' 与螺心的距离 $|OA'|$ 与扰动前的点 A 与螺心的距离 $|OA|$ 的比，则

$$k=\frac{|OA'|}{|OA|}=\frac{r_{A'}}{r_A}=e^{\varepsilon\Delta\theta}.$$

其中 r_A、r'_A 分别为点 A、A' 的极径.

最后由 $\Delta\theta$、k 以及螺线参数求出点 $A'(A'_i,A'_j,A'_r)$.

根据给定的参数，设三维数螺线扰动函数为 $F(k,\Delta\theta,O(x,y,z),V(a,b,c))$，则三维对数螺线扰动可以表示为

$$A'(A''_i,A'_j,A'_r)\leftarrow F(k,\Delta\theta,O(x,y,z),V(a,b,c))\times A(A_i,A_j,A_r).$$

将一个 $n(n>1)$ 维数据集 D 随机划分为一组不相交的二维、三维投影子集，二维、三维投影子集数分别为 i 和 j，其中 $n=2i+3j(i,j\geq 0,n>1)$. 然后对二维投影子集进行二维对数螺线扰动，对三维投影子集进行三维对数螺线扰动，令所有扰动的缩放参数 k 相同，将这种扰动称之为多维对数螺线扰动.

设二维投影子集为

$$(A_{b1},A_{b2})\cdots(A_{b\,2i-1},A_{b\,2i}).$$

设三维投影子集为

$$(A_{t1},A_{t2},A_{t3})\cdots(A_{t\,3j-2},A_{t\,3j-1},A_{t\,3j}).$$

然后对各投影子集进行相应的对数螺线扰动

$$A'(A'_{b1},A'_{b2})\leftarrow F(k,\Delta\theta_{b1},O(x_{b1},y_{b1}))\times A(A_{b1},A_{b2}).$$
$$\vdots$$
$$A'(A'_{b\,2i-1},A'_{b\,2i})\leftarrow F(k,\Delta\theta_{bi},O(x_{bi},y_{bi}))\times A(A_{b\,2i-1},A_{b\,2i}).$$
$$A'(A'_{t1},A'_{t2},A'_{t3})\leftarrow F(k,\Delta\theta_{t1},O(x_{t1},y_{t1},z_{t1}),V(a_{t1},b_{t1},c_{t1}))\times A(A_{t1},A_{t2},A_{t3}).$$
$$\vdots$$
$$A'(A'_{t\,3j-2},A'_{t\,3j-1},A'_{t\,3j})\leftarrow F(k,\Delta\theta_{tj},O(x_{tj},y_{tj},z_{tj}),V(a_{tj},b_{tj},c_{tj}))\times A(A_{t\,3j-2},A_{t\,3j-1},A_{t\,3j}).$$

在对数螺线扰动后，把扰动后的数据集替换扰动前的数据集. 有时为了增加数据隐藏的安全性，会对数据集进行多次的对数螺线扰动，称之为多重对数螺线扰动.

3　LSDP 算法

在聚类分析中，邻域关系是构成聚簇的基础，如

果扰动前后的数据具有相似的邻域关系,那么扰动前后的数据一定能够能够具有相似的聚类可用性.本文针对面向聚类应用的隐私保护数据发布问题,提出一种基于对数螺线的隐私保护数据扰动方法LSDP,在保持数据点邻域关系基本不变的情况下,通过设置适合的对数螺线扰动参数,对原始数据集进行干扰,得到扰动后的数据集,实现对保护数据隐私和维持聚类可用性的兼顾.

3.1 算法描述

LSDP 算法流程如下:

将多维数据集 D 中的多维属性随机划分成一组不相交的二维、三维投影子集,给定缩放参数 k.对于二维投影子集,给定螺心 $O(x,y)$ 和扰动角度 $\Delta\theta$,进行二维对数螺线扰动;对于三维投影子集,给定螺心 $O(x,y,z)$、螺轴向量 $\boldsymbol{V}(a,b,c)$ 和扰动角度 $\Delta\theta$,进行三维对数螺线扰动,最后将原数据集 D 替换为扰动后的数据集 D'.对数据集 D 进行 t 次这样的对数螺线扰动,最终得到 t 重对数螺线扰动后的数据集 $D^{(t)}$.

算法 1. LSDP.

输入:原始数据集 D,扰动重数 t,各重扰动缩放参数 $k_1 \sim k_t$

输出:扰动后的数据集 $D^{(t)}$

$D^{(0)} \leftarrow D$ /*用原始数据 D 初始化扰动数据集 $D^{(0)}$*/

For each perturbation (No. of perturbation is e: $1 \sim t$) do /*对数据 $D^{(e-1)}$ 进行扰动*/

$Pairs \leftarrow (A_i, A_j)$ or (A_i, A_j, A_r) in $D^{(e-1)}$ $(1 \le i, j, r \le n; i \ne j, j \ne r, i \ne r)$; /*划分投影子集*/

For each selected pair form $Pairs$ do /*对不同的投影子集运用不同的扰动方法*/

If the selected pair have two attributes: $A(A_i, A_j)$

$F(k_e, \Delta\theta_{ei}, O_{ei}(x,y)) \leftarrow [k_e, \Delta\theta_{ei}, O_{ei}(x,y)]$; /*确定二维对数螺线扰动的参数*/

$A'(A_i', A_j') \leftarrow F(k_e, \Delta\theta_{ei}, O_{ei}(x,y)) \times A(A_i, A_j)$; /*二维对数螺线扰动*/

If the selected pair have three attributes: $A(A_i, A_j, A_r)$

$F(k_e, \Delta\theta_{ei}, O_{ei}(x,y,z), V_{ei}(a,b,c)) \leftarrow [k_e, \Delta\theta_{ei}, O_{ei}(x,y,z), V_{ei}(a,b,c)]$; /*确定三维对数螺线扰动的参数*/

$A'(A_i', A_j', A_r') \leftarrow F(k_e, \Delta\theta_{ei}, O_{ei}(x,y,z), V_{ei}(a,b,c)) \times A(A_i, A_j, A_r)$; /*三维对数螺线扰动*/

End for

Combine the perturbed pairs into $D^{(e)}$; /*扰动后数据集 $D^{(e)}$*/

End for

Return $D^{(t)}$.

3.2 算法分析

LSDP 方法通过对数螺线扰动不仅可以保护数据的隐私,同时在扰动前后较好地维持了原始数据邻域关系的稳定,保持了原有数据的聚类效果,确保了扰动后数据有较好的聚类可用性.本节对隐藏算法 LSDP 的聚类可用性、隐私保护安全性以及算法效率进行理论分析.

定理 1. 对多维数据集进行对数螺线扰动,不会改变数据集的数据邻域关系.

证明. 设 A、B 为二维数据集上的两数据点,对其进行缩放参数为 k 和旋转角度为 $\Delta\theta$ 的对数螺线扰动,A'、B' 为扰动后的点,见图 4.

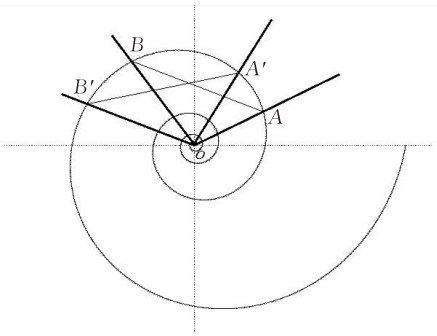

图 4 对 A、B 两点进行对数螺线扰动

因为 A、B 两点相对于螺心 O 的扰动角度相同,所以两点在扰动前后与螺心形成的夹角角度相同,即 $\angle AOB = \angle A'OB'$,又因为 $|OA'| = k|OA|$,$|OB'| = k|OB|$,根据相似三角形原理可推出 $|A'B'| = k|AB|$,数据点 A、B 之间的距离缩放了 k 倍.

由上可知,对二维数据集上的所有数据点应用参数相同的对数螺线扰动后,任意两个数据点之间的距离都缩放了 k 倍,所以二维数据集的数据邻域关系在扰动前后没有发生改变,数据集在扰动前后具有相似的聚类可用性.

同理可知,在三维对数螺线扰动中,任意两点的距离都缩放了 k 倍,所以三维数据集的数据邻域关系在扰动前后也没有发生改变,数据集在扰动前后具有相似的聚类可用性.

在二维和三维数据集中,对数螺线扰动都可以保证较好的聚类质量,同样也可以拓展到多维中.

将一个 $n(n>1)$ 维数据集 D 随机划分为一组不相交的二维、三维投影子集,然后对所有投影子集进行具有相同缩放参数的对数螺线扰动,二维或三维投影子集的任意两个数据点属性值之间的欧氏距离在扰动后都缩放了 k 倍.对所有投影子集进行对数螺线扰动后,原数据集中任意两数据点之间的欧氏距离都将缩放 k 倍,所以多维数据集中的数据邻域关系在对数螺线扰动前后没有发生改变,数据集在

扰动前后具有相似的聚类可用性. 证毕.

在此基础上，进一步考虑多重对数螺线扰动.

定理 2. 对数据集进行多重对数螺线扰动，不会改变数据集的数据邻域关系.

证明. 设原始数据集由 m 条具有 n 维属性的记录组成，由性质 1 可知，数据集的数据邻域关系在对数螺线扰动前后不发生改变，即对数螺线扰动后的数据集 $D'_{m\times n}$ 与原始数据集 $D_{m\times n}$ 中的数据邻域关系相同，那么对扰动后的数据集再进行对数螺线扰动得到的数据集 $D''_{m\times n}$ 与数据集 $D'_{m\times n}$ 中的数据邻域关系也相同，则数据 $D''_{m\times n}$ 与原始数据集 $D_{m\times n}$ 中的数据邻域关系也是相同的. 设对数据集 $D_{m\times n}$ 进行 t 重对数螺线扰动后得到数据集 $D^{(t)}_{m\times n}$，则数据集 $D^{(t)}_{m\times n}$ 与原始数据集 $D_{m\times n}$ 具有相同的数据邻域关系，即多重对数螺线扰动没有改变数据集的数据邻域关系. 证毕.

由定理 2 知，多重对数螺线扰动在保护数据隐私的同时还可以维持数据集的数据邻域关系.

定理 3. 多重对数螺线扰动可以增强隐私保护的安全性.

证明. 假设 $n(n\%2=0, n>1)$ 维数据集划分的投影子集全部是二维的，且各重投影子集的划分是独立并且随机的，即第 i 重投影子集 (A_{xi}, A_{yi}) 与第 j 重投影子集 (A_{xj}, A_{yj}) 相互独立. 每重扰动都是对 $n/2$ 个二维投影子集进行对数螺线扰动，且在每重的所有扰动函数 $F(k, \Delta\theta, O(x,y))$ 中，缩放参数 k 相同，且扰动角度 $\Delta\theta$ 和螺心 $O(x,y)$ 的取值是随机且独立的. 因为各重投影子集的划分相互独立，所以各重间的扰动函数参数也是相互独立的.

假设攻击者获取了部分原始数据，进行逆推猜测各重投影子集的划分和所有扰动函数的参数. 由于各重划分的投影子集都是二维的，因此在每重扰动中有 $n/2$ 个扰动函数 $F(k, \Delta\theta, O(x,y))$，所以 t 重扰动中的扰动函数共有 $tn/2$ 个. 根据排列组合理论，每重扰动的二维投影子集划分方式共有 $n!/2!^{\frac{n}{2}}$ 种，t 重的二维投影子集划分方式共有 $(n!/2!^{\frac{n}{2}})^t$ 种，则在无法获知扰动重数的前提下，攻击者进行逆推猜测可能需要计算所有的扰动函数 $F(k, \Delta\theta, O(x,y))$，又因为每个扰动函数可以在常数时间内解出，所以逆推猜测攻击的计算复杂度为 $F1$：

$$F1=O\left(\frac{n}{2}(n!/2^{\frac{n}{2}})+\frac{2n}{2}(n!/2^{\frac{n}{2}})^2+\cdots+\frac{tn}{2}(n!/2^{\frac{n}{2}})^t\right)$$
$$=O(nt(n!)^t\cdot\sqrt{2}^{-tn}).$$

同理，假设 $n(n\%3=0, n>2)$ 维数据集划分的投影子集全部是三维的，且各重投影子集的划分是独立并且随机的，则 t 重对数螺线扰动共有 $tn/3$ 个扰动函数 $F(k, \Delta\theta, O(x,y,z), V(a,b,c))$，且 t 重投影子集划分方式共有 $(n!/3!^{\frac{n}{3}})^t$ 种，所以逆推猜测攻击的计算复杂度为 $F2$：

$$F2=O\left(\frac{n}{3}(n!/6^{\frac{n}{3}})+\frac{2n}{3}(n!/6^{\frac{n}{3}})^2+\cdots+\frac{tn}{3}(n!/6^{\frac{n}{3}})^t\right)$$
$$=O(nt(n!)^t\cdot\sqrt[3]{6}^{-tn}).$$

现在，$n(n>1)$ 维数据集随机划分为不相交的 a 个二维和 b 个三维投影子集 $(a\geq 0, b\geq 0, 2a+3b=n)$，所以每重投影子集的划分方式共有 $\sum_{\substack{a\geq 0, b\geq 0\\ 2a+3b=n}} \frac{n!}{2!^a\cdot 3!^b}=O(n!/2^{\frac{n}{2}})$ 种.

$a\geq 0, b\geq 0, 2a+3b=n$ 有

$n/3\leq a+b\leq n/2, O(a+b)=O(n/2)$.

所以逆推猜测攻击的计算复杂度为 $F3$：

$$F3=O\left(\frac{n}{2}(n!/2^{\frac{n}{2}})+\frac{2n}{2}(n!/2^{\frac{n}{2}})^2+\cdots+\frac{tn}{2}(n!/2^{\frac{n}{2}})^t\right)$$
$$=O(nt(n!)^t\cdot\sqrt{2}^{-tn}).$$

因为攻击者无法获知二维和三维的投影子集数，所以 $F3\geq F1+F2$.

所以多重对数螺线扰动有效增强了隐私保护的安全性. 证毕.

由性质 3 知，多重扰动策略虽然使算法的计算复杂度略微增加，但是数据隐私的安全性却能获得指数级别的提高.

采用 LSDP 方法，缩放参数 k 和扰动角度 $\Delta\theta$ 决定了数据集的扰动强度，而 t 增强了数据扰动的安全性. 各重投影子集的划分相互独立，并且扰动角度 $\Delta\theta$、螺心和螺轴向量等参数的取值是独立且随机设定的，使得攻击者进行逆推猜测难以实行. 假设 k、$\Delta\theta$ 和 ε 的值都被攻击者获知，由于每重扰动的二、三维投影子集是随机划分的，所以要逆推猜测可能要遍历所有的划分组合，由性质 3 可知，在 t 取值合适的情况下，这种攻击付出的代价是巨大的.

与 RBT[8] 方法不同的是，LSDP 方法可以有效扩展到多维多重，即使泄露了部分原始数据或者部分扰动参数，也无法逆推出完全的原始数据，而且在 LSDP 方法中，相互独立而且随机设定的扰动角度 $\Delta\theta$、螺心和螺轴向量等参数只影响到扰动隐藏的属性值而没有影响其原有的邻域关系. LSDP 方法通过保证每重扰动时各投影子集具有相同的缩放参数，实现了对原始数据进行有效隐藏的同时能够很好地维持数据集的数据邻域关系.

4 实验与分析

为了对 LSDP 方法的聚类可用性和执行效率进行分析，本节将通过具体的实验来加以验证和说明. 实验环境为 Inter(R) Core(TM)i5 CPU 2.30GHz，2GB 内存，Windows 7 操作系统. 所涉及代码用 Visual C++(6.0)实现. 实验数据均采用实际数据，来自 http://www.ics.uci.edu/~mlearn. 第 1 个数据集：Breast Cancer Wisconsin (Diagnostic)，该数据集共有 569 个数据记录，每条记录有 32 个属性值；第 2 个数据集：Image Segmentation，该数据集共有 2100 个数据记录，每条记录有 19 个属性. 两数据集分别取其中的 8 个和 9 个属性，且这些属性值均为数值类型. 实验前需对实验数据进行适当的预处理，针对上述两个数据集的属性值，需要对其进行规格化处理，使得所有属性值的取值范围为 $[0,10]$.

对比算法的聚类可用性时，对实验数据集，分别应用 LSDP、NeNDS[9]、CAMP_CREST[10] 以及 RBT[8] 算法进行扰动，然后对原始数据集和扰动后的数据集分别用 k-means[16] 算法和 DBSCAN[17] 算法进行聚类，对比分析原始数据集和扰动后数据集的聚类质量，验证 LSDP 算法的聚类可用性.

本文采用文献[16]介绍的 F-measure 方法对比扰动算法的聚类可用性，C' 表示扰动后数据集生成的聚簇集合，其中 C_i、C'_j 分别表示 C、C' 中任意聚簇，$n_{ij}=|C_i \cap C'_i|$，则有

$$Recall(C_i,C'_j)=\frac{n_{ij}}{|C_1|},\ Precision(C_i,C'_j)=\frac{n_{ij}}{|C'_j|},$$

$$F(C_i,C'_j)=\frac{2\times Recall(C_i,C'_j)\times Precision(C_i,C'_j)}{Recall(C_i,C'_j)+Precision(C_i,C'_j)},$$

$$F(C')=\sum_{C_i\in C}\frac{C_i}{D}\max_{C_j\in C}\{F(C_i,C'_j)\}.$$

$F(C')$ 值越大，表明扰动前后数据集的聚类结果越相似，扰动算法的可用性越好；相反，则表明扰动算法的可用性越差.

实验中，在第 1 个数据集上，LSDP 算法中参数 $t,k,\Delta\theta$ 分别取 7，$[0.1,5]$，$[0.01\pi,0.5\pi]$，NeNDS 算法中参数 c 取 6，CAMP_CREST 算法中参数 b 取 5；在第 2 个数据集上，LSDP 算法中参数 $t,k,\Delta\theta$ 分别取 4，$[1,10]$，$[0.1\pi,\pi]$，NeNDS 算法中参数 c 取 4，CAMP_CREST 算法中参数 b 取 9. 图 5，6 显示了在两个实验数据集在扰动前后应用 k-means 算法进行聚类后得到的 F-measure 值与 k 的对应关系.

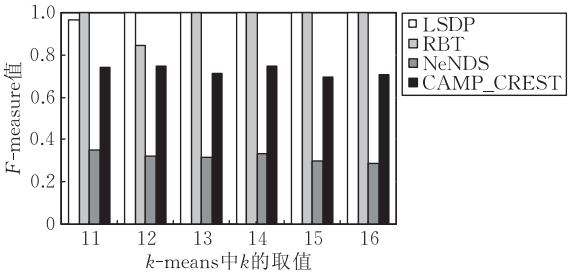

图 5 Breast Cancer Wisconsin 扰动前后 k-means 聚类质量对比

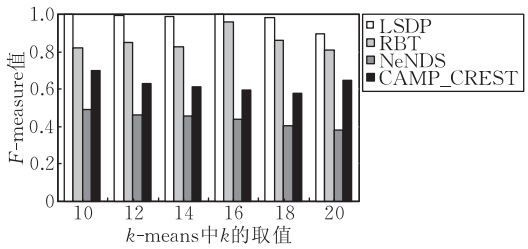

图 6 Image Segmentation 扰动前后 k-means 聚类质量对比

由图 5、图 6 可知，LSDP 算法与 RBT 算法具有相近的 F-measure 值，对比 NeNDS 和 CAMP_CREST 算法则具有更高的 F-measure 值. 因此扰动后的数据聚类结果与原始数据聚类结果比较相似，具有较好的可用性.

同时，表 1 给出了在两个原始数据集和扰动后的数据集上应用 DBSCAN 进行聚类后得到的 F-measure 值，在 Breast Cancer Wisconsin 数据集中 DBSCAN 算法参数 $epsilon=2.5$，$Minpts=20$，在 Image Segmentation 数据集中 DBSCAN 算法参数 $epsilon=1.5$，$Minpts=8$.

表 1 DBSCAN 聚类质量对比

数据集合	聚类后的 F-measure 值			
	LSDP	RBT	NeNDS	CAMP_CREST
Breat Cancer Wisconsin	0.91916	0.91916	0.77683	0.89803
Image Segmentation	0.91810	0.89407	0.76080	0.75579

由表 1 可知，LSDP 算法与 RBT 算法具有相近的 F-measure 值，略大于 NeNDS 和 CAMP_CREST 算法对应的 F-measure 值.

对扰动前后数据集应用 k-means 和 DBSCAN 聚类算法的结果分析表明，LSDP 扰动算法较好地维持了数据的聚类可用性.

对比算法的执行效率时，应用 LSDP、NeNDS、CAMP_CREST 以及 RBT 算法对 3～4 个不同大小的实验数据集进行扰动，对比算法执行所用的时间，验证 LSDP 算法的执行效率.

实验中，为了增强执行效率的可比性，LSDP 算法采用单重扰动. 实验分为两组：第 1 组实验对比算法在不同规模数据集上的执行效率，随机构造分别包含 1000、2000、4000 和 8000 条记录的 4 个 10 维浮点数据集；第 2 组实验对比算法在具有不同维度数据集上的执行效率，随机构造维度分别为 10、20、40 和 80 的浮点数据集，且数据集的记录数都设置为 1000. 为了增加数据的可比性，两组试验中，4 种算法的实现采用了相同的数据结构和运行环境.

试验中，LSDP 算法参数 t，k，$\Delta\theta$ 分别取 1，[0.1,10]，[0.01π,0.5π]，NeNDS 算法参数 c 取 6，CAMP_CREST 算法参数 b 取 8. 表 2、3 给出了 4 种扰动算法在不同规模的数据集上的执行时间(单位为 ms).

表 2　算法在不同记录规模数据集上的执行时间

记录数	执行时间/s			
	LSDP	RBT	NeNDS	CAMP_CREST
1000	16	15	32	295
2000	31	30	51	1170
4000	62	61	109	4560
8000	125	122	250	18601

表 3　算法在不同维度数据集上的执行时间

属性数	执行时间/s			
	LSDP	RBT	NeNDS	CAMP_CREST
10	16	15	32	295
20	31	31	63	317
40	64	63	125	326
80	141	132	249	328

由表 2、3 可知，LSDP 算法与 RBT 算法具有相近的执行效率，NeNDS 算法的执行效率比 LSDP 算法低近似一倍，且都与数据集的记录数和属性维度保持着近似线性关系. CAMP_CREST 算法的执行效率最低，且随着记录数的线性增长呈现指数级的增长，但 CAMP_CREST 算法的执行效率与数据集的维度呈弱相关性.

实验结果表明，LSDP 扰动算法具有较好的执行效率，在实际应用中具有较高的可行性.

5　结论与下一步工作

本文介绍了面向聚类的数据隐私保护的相关问题，提出一种基于对数螺线的隐私保护数据扰动方法 LSDP. LSDP 方法利用对数螺线的几何性质，对数据点的属性值进行扰动，在较好地保护数据隐私的同时维持了原始数据的 k 邻域关系. 理论分析和实验结果表明，LSDP 方法能够有效兼顾数据隐私的保护和数据聚类可用性的维持.

需要指出的是，LSDP 方法在对低维度数据点进行扰动时，投影子集划分的选择范围较小，导致隐私保护强度不高，下一步将对提高 LSDP 方法的安全性进行进一步研究.

参　考　文　献

[1] Agrawal R, Srikant R. Privacy preserving data mining//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Dallas, Texas, USA, 2000: 439-450

[2] Luo Yong-Long, Huang Liu-Sheng, Jing Wei-Wei, Yao Yi-Fei, Chen Guo-Liang. An algorithm for privacy-preserving Boolean association rule mining. Acta Electronica Sinaca, 2005, 33(5): 900-903(in Chinese)
(罗永龙，黄刘生，禁巍巍，姚亦飞，陈国良. 一个保护私有信息的布尔关联规则挖掘算法. 电子学报，2005，33(5)：900-903)

[3] Ge Wei-Ping, Wang Wei, Zhou Hao-Feng, Shi Bo-Le. Privacy preserving classification mining. Journal of Computer Research and Development, 2006, 43(1): 39-45(in Chinese)
(葛伟平，汪卫，周皓峰，施伯乐. 基于隐私保护的分类挖掘. 计算机研究与发展，2006，43(1)：39-45)

[4] Zhang Peng, Tong Yun-Hai, Tang Shi-Wei, Yang Dong-Qing, Ma Xiu-Li. An effective method for privacy preserving association rule mining. Journal of Software, 2006, 17(8): 1764-1774(in Chinese)
(张鹏，童云海，唐世渭，杨冬青，马秀莉. 一种有效的隐私保护关联规则挖掘方法. 软件学报，2006，17(8)：1764-1774)

[5] Zhou Shui-Geng, Li Feng, Tao Yu-Fei, Xiao Xiao-Kui. Privacy preservation in database applications: a survey. Chinese Journal Of Computers, 2009, 32(5): 847-861(in Chinese)
(周水庚，李丰，陶宇飞，肖小奎. 面向数据库应用的隐私保护研究综述. 计算机学报，2009，32(5)：847-861)

[6] Rajalaxmi R R, Natarajan A M. An effective data transformation approach for privacy. Journal of Computer Science, 2008, 4(4): 320-326

[7] Oliveira S R M, Zaïane O R. Achieving privacy preservation when sharing data for clustering//Proceedings of the International Workshop on Secure Data Management in a Connected World. Toronto, Canada, 2004: 67-82

[8] Oliveira S R M, Zaïane O R. Privacy preserving clustering by data transformation//Proceedings of the 18th Brazilian Symposium on Database. Manaus, Brazil, 2003: 304-318

[9] Parameswaran R, Blough D M. Privacy preserving data obfuscation for inherently clustered data. International Journal of Information and Computer Security, 2008, 2(1): 4-26

[10] Li Xiao-Bai, Sarkar Sumit. Data clustering and micro-perturbation for privacy-preserving data sharing and analysis//Proceedings of the ICIS 2010. Yamagata, Japan, 2010: 58

[11] Ni Wei-Wei, Chen Geng, Chong Zhi-Hong, Wu Ying-Jie.

Privary-presering data publication for clustering. Journal of Computer Research and Development, 2012, 49(5): 1095-1104(in Chinese)
(倪巍伟，陈耿，崇志宏，吴英杰. 面向聚类的数据隐藏发布研究. 计算机研究与发展，2012，49(5)：1095-1104)

[12] Liu L, Kantarcioglu M, Thuraisingham B. The applicability of the perturbation based privacy preserving data mining for real-world data. Data & Knowledge Engineering, 2008, 65(1): 5-21

[13] Li Feng, Ma Jin, Li Jian-Hua. Distributed anonymous data perturbation method for privacy-preserving data mining. Journal of Zhejiang University (Science A), 2009, 10(7): 952-963

[14] Yang Xiao-Chun, Wang Ya-Zhe, Wang Bin, Yu Ge. Privacy preserving approaches for multiple sensitive attributes in data publishing. Chinese Journal of Computers, 2008, 31(4): 574-587(in Chinese)
(杨晓春，王雅哲，王斌，于戈. 数据发布中面向多敏感属性的隐私保护方法. 计算机学报，2008，31(4)：574-587)

[15] Aggarwal G, Feder T, Kenthapadi K et al. Achieving anonymity via clustering//Proceedings of the ACM SIGMOD/PODS 2006. Chicago, Illinois, USA, 2006: 153-162

[16] Fung Benjamin C M, Wang Ke, Wang Lingyu et al. Privacy-preserving data publishing for cluster analysis. Data & Knowledge Engineering, 2009, 68(6): 552-575

[17] Ester M, Kriegel HP, Sander J et al. A density-based algorithm for discovering clusters in large spatial databases with noise//Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Oregon, USA, 1996: 226-231

HUANG Mao-Feng, born in 1987, M. S. candidate. His research interest is privacy preserving data application.

NI Wei-Wei, born in 1979, Ph. D., associate professor. His research interests include data mining and privacy preserving data application.

WANG Jia-Jun, born in 1987, M. S. candidate. His research interest is privacy preserving data application.

SUN Fu-Lin, born in 1988, M. S. candidate. His research interest is privacy preserving data application.

CHONG Zhi-Hong, born in 1969, Ph. D., associate professor. His current research interests include data streams and WebDB & P2P.

Background

Privacy preserving micro-data publishing has become a hot research issue, which requires seeking tradeoffs among maintaining data utility and preserving privacy of the dataset. In recent years, much of existed privacy preserving research is focused on association rule mining and classifying. Privacy preserving data obfuscation for clustering application isn't mentioned too much and still has many problems unresolved. Data perturbation is an effective technology for obfuscation, which perturbs the original dataset leveraging data transformation and generalization. Concerning privacy-preserving clustering, a novel logarithmic spiral based data obfuscation method LSDP (Logarithmic Spiral based Data Perturbation) is proposed in the paper. LSDP perturbs the original dataset exploiting geometrical properties of logarithmic spiral. Furthermore, LSDP maintains k neighborhood relationship of the original data while preserving privacy of the dataset. Further, this paper proposes a very effective privacy protection model to prevent attacks on the basis of research on LSDP.

Our work is supported by the Natural Science Foundation of China (61003057) with title Research of Clustering Information Nuggets Maintaining and Data Obfuscation for High Dimensional Dataset. Clustering heavily depends on the characteristics of individual records to segment dissimilar records into different clusters. On the contrary, it is a prevailing idea of obfuscation to suppress the individual characteristics for sake of avoiding leakage of individual privacy. Our project focuses on seeking tradeoff between clustering utility maintaining and individual privacy protection. Under this foundation, our team has published some high level article, in other hand, based on our results of theoretical studies, we have independently constructed a set of clustering-oriented privacy preserving models and proposed many of active data obfuscation algorithms for high dimensional dataset.