第 35 卷 第 2 期 计 算 机 工 程 2009 年 1 月
Vol.35 No.2 Computer Engineering January 2009
·安全技术· 文章编号：1000—3428(2009)02—0157—02 文献标识码：A 中图分类号：TP309

基于离散余弦变换矩阵的隐私数据保护方法

张国荣[1]，印 鉴[2]
(1. 广州美术学院计算机基础教研室，广州 510260；2. 中山大学信息科学与技术学院，广州 510275)

摘 要：针对聚类分析中隐私数据保护的问题，提出一种基于离散余弦变换矩阵的隐私数据保护方法(DCBT)，对随机选择的 k 个属性向量实施变换，直到所有属性都至少被变换一次且变换的次数达到初始设置值，选取隐私保护度最优的变换结果。实验结果表明，对于集中式数据，该方法能保持 2 点间距离不变，使数据较好地实现扭曲，保护隐私信息，对聚类结果基本没有影响。

关键词：数据挖掘；隐私保护；聚类分析；离散余弦变换

Privacy Data Preserving Method Based on Discrete Cosine Transform Martrix

ZHANG Guo-rong[1], YIN Jian[2]
(1. Computer Staff Room, Guangzhou Academy of Fine Arts, Guangzhou 510260;
2. School of Information Science & Technology, Sun Yat-sen University, Guangzhou 510275)

【Abstract】 Privacy preserving is an important direction for data mining research. This paper concentrates on the issue of protecting the underlying attribute values when sharing data for clustering and proposes a method called Discrete Cosine-Based Transformation(DCBT), random selects the k attributes and then distorts them with discrete cosine transformation. In the process of transformation, the goal is to find the proper chain of transform to satisfy the optimum privacy preserving requirement. For the centralized data, the experiments demonstrate that the method efficiently distorts attribute values, preserves privacy information and guarantees valid clustering results.

【Key words】 data mining; privacy preserving; clustering analysis; discrete cosine transform

1 概述

数据挖掘中的隐私保护主要考虑 2 个方面：敏感的原始数据和从数据库中提取的敏感知识。隐私保护挖掘的主要目的就是用某种技术改进已有的数据挖掘算法修改原始数据，使敏感数据和知识不被泄露[1]。聚类分析作为数据挖掘中的一个具有很强挑战性的领域，可作为一个单独的工具用来发现数据库中数据分布的一些深层的信息，也可作为数据挖掘算法中其他分析算法的一个预处理步骤。

对于聚类分析中集中分布的数据，常采用数据变换伪装的方法，通过变换初始数据，确保从变换后的数据中不能推算出初始数据，从而达到扰乱数据，隐藏私有信息的目的，同时变换后的数据对聚类的结果影响不大[2-3]。几何数据转换方法[2](GDTM)主要通过几何转换如平移、缩放和简单的旋转等转化初始数据达到聚类隐私保护的目的。但该方法容易改变数据的相似性，造成聚类的误差，对空间的维数也有限制；基于旋转的转换方法[3](RBT)在变换之前对数据进行规范化处理，使所有属性有相同的权重，同时标识实体的属性(如ID)被匿名保护，然后利用旋转变换对数据进行扭曲。在变换的过程中，由于随机选择了属性对，在要求的相对隐私保护度范围内随机选择变换的角度，因此很难从变换后的数据估计出原始数据，从而使原始数据的隐私得到较好的保护。但该方法在用户提出过高隐私要求时可能无法取得合适的旋转角度，并且该方法每次只能变换 2 个属性向量，也未讨论所有等距变换。

对于分布式存放的数据，可通过利用安全多方计算的相关方法，确保参与的各方只获得结果而不知道其他方输入的任何信息。文献[4]提出应用于垂直分布类型的聚类分析的方法，使每个站点在得到聚类结果时，不会泄露自己站点内部的事务属性值。文献[5]针对水平分布类型的聚类分析方法，提出在聚类挖掘的过程中使用基于健忘多项式计算和同态加密的协议，从而达到私有信息不被泄露的目的。

本文针对数据集中式分布的聚类分析隐私保护问题，在RBT[3]的基础上，提出一种利用离散余弦变换矩阵实现初始数据隐私保护的方法——DCBT。

2 基本概念

2.1 离散余弦变换矩阵

离散余弦变换(DCT)是数字图像处理等许多领域的重要数学工具，本文利用 DCT 作为预处理过程，对随机选取的属性向量进行变换，使数据在保持相似性的同时实现扭曲，确保从变换后的数据中不能推算出初始数据，从而达到保护私有信息的目的。DCT 变换矩阵为

基金项目：国家自然科学基金资助项目(60573097, 60773198)；广东省自然科学基金资助项目(05200302, 06104916)；广州市科技计划基金资助项目(2007Z3-D3071)；高等学校博士学科点专项科研基金资助项目(20050558017)；新世纪优秀人才支持计划基金资助项目(NCET-06-0727)

作者简介：张国荣(1977－)，男，讲师、硕士，主研方向：数据挖掘；印 鉴，教授、博士

收稿日期：2008-02-20 **E-mail：**chzhzgr@163.com

$$\boldsymbol{C}_N=\frac{\sqrt{2}}{\sqrt{N}}\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdots & \frac{1}{\sqrt{2}}\\ \cos\frac{\pi}{2N} & \cos\frac{3\pi}{2N} & \cdots & \cos\frac{(2N-1)\pi}{2N}\\ \cos\frac{2\pi}{2N} & \cos\frac{6\pi}{2N} & \cdots & \cos\frac{2(2N-1)\pi}{2N}\\ \vdots & \vdots & & \vdots\\ \cos\frac{(N-1)\pi}{2N} & \cos\frac{3(N-1)\pi}{2N} & \cdots & \cos\frac{(N-1)(2N-1)\pi}{2N}\end{bmatrix}$$

易验证 $\boldsymbol{C}_N\boldsymbol{C}_N^{\mathrm{T}}=I_N$，即$\boldsymbol{C}_N$是正交变换，对任意 2 点进行变换不改变 2 点之间的距离。

2.2 隐私保护度

本文借助数据库中评估数据安全的方法对隐私保护度进行评估。在实施变换前，数据用 z-score 方法进行规范化，因此，可利用如下公式评估隐私保护程度：

$$P=Var(X-X')$$

其中，X 表示扰乱前的属性值；X'表示扰乱后的属性值，且 $Var(X)=Var(x_1,x_2,\cdots,x_N)=\frac{1}{N}\times\sum_{i=1}^{N}(x_i-\bar{x})^2$（$\bar{x}$ 是 $x_1,x_2,\cdots,x_N$ 的平均值)。

在实施 DCBT 过程中，还须使用如下概念：

(1)最小隐私保护度：在实施一次变换后，p 称为本次变换的最小隐私保护度：

p=min($Var(X_1$-X_1'), $Var(X_2$-X_2'),⋯,$Var(X_m$-X_m'))

(2)最优隐私保护度：在实施 k 次变换后，$p_{\max}$ 称为 k 次变换的最优隐私保护度：

$$p_{\max}=\max(p_1,p_2,\cdots,p_k)$$

3 DCBT 方法

假设：(1)数据矩阵只包含在聚类之前必须进行转换保护的个人数值型数据；(2)存在可以唯一标识实体的属性(例如 ID)，这个属性联系着每一个实体并且已经被匿名保护；(3)数据矩阵已被 z-score 方法规范化。

3.1 方法描述

DCBT 主要对随机组合的属性向量对进行 DCT 变换，对每一个属性向量的组合，先选择相应 DCT 变换进行变换，直到最小隐私保护度不为 0，变换的次数达到初始设置值。最后选取隐私保护度最大的变换结果。

在进行变换时的关键步骤如下：

(1)随机选择属性向量组，随机产生属性向量组的数目 k，从数据矩阵 $\boldsymbol{D}_{m\times n}$ 中随机选取 k 个属性向量组成 1 组。

(2)变换属性向量组，选择 k 维离散余弦变换矩阵变换属性向量组，变换完成后，属性向量组必须返回数据矩阵 $\boldsymbol{D}_{m\times n}$ 替换原来的属性向量

(3)评估隐私保护程度，对每一个属性向量计算隐私保护度，选择最小的隐私保护度作为此次变换的最小隐私保护度，然后与以前变换的最小隐私保护度比较，较大的为最优隐私保护度。计算隐私保护度时，属性向量比较的是最原始的数据矩阵 $\boldsymbol{D}_{m\times n}$，而不是每次变换之前的数据矩阵。如果最小隐私保护度为 0，则表明有某些属性向量从来没有进行变换。

3.2 具体算法

输入 $\boldsymbol{D}_{m\times n}$, num

// $D_{m\times n}$ 是有 m 个属性向量的矩阵，num 是希望变换的次数

输出 $\boldsymbol{D}'_{m\times n}$

(1)sum=0, p=0, s=0　　//初始化

(2)Do

(3) k ← Rnd[m]，$V(A_1,\cdots,A_k)\leftarrow D_{k\times n}$, St ← St & k 个属性向量的编号

//2<k<m，随机选择 k 个属性向量，记录被选中属性的编号

(4) $V(A_1',A_2',\cdots,A_k')\leftarrow C_k\times V(A_1,A_2,\cdots,A_k)$

//用相应离散余弦变换矩阵变换

(5) $D_{k\times n}\leftarrow V(A_1',A_2',\cdots,A_k')$　//变换后重新放回

(6)计算最小隐私保护度 p

(7)If　$p>p_{max}$　then　$p_{max}=p$，　s=sum+1

//调整最优隐私保护度，记录下得到最优保护度时的变换次数

(8)sum=sum+1

(9)Until　$p\neq 0$　and　sum=num

//直到最小隐私保护度不为零且变换的次数达到初始设置值

(10) $D'_{m\times n}\leftarrow$ 根据 St 记录的属性向量编号，按顺序重新对 $D_{m\times n}$ 变换 s 次

(11)输出 $D'_{m\times n}$

3.3 讨论

变换之前数据经过了规范化处理，使所有属性有相同的权重，同时标识实体的属性(如 ID)被匿名保护，数据很难通过与外部数据库连接而被估计出来；当然，最重要的是利用 DCBT 对数据进行扭曲，在变换的过程中，因为随机选择了属性向量组，在多次变换后，很难从变换后的数据估计出原始数据，这样，原始数据得到很好的保护，从而也就保护了数据的隐私。

与RBT[3]不同的是，DCBT可同时变换多个向量属性，而不是每次变换向量属性对，同时，由于实施多次随机变换，原始数据能得到更好的保护

4 实验

医院为了研究的目的(例如关注一群患了相同疾病的病人)须共享一些数据。为符合隐私管理的要求，医院的安全管理员可能会从病人记录中删除一些标识(例如姓名、地址、电话号码等)。然而，释放的数据不一定被完全保护，病人的记录可能包含一些可以连接到其他数据库从而可以识别病人身份的信息，这就侵犯了病人的隐私。

本文选择关于心脏心率数据库[6]中的 451 条记录进行实验，分别运行 2 次程序，每次设置变换 30 次，第一次运行程序时，在第 13 次变换时最小隐私保护度可取最大值，即变换效果最好；而在第 2 次运行程序时，则在第 29 次变换时效果最好。因此，在设置适当的变换次数后，最小隐私保护度总能取得较好的值，得到较好的变换效果。程序运行界面如图 1 所示。

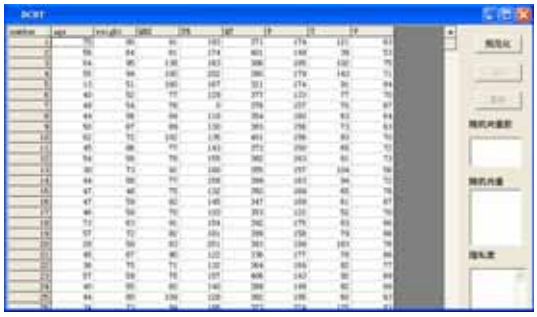

图 1　程序运行界面

5 结束语

本文利用离散余弦变换的正交性，主要考虑在集中式数据中应用 DCBT 变换，随机选择组合属性向量进行变换扰乱数据，使数据在保持相似性(距离不变)的同时实现扭曲，确保从变换后的数据中不能推算出初始数据，即保证其隐私度，

(下转第 161 页)