第 34 卷 第 21 期 计 算 机 工 程 2008 年 11 月
Vol.34 No.21 Computer Engineering November 2008
·安全技术· 文章编号：1000—3428(2008)21—0144—03 文献标识码：A 中图分类号：TP309.2

隐私保护聚类的独立噪音算法

王子亮，郑玉明，廖湖声
(北京工业大学计算机学院，北京 100022)

摘 要：数据挖掘技术具有很多优点，但存在隐私威胁等不足。该文针对聚类分析时如何保护隐私的问题，提出独立噪音思想并设计独立噪音算法(INA)。该算法对原数据叠加噪音以保护原始数据不被泄漏，所用噪音不会对数据分布造成严重影响，使后期挖掘工作可以在修改后的数据上直接进行。实验结果证明，INA 算法可以取得较高的隐私保护程度和挖掘正确率。

关键词：隐私保护；聚类挖掘；独立噪音

Independent Noise Algorithm for Privacy Preserving Clustering

WANG Zi-liang, ZHENG Yu-ming, LIAO Hu-sheng
(Collage of Computer, Beijing University of Technology, Beijing 100022)

【Abstract】 Data mining technique has a lot of merits, but it faces criticism like privacy threatening. This paper focuses on how to preserve privacy in clustering analysis and introduces the Independent Noise Algorithm(INA). This algorithm preserves original data by adding noises while keeping its distribution. The noises won't influence the data distributing badly so that the mining can be done directly on the amended data later. The experimental results demonstrate that INA is effective and provide acceptable values for balancing privacy and accuracy.

【Key words】 privacy preserving; clustering mining; independent noise

1 概述

数据挖掘是一种有效的知识发现方法，被众多学者深入研究并广泛应用。它可以从众多信息，如购物习惯、犯罪记录、病史、信用记录中，找出其内在联系，发现实用规律，用于决策或制定相应研究工作。但由于上述数据包含大量隐私信息，因此公开分析这些数据会造成隐私侵害。

基于保护隐私的数据挖掘是在不公开隐私信息的情况下，准确得出挖掘结果的方法，已成为数据挖掘研究的重要内容之一。隐私保护的方法主要有 2 种：(1)数据干扰，包括添加随机噪音、数据变换等方法，使变换后的数据不再代表实际的数值以保护原数据。(2)安全多方计算，以密码学为基础，常用于多个独立数据库之间的联合挖掘，保护各自的数据不被其他方得到。

本文提出独立噪音算法(Independent Noise Algorithm, INA)，此算法通过扰乱机密的数值型数据实现在聚类挖掘中对数据的保护，它不依赖任何特定聚类算法。

2 相关工作

文献[1]于 1996 年提出数据挖掘中的隐私问题，其他研究者随后提出了很多在数据挖掘中保护隐私的算法。本文主要研究应用于隐私保护聚类的算法和使用叠加噪音的算法。

文献[2]于 2000 年首次提出隐私保护数据挖掘算法。该算法通过叠加噪音保护隐私数据，并通过 Bayes 理论重建原始分布，利用重建结果，构造决策树。

对于聚类挖掘的隐私保护方法，文献[3]提出几何变换算法，通过对原数据进行几何变换，使变换后的数据偏离原数据。几何变换保证了数据之间的相对距离保持不变。

不同于几何变换算法，INA 通过向原数据叠加噪音来保护隐私，与文献[2]算法相比，INA 不需要还原原始数据分布，可以直接在干扰后的数据上进行聚类挖掘。

3 基本概念

3.1 数据干扰

数据干扰包括叠加噪音、数据变换等多种方法，本文只讨论叠加噪音的方法，也称为噪音算法。噪音算法最简单的形式是在原数据 X 上叠加噪音 Y 得到干扰后的数据 X'，可以描述为 $X'=X+Y$，其中，Y 是取自某种均值为 0 的概率分布(例如正态分布或均匀分布)。噪音算法主要应用于数值型或枚举型数据。

3.2 数据描述

设原数据库D由m个数值型属性$A_1, A_2,\cdots, A_m$构成；数据库D包含n个数据元组，记为$D=\{X_1, X_2,\cdots, X_n\}$；每个数据元组$X_i$记为$X_i=\{x_{i,1}, x_{i,2},\cdots, x_{i,m}\}$，其中，对任意的$i$=1, 2,$\cdots$, n, j=1, 2,$\cdots$, m，$x_{i,j}$是数值型数据；相应地，干扰后的数据库记为$D'=\{X_1', X_2',\cdots, X_n'\}$，其中，$X_i'=\{x_{i,1}', x_{i,2}',\cdots, x_{i,m}'\}$。

3.3 隐私度度量

根据原数据和干扰后的数据之间的差异程度来度量叠加噪音算法对隐私数据的保护程度，称为数据安全度[4]。若各元组在属性A_j上的取值记为$A_j=\{x_{1,j}, x_{2,j},\cdots, x_{n,j}\}$，对应的干扰后的数据记为$A_j'=\{x_{1,j}', x_{2,j}',\cdots, x_{n,j}'\}$，则属性$A_j$的数据安全度定义为$SEC_j=Var(A_j-A_j')/Var(A_j)$。函数$Var(A_j)$表示取值集合$\{x_{1,j}, x_{2,j},\cdots, x_{n,j}\}$的统计方差。对于$m$个数值型属性$A_1, A_2,\cdots, A_m$，定义它们的最低数据安全度为

基金项目：北京市自然科学基金资助项目(4052006)
作者简介：王子亮(1981－)，男，硕士研究生，主研方向：数据挖掘；郑玉明，教授；廖湖声，教授、博士生导师
收稿日期：2007-12-24 **E-mail：**wangziliang1022@yahoo.com.cn

$$mSec = \min_{i=1,2,\cdots,m}(Sec_i)$$

3.4 有效性度量

隐私保护算法的有效性是指对干扰前后的数据分别进行数据挖掘所得结果的偏差程度，偏差越大，有效性越低。对于聚类算法，一般采用误分类率作为偏差的度量。设一个聚类算法在原数据上得到u个簇，它们是$\{C_1, C_2,\cdots, C_u\}$；同样的算法在干扰后的数据上得到v个簇，它们是$\{C_1', C_2',\cdots, C_v'\}$。其中，若$C_w$与$C_w'$均存在，则$C_w$与$C_w'$被识别为一对相对应的簇。据此定义函数$f_{Mc}(X_i)$如下：若在原数据上经聚类运算，数据元组$X_i \in C_p$，而在干扰后的数据上有$X_i \in C_q'$，则只有在$q=p$时，函数取值为 0，否则，函数取值为 1。定义误分类率为 $Mc = \frac{1}{n}\sum_{i=1}^{n} f_{Mc}(X_i)$。可见，误分类率越高，隐私保护算法的有效性越低。

4 独立噪音

为了获得适当的有效性，噪音算法必须控制噪音的大小。INA 选用$(-d,+d)$之间的均匀分布作为随机噪音，称为大小为d的噪音。

传统噪音算法对不同数据元组在同一属性上的数据叠加同样大小的噪音。实验表明，只要噪音稍微偏大，就可能使很多数据元组被误分类，而如果噪音变小，虽然误分类率可以降低，但数据安全度也会下降。如果把数据库看作 m 维的欧几里德空间，把数据元组看作空间中的点，则产生误分类的元组全部集中在一个簇的边缘。其原因是处在簇中央的元组，即使叠加了较大的噪音，也不足以使它们偏移到该簇以外，而处在簇边缘的元组，只要噪音稍微偏大，就会使它们偏移到另一个簇的范围内。

上述不平衡现象提示了解决隐私度和有效性之间矛盾的方法，即增大叠加在簇中央的元组上的噪音，可以在不增加误分类的前提下提高隐私度，而减小叠加在簇边缘的元组上的噪音，可以在较大隐私度条件下获得较高聚类运算有效性。

因此，可得独立噪音主要思想如下：(1)对处在簇中央的数据元组叠加较大噪音；(2)对处在簇边缘的数据元组叠加较小噪音；(3)噪音的大小不使元组偏移出原来簇的范围。

5 独立噪音算法设计与分析

实现独立噪音的关键是估算每个数据元组在簇内的位置及其到簇边界的距离。INA 采用一种称为直接归类的算法。利用该算法可以从原数据库的样本数据中得到一些参考点，对每个数据元组与这些参考点进行比较，可以估算出数据元组在簇中的位置，并计算出将要对它叠加的噪音大小。

5.1 独立噪音算法设计

如图 1 所示，INA 分为 2 个部分：(1)直接归类，在样本数据上执行直接归类得到若干个参考点；(2)独立噪音的计算与叠加，根据原数据元组与参考点之间的距离，计算噪音的大小并对数据元组叠加噪音。

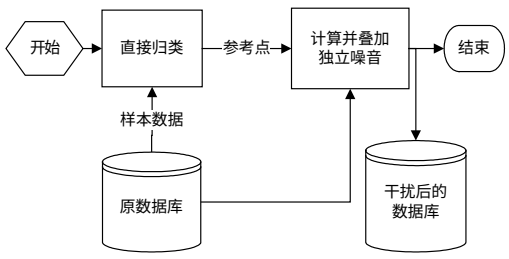

图 1　独立噪音算法

INA涉及不同数据元组之间距离的计算。由于语义不同的属性具有不同域和度量单位，因此 2 个数据元组X_1, X_2之间的距离可以定义为

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{m}\left(\frac{x_{1,i} - x_{2,i}}{\alpha_i}\right)^2}$$

其中，α_i 是为了规范不同属性的域和度量单位而引入的规范化因子。

在INA的第 1 部分中，样本数据是从原数据库中经过随机抽取部分数据元组形成的。设样本数据通过执行直接归类算法，被归并成b个小类，每个小类中包含的数据元组个数由$count_i$指示$(i=1,2,\cdots,b)$，每个小类的中心坐标$Q_1, Q_2,\cdots, Q_b$即所求参考点。直接归类算法的具体过程如下：

(1)获取样本数据所有数据元组在各属性上的最大值和最小值$\max_1, \max_2,\cdots, \max_m$和$\min_1, \min_2,\cdots, \min_m$，并令$P_{\max}=(\max_1, \max_2,\cdots, \max_m)$, $P_{\min}=(\min_1, \min_2,\cdots, \min_m)$。

(2)令$r_0=dist(P_{\max}, P_{\min})/\beta$。

(3)从样本数据中取出一个元组$X=(x_1, x_2,\cdots, x_m)$。

(4)初始化，令$b=1$，$Q_1=(x_1, x_2,\cdots, x_m)$, $count_1=1$。

(5)重复以下步骤直到样本数据集为空：

1)从样本数据中另取一个数据元组$X=(x_1, x_2,\cdots, x_m)$；

2)从$Q_1, Q_2,\cdots, Q_b$中找到与X距离最近的点，设该点为Q_i，它与X的距离为r；

3)如果$r<r_0$，则更新Q_i，$count_i=count_i+1$；否则$b=b+1$, $Q_b=(x_1, x_2,\cdots, x_m)$, $count_b=1$。

(6)输出参考点$Q_1, Q_2,\cdots, Q_b$。

步骤(5)的 3)中所述“更新Q_i”，是指对Q_i的每个分量$q_{i,j}(j=1,2,\cdots,m)$做如下更新运算：

$$q_{i,j} = \frac{q_{i,j} \times count_i + x_j}{count_i + 1}$$

为了限定每个小类的空间大小，INA 引入 β 参数，β 越大，参考点的数量越多。

利用由直接归类算法得到的参考点$Q_1, Q_2,\cdots, Q_b$，可以计算原数据库D中每个元组的独立噪音大小，并完成噪音叠加，生成干扰后的数据库D'，算法过程如下：

(1)令 D'为空。

(2)重复以下步骤，直到处理完 D 中的所有数据元组：

1)顺序读取一个数据元组$X=(x_1, x_2,\cdots, x_m)$；

2)从$Q_1, Q_2,\cdots, Q_b$中找到与该元组距离最近的 2 个参考点。设这 2 个参考点与X的距离分别为r_1, r_2，且$r_1<r_2$；

3)对于$i=1,2,\cdots,m$，依次计算$d_i=\alpha_i\times\frac{r_2-r_1}{2}$，$x_i'=x_i+random(-d_i,d_i)$；

4)将$X'=(x_1', x_2',\cdots, x_m')$加入$D'$；

(3)输出 D'。

由实验可知，由于样本数据的容量远小于数据库的数据量，因此 INA 的 I/O 开销与一遍读取数据库数据的开销相近。

INA 的时间开销主要取决于计算数据元组与参考点之间距离的时间。在算法第 1 部分中，样本数据的每个元组要与当前已得到的参考点计算一次距离。在算法第 2 部分中，数据库的每个元组要与每个参考点计算一次距离。因此，算法的时间开销为 $O(bn)$。

5.2 独立噪音算法思想分析

执行 INA 时，当处于簇中心的数据元组 X 较接近某个参

考点时，$(r_1 - r_2)$ 较大，数据元组叠加了较大噪音；当处于簇边缘的数据元组 X 远离所有参考点时，$(r_1 - r_2)$ 较小，数据元组叠加了较小噪音。可见，INA 满足第 4 节所述独立噪音主要思想的第(1)条和第(2)条。

INA 满足第(3)条独立噪音主要思想，证明如下：

(1)INA给数据元组X叠加的噪音不改变与X距离最近的参考点。设Q_1, Q_2分别是与数据元组X距离最近和次近的 2 个参考点，其距离分别为r_1, r_2。上述 3 个数据点经过规范化后的图像如图 2 所示，其中，直线l是线段Q_1Q_2的中垂线；X'是X关于直线l的对称点；线段XX'与l相交于点P。

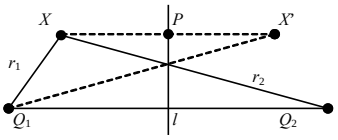

图 2　独立噪音算法分析

根据对称性和三角形三边关系，可以得到

$$d=\frac{r_2-r_1}{2}=\frac{|Q_2X|-|Q_1X|}{2}=\frac{|Q_1X'|-|Q_1X|}{2}<\frac{|XX'|}{2}=|XP|$$

因此，给数据元组X叠加大小为d的噪音时，X仍将驻留在直线l的左侧，距离X较近的参考点仍然是Q_1。对其他距离较远的参考点而言，此结论同样成立。可见，给数据X叠加大小为d的噪音，可以保证与X距离最近的参考点保持不变。

(2)若以“距离哪个参考点最近”为标准，把数据元组空间分割成一些小的区间。对比实际聚类结果可以发现，上述小区间大部分是聚类所得簇的更细的划分，也有极少数区间会跨越 2 个或多个簇，但这些区间普遍很小。因此，只要保证数据元组在叠加噪音后还能驻留在原来的区间，就保证了大部分元组不会偏移出原来簇的范围。

综上所述，INA 满足了独立噪音思想的所有要求。

6　实验结果及分析

6.1　实验策略及结果

算法检验采用二维平面上的点坐标作为数据元组，不考虑噪音，共生成 5 个原数据库，分别对应簇个数 k 为 2~6 的聚类。每个原数据库包含 10 000 个点的数据；算法中 β 值取 5，样本数据分别取 50, 100, 200, 1 000 个数据元组；聚类算法选用 k-means 算法和 Chameleon 算法。在实验中，对每个原数据库分别在不同样本数据大小下，各生成 10 个干扰后的数据库，以这 10 个数据库的安全度和误分类率的平均值作为结果。实验结果如表 1~表 3 所示。

表 1　不同样本数据大小下的最低数据安全度　(%)

样本大小	k=2	k=3	k=4	k=5	k=6
50	2.05	1.80	3.83	4.44	3.22
100	1.61	2.07	3.94	3.66	3.37
200	1.47	1.77	4.16	3.45	3.40
1 000	1.47	0.93	0.90	0.84	1.21

表 2　应用 k-means 算法的误分类率　(%)

样本大小	k=2	k=3	k=4	k=5	k=6
50	0.71	1.23	1.01	1.63	3.09
100	0.48	1.83	0.72	1.02	1.58
200	0.36	1.55	0.61	0.75	1.30
1 000	0.28	1.32	0.62	0.94	1.13

表 3　应用 Chameleon 算法的误分类率　(%)

样本大小	k=2	k=3	k=4	k=5	k=6
50	0.90	0.63	1.23	2.08	3.87
100	0.51	0.48	0.80	1.18	1.30
200	0.44	0.53	0.65	0.93	1.57
1 000	0.40	0.73	0.75	1.13	1.15

6.2　结果分析

由表 1 可见，INA 的数据安全度随着样本数据大小的增大而降低，在表 2 和表 3 中，随着样本数据大小的增大，误分类率呈现下降趋势。其原因是随着样本数据的增大，通过直接归类得到的参考点会增多，减小了大部分数据元组的噪音大小。

由表 2 和表 3 可以看出，INA的误分类率随簇个数k的增加而增大。由于容易发生误分类的数据元组集中在跨越多个簇的区间内，而对应较大簇个数k的原数据库，其数据分布较松散，因此不同簇之间的边界不是很清晰。在此情况下，跨簇区间出现的几率增加，算法的误分类率就会随之变大。

对相同测试数据，传统噪音算法在数据安全度达 2%时，误分类率为 4%以上。INA 在大部分数据安全度为 3%以上的情况下，误分类率大多在 2%以下。可见，INA 对传统噪音算法的改进效果显著。

7　结束语

叠加噪音是隐私保护数据挖掘中的一个基本方法，被应用于各种数据挖掘工作。本文在传统噪音算法的基础上，提出独立噪音思想并设计独立噪音算法。该算法能在保护隐私与正确挖掘结果之间取得较好平衡。

参考文献

[1] Clifton C, Marks D. Security and Privacy Implications of Data Mining[C]//Proc. of ACM Workshop on Data Mining and Knowledge Discovery. [S. l.]: ACM Press, 1996.

[2] Lindell Y, Pinkas B. Privacy Preserving Data Mining[J]. Cryptology, 2002, 15(3): 177-206.

[3] Oliveira S R M, Zaiane O R. Privacy Preserving Clustering by Data Transformation[C]//Proc. of the 18th Brazilian Symposium on Databases. Manaus, Brazil: [s. n.], 2003.

[4] Muralidhar K, Parsa R, Sarathy R. A General Additive Data Perturbation Method for Database Security[J]. Management Science, 1999, 45(10): 1399-1415.

（上接第 143 页）

[4] Xia Xianggen. Discrete Chirp-Fourier Transform and Its Application to Chirp Rate Estimation[J]. Trans. on Signal Processing, 2000, 48(11): 3122-3133.

[5] 孙泓波, 郭　欣, 顾　红, 等. 修正离散 Chirp-Fourier 变换在运动目标检测中的应用[J]. 电子学报, 2003, 31(1): 25-28.