一种基于决策树的隐私保护数据流分类算法

陈 煜,李玲娟
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:隐私保护的决策树挖掘方法主要是基于数据扰动的方法和基于安全多方计算的方法。由于数据流高速、连续无限和动态的特性,这些隐私保护方法在数据流挖掘应用上有所不足。针对当前数据流挖掘应用中的隐私泄露问题,提出了一种基于决策树的隐私保护的数据流分类算法—PPFDT。该算法通过采用添加随机噪声的方法对数据加以隐私保护,改进经典的数据流挖掘算法—VFDT,并使用阈值算法找到扰动数据流的最佳分裂属性和最佳分裂点,从而直接在扰动数据流上建立决策树,通过使用该决策树对初始数据流和扰动数据流分类得到较精准的结果。从 PPFDT 算法的隐私保护程度和在直接扰动的数据流上的分类性能两方面,基于 UCI 的 WaveForm 数据集进行了实验验证。实验结果表明,该算法在数据流上快速准确分类的同时,具有一定的隐私保护程度。

关键词:决策树;隐私保护;数据流;分类
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2017)07-0111-04
doi:10.3969/j.issn.1673-629X.2017.07.026

A Decision Tree-based Privacy Preserving Classification Mining Algorithm for Data Streams

CHEN Yu, LI Ling-juan
(School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Privacy preserving data mining methods are mainly based on perturbation and randomization approaches and secure multi-party computation approaches. Due to the high-speed data streams with unlimited continuous and dynamic characteristics, these methods are still inadequate. In order to solve privacy leaking problem on current data streams mining application, a privacy preserving fast decision tree mining algorithm for data streams named as PPFDT has been designed and implemented. It adds random noises to protect data privacy and improves the data mining algorithm named VFDT, and uses threshold method to find the best split attribute and the best split point of perturbed data streams, so that a decision tree is directly built on perturbed data streams. Then the decision tree is used to classify original data streams and perturbed data streams for getting accurate results. From the aspects of the privacy protection degree of the PPFDT algorithm and the classification performance on the direct perturbed data stream, the algorithm has been experimentally verified on the Waveform dataset of UCI. The experimental results show that the algorithm can achieve certain degrees of privacy protection, and at the same time, classify data streams fast and accurately.

Key words: decision tree; privacy preserving; data stream; classification

0 引 言

随着学术界和政府对隐私保护越来越重视,隐私保护的数据挖掘算法已经成为一个研究热点。其中决策树是一种重要的分类算法,也是隐私保护分类挖掘的首选模型算法。隐私保护的决策树挖掘技术按照实现技术主要分为两大类,即基于数据扰动的方法和基于安全多方计算的方法[1]。

在基于数据扰动的隐私保护的决策树挖掘方法中,用户隐私数据通过扰动加以隐私保护,并且被扰动的数据在不泄露隐私的情况下用来重构用户隐私数据的分布,然后分类算法和关联算法通过重构的数据分布进行挖掘。

收稿日期:2016-07-16 **修回日期**:2016-10-26 **网络出版时间**:2017-04-28
基金项目:国家自然科学基金资助项目(61302158,61571238)
作者简介:陈 煜(1992-),男,硕士研究生,CCF 会员(E200052164G),研究方向为流数据挖掘与信息安全;李玲娟,教授,CCF 会员(E200015276M),研究方向为数据挖掘、信息安全、分布式计算。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20170428.1703.072.html

2000 年,Rakesh Agrawal 等[2]提出用加随机噪声的方法进行隐私保护的决策树挖掘。设真实数据为 X,随机生成分布已知的噪声 R,扰动后数据 $Y = X + R$。真实数据 X 是不公开的,数据的使用者得到的是扰动后数据 Y 的值,以及噪声 R 的分布。因为 R 是一个随机数,使用者仅知道 R 的分布而不知道其具体值,因此无法得到真实数据 X 的值。然后,Rakesh Agrawal 等使用一种基于贝叶斯理论的迭代方法来估计 X 的分布,利用该分布生成决策树,并使用一种基于区间估计的方法来度量隐私保护程度。

随后,这种加随机噪声的方法的安全性受到质疑,H. Kargupta 等[3-4]基于随机矩阵理论,提出了一种从扰动后的数据上估计真实数据的方法。此后,又有基于随机响应的数据扰动、奇异值分解和 K 匿名化等隐私保护方法被提出[5-8]。

2009 年,Li Liu 等[9]开发了直接在扰动后数据上生成决策树的算法。设样本集合为 S,分裂后为集合 S_1 和 S_2,该方法的核心思想是对每个样本,计算它属于集合 S_1 和 S_2 的概率 p_1 和 p_2,并利用计算出的概率值生成决策树,从而避免了估计真实数据的分布这一复杂的工作。

由于基于数据扰动方法的通信开销和时间复杂度较小,采用数据扰动方法,对数据流分类挖掘算法进行了改进,设计并提出了一种基于决策树的隐私保护的数据流分类算法—PPFDT。该算法是在扰动数据上直接进行挖掘,而不是通过扰动数据预估初始数据的大体分布来建立决策树,从而可以得到高质量的挖掘结果。更为重要的是,通过改进 VFDT 算法,在随机扰动的数据流上直接建立决策树模型,并且可以增量更新决策树模型,来分类原始数据或者扰动数据。

1 流数据分类挖掘算法—VFDT

VFDT 算法[10]是增量式算法,在处理每一个决策树的节点时仅依赖于整个数据的部分子样本,对整个流数据仅作一次扫描,得到一个近似解。与传统算法相比,VFDT 具有较高的时空效率,分类器的性能可以渐近于传统算法生成的分类器。

VFDT 算法是通过对 Hoeffding 树改进实现的,Hoeffding 树通过不断将叶子节点转换为内部节点而生成,其中每个叶节点都保存有关于属性值的统计信息,这些统计信息用于计算属性的信息增益。VFDT 算法是采用信息熵或者 Gini 指标作为选择分裂属性的标准,以 Hoeffding 不等式作为判定节点分裂的条件进行的。

VFDT 算法的有关定义如下:

(1)Hoeffding 边界值 ε[11]:代表在 Hoeffding 树的每一个节点上,正确的属性被选择的概率。描述如下:

对一个真值随机变量 r,其取值范围为 R。假定对 r 取 n 个独立的观察值,并计算它们的平均值 $\bar{r}$,其 Hoeffding 约束对于可信度 $1-\delta$,变量 r 的真实值至少是 $\bar{r}-\varepsilon$。其中,$\varepsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n}}$,$r$ 为信息增益,R 的取值范围是 $\log_2$#Classes,Classes 是属性类别取值的数目。

(2)信息增益[12]:叶子节点 l 中存储训练样本集 D 的统计信息,则对样本集 S 分类所需的期望信息如下:

$$\mathrm{Info}(S) = -\sum_{j=1}^{k} \frac{\mathrm{freq}(C_j,S)}{|S|} \times \log_2\left(\frac{\mathrm{freq}(C_j,S)}{|S|}\right) \tag{1}$$

其中,$|S|$ 为 S 中的实例数目;$\mathrm{freq}(C_j,S)$ 为 j 从 1 到 k,属于类 j 的实例数目。

则对于叶子节点可能的分裂属性 A,有 n 个取值,对于将属性 A 划分成对样本集 S 的分类所需要的期望信息如下:

$$\mathrm{Info}_A(S) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \mathrm{Info}(S_i) \tag{2}$$

属性 A 的信息增益为:

$$\mathrm{Gain}(A) = \mathrm{Info}(S) - \mathrm{Info}_A(S) \tag{3}$$

(3)主动分裂系数 τ[13]:τ 作用在于,当几个属性的信息增益 G 几乎相等时,此时可能需要更多的样本来决定叶子节点的决策属性,设定 τ 值用以主动选择属性并实现叶子节点分裂。当满足 $\Delta G < \varepsilon < \tau$ 时,选择 ΔG 中信息增益最大或者次大的属性作为该叶子节点的决策属性。

(4)节点分裂准则:判断节点分裂的有效条件之一是:

$$(\overline{G_l(X_a)} - \overline{G_l(X_b)}) > \varepsilon \text{(或者 } \varepsilon < \tau) \tag{4}$$

其中,$\overline{G_l(X)}$ 是在叶节点中属性为 X 的信息增益;X_a、X_b 分别为具有最大及次大信息增益的属性;τ 为用户设定的 Tie 值,用于确定叶节点的决策属性值。

2 PPFDT 算法设计

在此,将描述 PPFDT 算法如何在扰动数据流上建立 VFDT 决策树分类器,并对初始数据和扰动数据进行分类。

2.1 算法设计思想

对于以 j 作时间戳,X_j 表示 j 时刻到达的数据向量,形式为 $\{\cdots X_{j-1}, X_j, X_{j+1}\cdots\}$ 的数据流,采用加随机噪声的方法,形成表示为 $\{\cdots W_{j-1}, W_j, W_{j+1}\cdots\}$ 的扰动数据流,因为添加了随机噪声,不知道 X_j 的值,因此达到了隐私保护的效果。

随后,对于 VFDT 算法只能处理离散型数据的问题,在 PPFDT 算法中,对于连续数值型属性 A,其值为 $\{\cdots X_{t-1}, X_t, X_{t+1}\cdots\}$,然后以 X_t 的值划分为两个集合 S_1、S_2。那么以 $\leqslant X_t$ 的值划分为集合 S_1,对 $> X_t$ 的值化分为集合 S_2。对每个划分,计算其信息增益,并选择信息增益最大的划分。对于扰动后的数据 $W_j = X_j + R$ (R 是已知分布的随机噪声),对每个划分,计算 $P_s(W_j \mid X_t)$ 的概率,从而选择最佳的 W_t 代替原始数据 X_t 划分。对于 $P_s(W_j \mid X_t)$ 的概率的计算,由于 $P_s(W_j \mid X_t) = P(W_j \in S_1) = P(W_j - X_t \leqslant R)$,通过 R 的分布和 X_t 的值,可以计算出 $P(W_j - X_t \leqslant R)$,所以可以计算得到 $P_s(W_j \mid X_t)$ 的概率值。

PPFDT 算法的核心内容如下:

(1)阈值算法。

设定一个阈值,统计 W_j 有类标签 C_j 并且高于阈值的个数,那么公式为:

$$\text{freq}'(C_j, S_i) = \sum_{W_j \in S_i} (I_{W_j \in C_j}, P_{S_i}(W_j) > \text{阈值}) \quad (5)$$

基于式(5),使用 $P_{S_i}(W_j)$ 计算扰动数据 W_j 属于 S_i 的概率,$I_{W_j \in C_j}$ 是一个指标函数,当扰动数据 W_j 有类标签 C_j 时返回 1。

$$\text{Info}(S) = -\sum_{j=1}^{k} \frac{\text{freq}'(C_j, S)}{|S|} \times \log_2\left(\frac{\text{freq}'(C_j, S)}{|S|}\right) \quad (6)$$

其中,$|S|$ 为样本集 S 中的实例数目;$\text{freq}'(C_j, S)$ 为扰动数据 W_j 有类标签 C_j 并且高于阈值的个数。

使用式(5)、式(6)和式(2)~(4),能够找到最佳分裂属性和连续属性的最佳分裂点。

(2)使用阈值算法划分训练样本。

在训练数据上改进的分裂准则更直观:

$$\text{splitThreshold}(W_j, X_t) = \begin{Bmatrix} S_1, P_{S_i}(W_j \mid X_t) > \text{阈值} \\ S_2, P_{S_i}(W_j \mid X_t) < \text{阈值} \end{Bmatrix} \quad (7)$$

设定合适阈值建立一个成功的决策树并不容易,$P_S(W_j \mid X_t)$ 是 W_j 在给定分裂点 X_t 下属于 S 的概率。当分裂点 X_t 和扰动实例 W_j 给出后,唯一的不确定性是随机噪声 R。换句话说,阈值的选择跟随机噪声的分布有关,在实验中,使用了均匀分布和高斯分布的噪声。阈值在这两种分布下并不相同。对于高斯分布,阈值设为 0.3,分类器有最高的准确度;对于均匀分布,阈值设为 0.5,分类器有最高的准确度。

2.2　算法描述

建立算法 PPFDT 的流程伪代码如下:

输入:T,扰动数据流;X,数值连续属性集;δ,预设置信度;τ,用户设定 Tie 值;$n_{\min}$,叶节点分裂测试所需样本数目的最小值;threshold,阈值。

输出:一棵决策树 PPFDT。

1:令 PPFDT 是一棵只有一个根节点 l_1 的树,令 $X_l = X \cup X_\Phi$

2:对每个扰动实例 W_j,令 $n_{ijk}(l_1) = 0$(n_{ijk} 是某叶节点第 j 个可能属性的第 i 个取值的属于类别 k 的数目),然后从决策树根节点开始,根据扰动实例 W_j 在该节点属性的取值,计算其属于集合 S_1、S_2 的概率,根据分裂准则,从上到下遍历进入不同的节点分支,直到叶节点

3:当扰动实例到达叶节点 l,将扰动实例 W_j 在该叶节点属性的取值插入到该叶节点的二叉排序树中

4:该叶节点 l 的统计值 $n_{ijk}(l)+1$

5:将叶节点 l 的标签标记为到 l 为止可见的扰动实例所属的最多的类

6:if($n_l \bmod n_{\min} = 0$ and 不是所有的实例都属于同一类)

7:对于 $X_a \in X_l - \{X_\Phi\}$,对叶子节点 l 及其属性 X_a,根据式(2)、式(3)、式(5)及式(6)计算信息增益 $G_l(X_a)$,选择最大的属性 X_a 和次大的属性 X_b,选择信息熵最小的 W_t 作为最佳分裂点

8:计算 Hoeffding 值 $\varepsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n}}$

9:if($X_a \neq X_\Phi$ and 符合节点分裂准则即满足式(4))then

10:将叶子节点 l 转化为内部节点,属性 X_a 为该节点的决策属性

11:内部节点 l 生成两个新的叶子节点

12:end if

13:end if

14:end

3　实验及结果分析

为了验证 PPFDT 算法的隐私保护程度和在直接扰动的数据流上分类方面的性能,设计了两组实验,分别用于检测算法的运行时间和分类精度。算法参数设置为 $\delta = 10^{-7}$,$\tau = 5\%$,$n_{\min} = 200$。实验在 2.13 GHz,4 GB PC 上进行,操作系统为 Win7。

实验采用 UCI 的 WaveForm 数据集,原因是 UCI 提供了数据生成器,可以生成大规模的数据集,便于考察算法处理大规模数据的能力。选择 WaveForm 第一种版本的数据集,其包含 21 个连续数值型属性,取值范围是[0,6],类别取值个数为 3。使用 WaveForm 数据生成器生成 200 K(1 K=1 000 条)的数据作为训练数据,100 K 的数据为测试数据,然后使用文献[14]提出的方法,给每个属性添加高斯分布和均匀分布的随机噪声,当使用高斯分布的随机噪声时,知道其方差 σ^2 能在很大程度上影响结果,所以采用 4 种不同方差值的高斯分布的噪声数据,并且使用信噪比(Signal-to-Noise Ratio,SNR)来衡量噪声数据对实际数据的影响。2001 年,文献[15]提出一个基于微分熵的隐私度量来衡量隐私信息损失,采用该隐私度量来衡量算法的隐私保护程度。

表 1 显示了 5 种随机扰动数据集的 SNR 值和隐私度量。

表 1 不同数据集的 SNR 值和隐私度量

数据集	噪声分布	信噪比	隐私信息损失
Data1	Gaussian	1.7	0.207
Data2	Gaussian	1.3	0.187
Data3	Gaussian	1.0	0.154
Data4	Gaussian	0.5	0.103
Data5	Uniform	N/A	0.25

从表 1 可以看出,SNR 值越小,隐私信息损失越少,算法的隐私保护程度越高。

3.1 PPFDT 算法的准确度

针对初始数据集和扰动数据集,提出的 PPFDT 算法的分类准确度如表 2 所示。

表 2 PPFDT 算法的分类准确度

数据集	初始数据集上的准确度/%	扰动数据集上的准确度/%
Data1	79.14	75.09
Data2	78.69	75.14
Data3	75.63	73.41
Data4	76.01	75.03
Data5	79.29	79.52

使用 VFDT 算法在初始训练数据集上建立决策树,并用初始分类数据集分类的准确度为 81%,在扰动数据集上建立决策树,用初始数据分类的准确度低于 60%。

从表 2 可以得出,PPFDT 算法直接在大规模扰动数据集上建立的决策树在对初始数据集和扰动数据集分类时仍可以得到较精准的结果。

3.2 阈值对 PPFDT 算法准确度的影响

分析阈值在添加均匀分布噪声和高斯分布噪声的情况下,对 PPFDT 算法准确度的影响。添加均匀分布噪声和高斯分布噪声的算法准确度如图 1、图 2 所示。

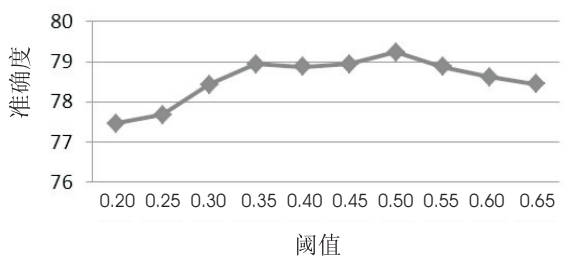

图 1 添加均匀分布噪声的数据集上 PPFDT 算法在不同阈值下的准确度

由图 1、图 2 可知:对于均匀分布,阈值设为 0.5,分类器有最高的准确度,而对于高斯分布,阈值设为 0.3,分类器有最高的准确度。

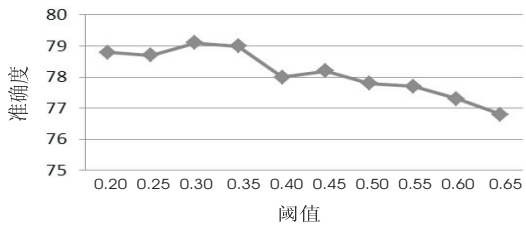

图 2 添加高斯分布噪声的数据集上 PPFDT 算法在不同阈值下的准确度

3.3 PPFDT 算法的执行时间

使用 WaveForm 数据生成器生成 100 K(1 K = 1 000)到 1 000 K 的数据作为训练数据,并使用同样添加均匀分布噪声的方法扰动初始数据集,并测试 PPFDT 算法在扰动数据集上建立决策树的时间,如图 3 所示。

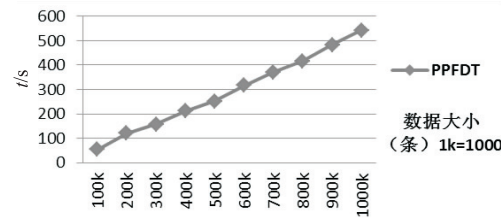

图 3 PPFDT 算法在不同数据量下建树的时间

4 结束语

文中研究了基于决策树的数据流分类算法的隐私保护问题。提出的方法采用数据扰动技术,改进经典的数据流挖掘算法—VFDT,并使用阈值算法找到扰动数据流的最佳分裂属性和最佳分裂点,从而直接在扰动数据流上建立决策树。基于该方法的算法实现和实验结果表明,提出的方法在快速准确地进行数据流分类的同时,具有一定的隐私保护程度。

参考文献:

[1] 李 光,王亚东,苏小红. 隐私保持的决策树分类挖掘[J]. 电子学报,2010,38(1):204-212.

[2] Agrawal R. Privacy-preserving data mining[C]//Eurographics/ACM SIGGRAPH symposium on geometry processing. [s. l.]:ACM,2000:97-115.

[3] Kargupta H,Datta S,Wang Q,et al. On the privacy preserving properties of random data perturbation techniques[C]//Third IEEE international conference on data mining. [s. l.]:IEEE,2003:99-106.

[4] Kargupta H,Datta S,Wang Q,et al. Random-data perturbation techniques and privacy-preserving data mining[J]. Knowledge and Information System,2005,7(4):387-414.

[5] 葛伟平,汪 卫,周皓峰,等. 基于隐私保护的分类挖掘[J]. 计算机研究与发展,2006,43(1):39-45.

[6] 杨晓春,刘向宇,王 斌,等. 支持多约束的 K-匿名化方法

(下转第 119 页)

参考文献:

[1] 翟　胜,师五喜,修春波. 基于模糊贝叶斯网的危害性分析方法[J]. 计算机应用,2014,34(12):3446-3450.

[2] 贾兴利,许金良. 基于云模型的地震区公路震害风险评估[J]. 同济大学学报:自然科学版,2014,42(9):1352-1358.

[3] 蒋明敏. 自媒体时代网络舆论风险的特点、成因及其治理[J]. 西南民族大学学报:人文社会科学版,2015,36(3):173-177.

[4] 王常柱,高晓宇. 网络舆论的民意属性及其诉求三维度—网络舆论民意属性的政治伦理审视[J]. 济南大学学报:社会科学版,2014,24(1):71-75.

[5] 曾润喜,杜换霞,王君泽. 网络舆情指标体系、方法与模型比较研究[J]. 情报杂志,2014,33(4):96-101.

[6] 屈正庚. 层次分析法在旅游评价体系中的研究[J]. 计算机技术与发展,2016,26(7):169-172.

[7] 申　楠,杨　琳. 复杂背景下网络舆论引导与网络环境治理探析[J]. 西安交通大学学报:社会科学版,2014,34(4):96-101.

[8] 张军玲. 基于层次分析法的企业网络舆情危机应对评价研究[J]. 经济数学,2015,32(3):60-63.

[9] 屈正庚. 层次分析法在大学生课堂上玩手机中的研究[J]. 系统仿真技术,2016,12(1):66-70.

[10] 郭正红,马辛华,兰安怡. 基于层次分析法权重和灰色服务器负载预测的云计算 on-line 迁移策略[J]. 计算机测量与控制,2015,23(3):1002-1004.

[11] 高　炜,张庆普,敦晓彪,等. 基于改进的可拓层次分析法和动态加权的航天高技术综合评价研究[J]. 系统工程与电子技术,2016,38(1):102-109.

[12] 屈正庚. 层次分析法在应用型人才培养体制中的研究[J]. 计算技术与自动化,2015,34(2):104-108.

[13] 范亚琼,燕雪峰,陈海燕. 基于改进离差最大化方法的梯形灰云评估模型[J]. 计算机技术与发展,2016,26(4):20-24.

[14] Wu Liang. On regulation to hostile environment sexual harassment speech in the University of U. S. A[J]. Comparative Education Review,2015,306(7):51-56.

[15] Anneliese A. The use of Popular Opinion Leader (POL) groups and the reduction of "gay bullying" in middle school:a case study inquiry of group leader experiences[J]. Journal for Specialists in Group Work,2013,38(3):184-206.

[16] Friedman A L,Oruko K O,Habel M A. Preparing for human papillomavirus vaccine introduction in Kenya: implications from focus-group and interview discussions with caregivers and opinion leaders in Western Kenya[J]. BMC Public Health,2014,14(1):855-858.

[17] Habel P D. The dynamics of influence among media opinion, the public,and politicians[J]. Political Communication,2012,29(3):257-277.

[18] Dougherty T. Freedom of expression and the internet[M]. [s. l.]:Lucent Books,2010.

(上接第 114 页)

[J]. 软件学报,2006,17(5):1222-1231.

[7] 韩建民,岑婷婷,虞慧群. 数据表 k-匿名化的微聚集算法研究[J]. 电子学报,2008,36(10):2021-2029.

[8] Sharma V. Methods for privacy protection using k-anonymity[C]//International conference on optimization,reliability,and information technology. [s. l.]:IEEE,2014:149-152.

[9] Liu L,Kantarcioglu M,Thuraisingham B. Privacy preserving decision tree mining from perturbed data[C]//42nd Hawaii international conference on system sciences. Hawaii:IEEE,2009:1-10.

[10] 王　涛,李舟军,颜跃进,等. 数据流挖掘分类技术综述[J]. 计算机研究与发展,2007,44(11):1809-1815.

[11] Matuszyk P,Krempl G,Spiliopoulou M. Correcting the usage of the hoeffding inequality in stream mining[C]//International symposium on intelligent data analysis. Berlin:Springer,2013:298-309.

[12] 王　涛,李舟军,胡小华,等. 一种高效的数据流挖掘增量模糊决策树分类算法[J]. 计算机学报,2007,30(8):1244-1250.

[13] 蒋良孝,蔡之华,刘　钊. 一种基于信息增益的分类规则挖掘算法[J]. 中南大学学报:自然科学版,2003,34(z1):69-71.

[14] Agrawal R,Srikant R. Privacy-preserving data mining[J]. ACM SIGMOD Record,2000,29(2):439-450.

[15] Agrawal D,Aggarwal C C. On the design and quantification of privacy preserving data mining algorithms[C]//Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems. [s. l.]:ACM,2001:247-255.