第 38 卷 第 3 期
Vol.38 No.3

计 算 机 工 程
Computer Engineering

2012 年 2 月
February 2012

• 博士论文 •

文章编号：1000—3428(2012)03—0012—02 文献标识码：A 中图分类号：TP391.4

基于概率论的隐私保持分类挖掘

李 光，王亚东，苏小红

(哈尔滨工业大学计算机科学与工程系，哈尔滨 150001)

摘 要：在现有的基于数据扰动的隐私保持分类挖掘算法中，扰动数据和原始数据相关联，对隐私数据的保护并不完善，且扰动算法和分类算法耦合度高，不适合在实际中使用。为此，提出一种基于概率论的隐私保持分类挖掘算法。扰动后可得到一组与原始数据独立同分布的数据，使扰动数据和原始数据不再相互关联，各种分类算法也可直接应用于扰动后的数据。

关键词：数据挖掘；隐私保持；数据扰动；随机噪声；决策树

Privacy-preserving Classification Mining Based on Probability Theory

LI Guang, WANG Ya-dong, SU Xiao-hong

(Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 In the existed privacy-preserving classification mining methods based on data perturbation, the privacy data is not protected perfectly because the perturbed data and the original data have been related. The classification algorithm and the data perturbation algorithm have high coupling It is not easy to use these methods in practice. To solve these problems, it proposes a privacy-preserving classification mining algorithm based on probability theory. The perturbed data is independent from the original data and they have the same distribution. This proposed method overcomes the shortcomings of others. The perturbed data is no relation with the original data and the classification methods can be used on the perturbed data directly.

【Key words】 data mining; privacy protection; data perturbation; random noise; decision tree

DOI: 10.3969/j.issn.1000-3428.2012.03.005

1 概述

随着数据挖掘应用的不断普及，越来越需要在隐私数据上进行挖掘，隐私保持逐渐成为数据挖掘应用中的重要问题。隐私保持的数据挖掘由此被提出，并且引起了研究界的普遍关注[1-4]。

目前，很多隐私保持的数据挖掘方法都是基于数据扰动实现的[1-3]。基于数据扰动的方法不公开原始数据，而是公开一组由原始数据经扰动而得的新数据。用户通过处理扰动数据得到原始数据上的挖掘结果。

很多基于数据扰动的隐私保持的分类挖掘方法都是基于概率论设计的。这些方法的一个基本观点是分类挖掘关心数据的总体趋势而非具体取值。而隐私泄漏却需要得到隐私数据的具体取值。因此，如果仅知道原始数据的某些重要的分布信息，就可以进行分类挖掘而且不会泄漏隐私。

文献[1-2]通过加随机噪声的方法进行数据扰动。设原始数据为 X，扰动数据为 Y，则 $Y=X+r$，其中，r 是一个分布已知的随机数。文献[1-2]给出了在已知 Y 的情况下重构 X 的分布并生成决策树的算法。文献[5]指出这种加随机噪声的方法在安全性上存在一定的问题，提出了从扰动后数据中恢复原始数据的方法。

文献[6]对布尔型数据给出了一种基于随机响应的扰动方法。在该方法中，设原始数据为 X，X=0 或 1，设定参数 $p(0\leqslant p\leqslant 1)$，$X$ 以 p 的概率保持不变，以 $1-p$ 的概率取反，从而得到扰动数据 Y。文献[6]给出在 Y 上生成决策树的算法。

文献[7]将文献[6]的方法拓展到离散型数据上去。设原始数据 X 一共有 m 个取值，分别为 $x_1,x_2,\cdots,x_m$。使用一个可逆的转移概率矩阵 $\boldsymbol{P}_A$ 对 X 进行扰动。设 p_{ij} 为 $\boldsymbol{P}_A$ 中第 i 行第 j 列的元素。若 $X=x_i$，则扰动数据 $Y=x_j$ 的概率为 p_{ij}。文献[7]也给出了在扰动数据上生成决策树的算法。

综上所述，目前已设计出多种基于概率论的数据扰动方法，用于隐私保持的分类挖掘。但在这些方法中，扰动数据和原始数据依然存在联系，并没有做到仅保留分布信息。正因如此才出现了一些从扰动数据上恢复原始数据的方法[5]。而且在这些方法中，挖掘算法和数据扰动方法耦合程度较高，造成了这些方法往往仅适用于某一特定分类算法，不方便在实际中使用。

为了克服以上缺点，本文提出了一种新的基于概率论的隐私保持分类挖掘方法。

2 算法描述

假设原始数据为数值型，并且已经处理为标准的关系型数据库的形式，即已经存储在一个二维表格中了。如果数据不是数值型的，则可以对它进行编码，转化为数值型数据。设原始数据中的隐私数据一共有 m 个属性、n 个元组，排列为矩阵 $\boldsymbol{A}$，$\boldsymbol{A}$ 的列代表属性，行代表元组，将第 i 个元组在第 j 个属性上的值记为 a_{ij}。

基金项目：国家“863”计划基金资助项目(2007AA02Z329)

作者简介：李 光(1982－)，男，博士研究生，主研方向：数据挖掘，生物信息学；王亚东、苏小红，教授、博士生导师

收稿日期：2011-07-25 **E-mail：**hit6006@126.com

首先，将各个隐私属性看成是独立的，分别对每个属性进行统计，得到每个属性的经验分布函数，并由此函数独立地生成该属性的新数据。详细过程如下。

对于 $\boldsymbol{A}$ 的第 i 列，设该列元素构成集合 $X_i=\{a_{1i},a_{2i},\cdots,a_{ni}\}$，生成区间划分点 $x_1,x_2,\cdots,x_k$，满足：

$$\min\{X_i\}=x_1\leqslant x_2\leqslant\cdots\leqslant x_k=\max\{X_i\}$$

在本文中采用等分区间的方法，即令：

$$x_i=\min\{X_i\}+(\max\{X_i\}-\min\{X_i\})(i-1)/(k-1)$$

令落入区间$(x_{i-1},x_i]$的样本的个数为 $n_i(i\geqslant 2)$，落入区间$(-\infty,x_1]$的样本个数为 n_1，若 $n_i=0$，则令其为 1 以进行平滑。由统计结果可以直接估计 X_i 的分布函数在分割点处的值，分布函数在其他点处的值由线性插值方法得到。X_i 的分布函数 $F(x)$定义如下：

$$\begin{cases}0 & x<x_1\\ \dfrac{\sum_{i=1}^{j}n_i}{\sum_{i=1}^{k}n_i} & x=x_j\\ F(x_{j-1})+\dfrac{F(x_j)-F(x_{j-1})}{x_j-x_{j-1}}(x-x_{j-1}) & x_{j-1}<x<x_j\\ 1 & x>x_k\end{cases}$$

生成满足此分布函数的 n 个数，记为 $Y_i=\{b_{1i},b_{2i},\cdots,b_{ni}\}$。

随后，利用序关系重构不同属性之间的联系。分别对集合 $X_1,X_2,\cdots,X_m$ 和 $Y_1,Y_2,\cdots,Y_m$ 中的元素进行排序，排序后的结果以向量形式分别表示为 $\boldsymbol{X}'_1,\boldsymbol{X}'_2,\cdots,\boldsymbol{X}'_m$ 和 $\boldsymbol{Y}'_1,\boldsymbol{Y}'_2,\cdots,\boldsymbol{Y}'_m$，其中，设向量 $\boldsymbol{X}'_i=(a'_{1i},a'_{2i},\cdots,a'_{ni})$，向量 $\boldsymbol{Y}'_i=(b'_{1i},b'_{2i},\cdots,b'_{ni})$，由于是排序后结果，因此有 $a'_{1i}\leqslant a'_{2i}\leqslant\cdots\leqslant a'_{ni}$，以及 $b'_{1i}\leqslant b'_{2i}\leqslant\cdots\leqslant b'_{ni}$。若原始隐私数据矩阵 $\boldsymbol{A}$ 中某一行为 $(a'_{p_11},a'_{p_22},\cdots,a'_{p_nn})$，则 $(b'_{p_11},b'_{p_22},\cdots,b'_{p_nn})$ 为生成的隐私数据中的一行，$(a'_{p_11},a'_{p_22},\cdots,a'_{p_nn})$ 所对应的非隐私数据直接添加到 $(b'_{p_11},b'_{p_22},\cdots,b'_{p_nn})$ 中去，构成生成数据的一个元组。

例如，假设原有的数据为：

$$\begin{bmatrix}2&6&3&9&7\\4&7&1&6&8\\0&1&1&0&0\end{bmatrix}^{\mathrm{T}}$$

其中，第 3 列为分类标记，不参与扰动，仅对前 2 列数据进行扰动。假设新生成的 2 列数据分别表示为单列的矩阵 $\boldsymbol{Y}_1=[1\ 2\ 8\ 7\ 5]^{\mathrm{T}}$ 以及 $\boldsymbol{Y}_2=[3\ 4\ 6\ 9\ 5]^{\mathrm{T}}$，在恢复属性间关联时，因为 $\boldsymbol{Y}_1$ 中的元素依次是 $\boldsymbol{Y}_1$ 中第 1、第 2、第 5、第 4、第 3 小的元素，而在原始数据中，第 1 列第 1、第 2、第 5、第 4、第 3 小的元素分别对应第 2 列第 2、第 1、第 3、第 5、第 4 小的元素。在 $\boldsymbol{Y}_2$ 中，第 2、第 1、第 3、第 5、第 4 小的元素分别为 4、3、5、9、6，因此，扰动数据中 $\boldsymbol{Y}_2$ 应排列为 $[4\ 3\ 5\ 9\ 6]^{\mathrm{T}}$。又因为在原始数据中，第 1 列第 1、第 2、第 5、第 4、第 3 小的元素对应第 3 列的元素值分别为 0、1、0、0、1，因此，根据序关系连接后的数据为：

$$\begin{bmatrix}1&2&8&7&5\\4&3&5&9&6\\0&1&0&0&1\end{bmatrix}^{\mathrm{T}}$$

3　算法分析

分别在原始数据和本文方法扰动后的数据上训练决策树、最近邻、人工神经网络和支持向量机，并比较得出的分类器在测试样本上的分类正确率。实验使用文献[1]的数据。实验结果如表 1 所示。其中，$F1$~$F5$ 表示 5 个不同的分类函数。可以看出，各种分类算法在原始数据和本文方法扰动后的数据上都能得到分类正确率相差不大的分类器。本文方法具有很好的准确性。

表 1　分类器分类正确率比较　(%)

挖掘方法	训练数据	$F1$	$F2$	$F3$	$F4$	$F5$
决策树	原始数据	100.00	87.18	99.98	99.14	98.16
	扰动数据	99.78	86.85	99.46	99.13	98.16
最近邻	原始数据	100.00	99.88	99.84	99.86	98.28
	扰动数据	100.00	99.88	99.82	99.78	98.68
神经网络	原始数据	99.78	94.92	91.80	98.16	91.66
	扰动数据	99.74	93.98	91.56	98.48	92.88
SVM	原始数据	99.98	99.46	99.04	99.96	98.00
	扰动数据	100.00	99.60	99.18	99.90	98.40

使用基于信息论的度量指标[2]衡量隐私性，仍然使用文献[1]的数据。假设原始数据为 X，扰动后数据为 Y，则使用 $Q(X|Y)=1-2^{-I(X,Y)}$表示隐私保持程度，其中，$I(X,Y)$表示 X 与 Y 的互信息。$Q(X|Y)$越小，说明隐私保持程度越高。图 1 显示了分别在本文方法和加随机噪声方法扰动后的数据上计算 $Q(X|Y)$的结果。

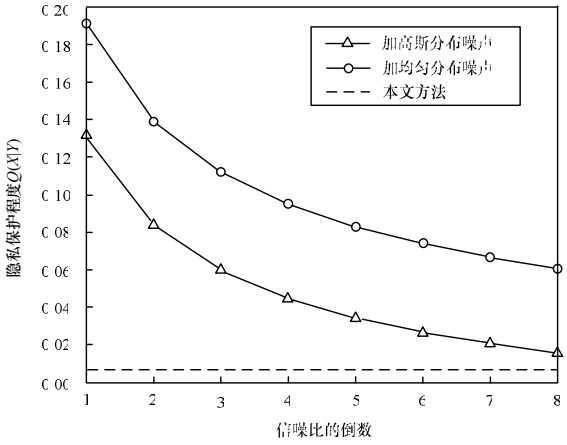

图 1　隐私性对比

在加随机噪声的方法中，$Y=X+r$，其中，X 为原始数据；Y 为扰动数据；r 是噪声。在本实验中，分别向原始数据添加不同强度的服从均匀分布和高斯分布的噪声。噪声强度由信噪比 SNR 决定。SNR 定义为 $D(X)/D(r)$，即原始数据和噪声的方差之比。由图 1 可以看出，与加随机噪声的方法相比，本文方法提供了更好的隐私保持。

图 2 显示的是使用文献[5]提出的基于主成分分析(PCA)的还原方法分别对本文方法和加随机噪声方法扰动后的数据进行还原的结果。

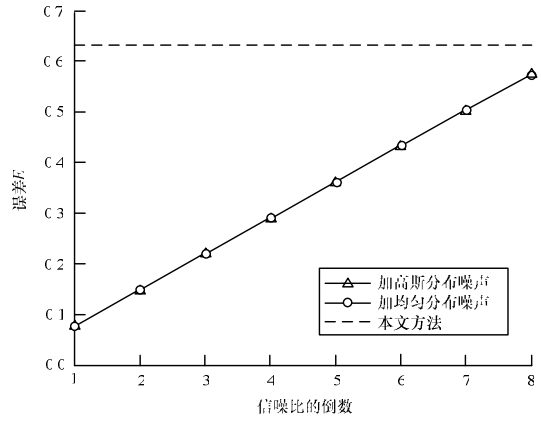

图 2　基于 PCA 的还原方法效果比较

(下转第 18 页)