朴素贝叶斯分类中的隐私保护方法研究

张 鹏[1)] 唐世渭[2)]

[1)](中国电信股份有限公司北京研究院 北京 100035)

[2)](北京大学信息科学技术学院 北京 100871)

摘 要 数据挖掘中的隐私保护方法,试图在不精确访问原始数据详细信息的条件下,挖掘出准确的模式与规则.围绕着分类挖掘中的隐私保护问题展开研究,给出了一种基于数据处理和特征重构的朴素贝叶斯分类中的隐私保护方法.分别提出了一种针对枚举类型的隐私数据处理与特征重构方法——扩展的部分隐藏随机化回答(Extended Randomized Response with Partial Hiding,ERRPH)方法和一种针对数值类型的隐私数据处理与特征重构方法——转换的随机化回答(Transforming Randomized Response,TRR)方法,并在此基础上实现了一个完整的隐私保护的朴素贝叶斯分类算法.理论分析和实验结果均表明:朴素贝叶斯分类中基于 ERRPH 和 TRR 的隐私保护方法具有很好的隐私性、准确性、高效性和适用性.

关键词 数据挖掘;隐私保护;朴素贝叶斯分类;随机处理;特征重构

中图法分类号 TP311

Privacy Preserving Naive Bayes Classification

ZHANG Peng[1)] TANG Shi-Wei[2)]

[1)](*Beijing Research Institute, China Telecom Corporation Limited, Beijing* 100035)

[2)](*School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871)

Abstract Privacy preserving data mining is to discover accurate patterns without precise access to the original data. This paper focuses on privacy preserving classification, and presents a privacy preserving Naive Bayes classification approach based on data randomization and feature reconstruction. An ERRPH (Extended Randomized Response with Partial Hidding) method and a TRR (Transforming Randomized Response) method are respectively presented for enumerated data and numerical data. Then, a privacy preserving Naive Bayes classification algorithm is implemented based on those methods. Theoretical analyses show that it can provide better privacy, accuracy, efficiency, and applicability. The effectiveness is also verified by experiments.

Keywords data mining; privacy preservation; Naive Bayes classification; data randomization; feature reconstruction

1 引 言

随着信息技术,特别是网络技术、数据存储技术和高性能处理器技术的飞速发展,海量数据的收集、管理和分析变得越来越方便.包括分类挖掘在内的各种数据挖掘技术,更是在一些深层次的应用中发挥了非常积极的作用.但与此同时,也带来了隐私保护方面的诸多问题.例如,通过对电信客户的基本信息和消费行为数据进行挖掘,可以预测新增客户的

收稿日期:2007-03-05;修改稿收到日期:2007-05-22.本课题得到国家自然科学基金(60403041)和北京市科学技术委员会博士论文专项基金(ZZ6027)资助.**张 鹏**,男,1978 年生,博士,工程师,主要研究方向为数据仓库、数据挖掘、联机分析处理. E-mail: zhangpeng@ctbri.com.cn. **唐世渭**,男,1939 年生,教授,博士生导师,主要研究领域为数据模型、数据库系统、数据仓库、数据挖掘.

服务偏好和价值走向,但在使用一般方法进行挖掘的过程中,不可避免地会使客户数据暴露,从而造成客户隐私的泄露.于是,如何在数据挖掘的过程中解决好隐私信息保护的问题,成为了数据挖掘界的一个研究热点[1].数据挖掘有一个重要特征,就是从大量数据中挖掘出来的模式或者规则,通常是针对综合数据而非细节数据.那么,能否基于非精确的原始数据而抽取出准确的模式与规则呢?实现隐私数据的合理保护和基于统计数据的模式抽取两者兼得,正是数据挖掘中隐私保护方法研究的出发点和最终目标.

数据挖掘中隐私保护的概念是由 Agrawal 等人[2]提出的,他们专门针对决策树分类问题,采用随机变换[3]的数据干扰策略,实现了数值类型数据的变换和重构以及隐私保护的决策树分类算法;在此基础上,Agrawal 等人又提出了一种基于期望最大化的分布重构方法[4].但是,这些方法都不能处理布尔类型和枚举类型的数据.之后,Du 等人针对布尔类型的数据,提出了一种决策树分类中基于随机化回答的隐私保护方法[5],但该方法与此前提出的随机处理方法[6]一样,利用的都是统计学中的沃纳模型,不仅由于变换后的所有数据都与真实的原始数据直接相关,而使得对隐私信息的保护程度不高,并且随机化参数的选择也受到限制,必须偏离 0.5.与此同时,Johnsten 等人还提出了利用数据隐藏策略,实现隐私保护的分类方法[7],虽然一部分敏感信息得到了保护,但由于数据所有者提供的数据全部都是真实数据,所以此类方法对整个数据集的隐私保护程度并不理想.Zhang 等人则将数据干扰和查询限制的隐私保护策略相结合,提出了一种部分隐藏的随机化回答(Randomized Response with Partial Hiding,RRPH)方法[8],进行数据的变换和隐藏;然后针对经过 RRPH 方法处理的数据,设计了一种简单、高效的频繁项集生成算法,进而实现了关联规则挖掘中的隐私保护.与之相比,分类算法所要处理的数据类型更加广泛,不但包括布尔类型的数据,而且还包括枚举类型和数值类型的数据.在分布式环境下的隐私保护方法中,文献[9-11]分别通过安全多方计算协议,实现了决策树构建中的隐私保护方法;Vaidya 等人则相继提出了朴素贝叶斯分类中基于水平数据划分[12]和垂直数据划分[13]的隐私保护方法;Clifton 等人还提供了一组分布式环境下,支持数据挖掘中隐私保护计算的工具集[14].但是,这些方法都只能用于分布式数据库,并且所有的数据提供者都必须参与到计算中来,单点故障就会导致错误结果的产生,甚至造成整个挖掘的无法进行.

可以看出,分类作为数据挖掘中一种重要的方法,目前在隐私保护方面的研究大多数都是针对决策树的构建方法展开的,包括朴素贝叶斯在内的其它分类方法只在分布式环境下有少量的应用.然而,在属性相互独立的条件下,朴素贝叶斯分类具有实现简单、结果准确的特点,还可以针对枚举类型的数据实现增量挖掘.于是,本文紧紧围绕分类挖掘中的隐私保护问题展开研究,给出了一个基于数据处理和特征重构的朴素贝叶斯分类中的隐私保护方法,主要研究内容和创新成果包括:(1)提出了一种扩展的部分隐藏随机化回答(ERRPH)方法,实现了枚举类型隐私数据的处理和特征重构;(2)提出了一种转换的随机化回答(TRR)方法,实现了数值类型隐私数据的处理和特征重构;(3)在 ERRPH 方法和 TRR 方法的基础上,实现了一个完整的隐私保护的朴素贝叶斯分类算法;(4)通过理论分析和实验结果,说明了朴素贝叶斯分类中基于 ERRPH 和 TRR 的隐私保护方法具有很好的隐私性、准确性、高效性和适用性.

本文第 2 节首先给出分类挖掘中隐私保护问题的描述,然后提出解决方法的总体架构;第 3 节提出一种 ERRPH 方法,实现枚举类型数据的随机处理和特征重构;第 4 节再提出一种 TRR 方法,实现数值类型数据的随机处理和特征重构;第 5 节说明如何使用经过 ERRPH 和 TRR 方法处理得到的发布数据集,构造朴素贝叶斯分类器,预测未知样本的类标号;对方法的详细分析和评价以及同原有方法的比较在第 6 节中给出;第 7 节中展示相关的实验结果;最后的第 8 节给出全文的总结.

2 问题与架构

本节中,我们将首先给出分类挖掘中隐私保护问题的描述,然后提出解决方法的总体架构.

2.1 问题描述

给定的数据集 D 包含 n 个属性,分别记作 A_1,A_2,…,A_n,且 $A_i(i=1,2,\cdots,n)$的值域为 $dom(A_i)$.D 中的每个样本用一个 n 维特征向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$表示,其中 $x_i\in dom(A_i)$,$i=1,2,\cdots,n$,分别表示该样本中属性 A_i的取值.而且数据集 D 中的全体样本被分成 m 个类,并分别用 $C_1,C_2,\cdots,C_m$

进行标记.那么,给定一个未知的数据样本 Z(类标号未知),分类的目的就是要根据数据集 D 来预测出样本 Z 的类标号.而分类挖掘中的隐私保护问题,就是要在不精确访问原始数据集 D 的条件下,尽可能准确地预测出 Z 的类标号.

2.2　总体架构

为解决上述问题,本文提出了一种朴素贝叶斯分类中的隐私保护方法.该方法基于 KD^3(Knowledge Discovery in Distorted Database)架构[15],分两阶段进行:首先用随机化回答的方法对原始数据进行处理;然后再利用特征重构的方法,借助经过处理的数据来对未知样本进行朴素贝叶斯分类.特别要指出的是,我们会根据不同的数据类型来选择具体的随机处理和特征重构方法.整个方法的总体架构如图 1 所示.

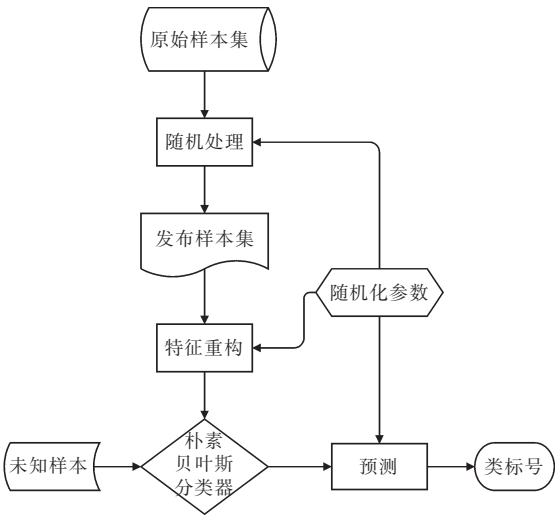


图 1　朴素贝叶斯分类中基于 ERRPH 和 TRR 的隐私保护方法

第一个阶段针对枚举类型和数值类型的数据,分别利用 ERRPH 方法和 TRR 方法来对原始数据进行处理,具体的方法在第 3 和第 4 节中给出.第二个阶段则将基于经过 ERRPH 方法和 TRR 方法处理的发布数据,对原始数据的分布特征进行重构,以此估算未知样本的后验概率,进而预测出分类标号,详细的过程在第 5 节中介绍.整个方法中,连接两个阶段操作的除了经过 ERRPH 方法和 TRR 方法处理的数据以外,还有一组随机化参数,在第一个阶段描述随机处理的规则;在第二个阶段指导分布的重构和后验概率的计算.而且,参数值的选择对于隐私的保护程度和挖掘结果的准确性,都具有非常重要的影响.

3　扩展的部分隐藏随机化回答(ERRPH)方法

本节将提出一种 ERRPH 方法,实现枚举类型隐私数据的处理和特征重构.

3.1　数据处理

随机化回答方法的基本思想是:数据提供者依照给定的随机化参数对原始数据进行变换,再提供给数据使用者.在使用者得到的信息中,虽然详细信息被提供者进行了处理,但在数据量比较大的情况下,统计信息和聚集信息仍然可以被相当精确地估算出来.由于样本分类,特别是朴素贝叶斯分类,是基于一个数据集合的聚集信息值,而非一个个详细的数据项,因此随机化回答方法可以很好地用于朴素贝叶斯分类中的隐私保护问题.本节提出了一种处理枚举类型隐私数据的 ERRPH 方法,其具体描述如下:

设属性 A 有 k 个可能的取值,其值域 $dom(A)=\{a_1,a_2,\cdots,a_k\}$.给定一组随机化参数 $0\leq p_0,p_1,\cdots,p_k\leq 1$,使得 $p_0+p_1+\cdots+p_k=1$.对任意的 $x\in dom(A)$,令 $r_0=x$, $r_j=a_j$ $(j=1,2,\cdots,k)$,则随机化函数 $r(x)$ 以 p_j $(j=0,1,\cdots,k)$ 的概率返回 r_j.数据集 D 中的样本可以用向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$ 来表示,其中 $x_i\in A_i$,且 A_i 的值域 $dom(A_i)=\{a_{i1},a_{i2},\cdots,a_{ik_i}\}$.于是,随机处理后的样本 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)$ 可以通过 $\boldsymbol{Y}=R(\boldsymbol{X})$ 计算得到,其中 $y_i=r(x_i)$.也就是说,y_i 以 p_0 的概率取值为 x_i,而以 p_j 的概率取值为 a_{ij}.

这样,对数据集 D 中的每一个样本 $\boldsymbol{X}$,经过随机化函数 R 的处理得到 $\boldsymbol{Y}$,而且由于 $\boldsymbol{Y}$ 在形式上仍然是一个与 $\boldsymbol{X}$ 相似的数据样本,所以就可以作为一个伪造的样本,加入到伪造的数据集 D' 中.

需要说明的是,在 ERRPH 方法中,我们可以对不同的属性分别使用不同的随机化参数进行干扰.但为了简单起见,我们将对所有属性使用相同的随机化参数进行数据处理.

3.2　特征重构

那么,如何利用经过 ERRPH 方法处理得到的发布数据集 D' 来重构原始数据集 D 中属性的分布情况呢?设 π_j 表示原始数据集 D 中属性 A 的取值为 a_j 的样本所占的比例;λ_j 表示发布数据集 D' 中属性 A 的取值为 a_j 的样本所占的比例,其中 $j=1,2,\cdots,k$. D 中的样本 $\boldsymbol{X}$ 经过 ERRPH 方法处理,变

成 D' 中的样本 $\boldsymbol{Y}$，而 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 对应属性 A 的取值分别用 $(a_1,a_2,\cdots,a_k)$ 表示，则 $\boldsymbol{X}_A$ 和 $\boldsymbol{Y}_A$ 的取值与映射概率如表 1 所示.

表 1 ERRPH 方法处理数据的映射概率

	映射概率					
	a_1	a_2	$\cdots$	a_i	$\cdots$	a_k
a_1	p_0+p_1	p_2	$\cdots$	p_i	$\cdots$	p_k
a_2	p_1	p_0+p_2	$\cdots$	p_i	$\cdots$	p_k
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
a_i	p_1	p_2	$\cdots$	p_0+p_i	$\cdots$	p_k
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
a_k	p_1	p_2	$\cdots$	p_i	$\cdots$	p_0+p_k

由表 1 可得 $\lambda_i=p_0\cdot\pi+p_j$. 于是，

$$\pi_j=\frac{\lambda_j-p_j}{p_0} \tag{1}$$

也就是说，我们可以首先计算出发布数据集 D' 中属性 A 的取值为 a_j 的样本所占的比例 λ，然后再利用式(1)估算出原始数据集 D 中属性 A 的取值为 a_j 的样本所占的比例 π.

为了简单起见，我们通常会假设除了 p_0 以外的 k 个参数都相等，也就是说，$p_1=p_2=\cdots=p_k=\frac{1-p_0}{k}$. 这样，在原始数据集中对应属性 A 的不同取值的样本就可以通过相同的式(1)计算得到.

4 转换的随机化回答(TRR)方法

本节将基于统计学中的转换模型，提出一种针对数值类型数据的 TRR 随机处理与特征重构方法.

4.1 数据处理

转换的随机化回答(TRR)方法，首先需要随机生成一个满足给定特征的变换函数，然后利用这个函数对原始的数值类型数据进行变换，并将变换后的数值作为随机化回答的结果. 方法的具体描述如下：

给定集合 $A=\{a_1,a_2,\cdots,a_l\}$，且设其中元素的均值和方差分别为

$$\overline{A}=\frac{1}{l}\sum_{i=1}^{l}a_i,\quad \sigma_A^2=\frac{1}{l-1}\sum_{i=1}^{l}(a_i-\overline{A})^2;$$

再给定集合 $B=\{b_1,b_2,\cdots,b_m\}$，且设其中元素的均值和方差分别为

$$\overline{B}=\frac{1}{m}\sum_{i=1}^{m}b_i,\quad \sigma_B^2=\frac{1}{m-1}\sum_{i=1}^{m}(b_i-\overline{B})^2.$$

对数值类型的数据 x，随机化函数 $r(x)=ax+b$，其中 $a\in A$，是从 A 中随机选取的一个元素，而 $b\in B$，是从 B 中随机选取的一个元素. 这样，变换后的数据就可以通过 $y=r(x)$ 计算得到. 设原始数据集 D 中包含 n 条记录，这 n 条记录所对应的属性 At 的取值分别为 $X=\{x_1,x_2,\cdots,x_n\}$. 那么，发布数据集 D' 中的 n 条记录所对应的属性 At 的取值 $Y=\{y_1,y_2,\cdots,y_n\}$ 就可以通过 $Y=R(X)$ 计算得到，其中 $y_i=r(x_i)$，$i=1,2,\cdots,n$.

4.2 特征重构

按照上述的 TRR 方法，我们将原始数据 x 转换成 $y=r(x)=ax+b$. 当 $\overline{A}\neq 0$ 时，令 x 估计量 $\hat{x}=\frac{y-\overline{B}}{\overline{A}}$，且 $\hat{x}$ 是 x 的无偏估计量. 原始数据集 D 中的 n 条记录所对应的属性 At 的取值 $X=\{x_1,x_2,\cdots,x_n\}$，被转换为了发布数据集 D' 中的 n 条记录所对应的属性 At 的取值 $Y=\{y_1,y_2,\cdots,y_n\}$.

为了进行朴素贝叶斯分类，我们假设 X 符合高斯分布 $N(\mu_x,\sigma_x^2)$. 我们知道，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 是 μ_x 的无偏估计量. 但 x_i 是未知的，故我们用 $\hat{x}_i$ 代替 x_i，得到估计量 $\bar{\hat{x}}=\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i-\overline{B}}{\overline{A}}=\frac{\bar{y}-\overline{B}}{\overline{A}}$.

通常，为了计算简便，可以设计 $\overline{B}=0$，进而得到 x 均值的估计值

$$\bar{\hat{x}}=\frac{\bar{y}}{\overline{A}}=\frac{1}{n\overline{A}}\sum_{i=1}^{n}y_i \tag{2}$$

另一方面，$Var(y)=Var(ax+b)=(\sigma_A^2+\overline{A}^2)\cdot Var(x)+\sigma_A^2\cdot\bar{x}^2+\sigma_B^2$. 因此，

$$Var(x)=\frac{Var(y)-\sigma_A^2\cdot\bar{x}^2-\sigma_B^2}{\sigma_A^2+\overline{A}^2}.$$

但 $\bar{x}$ 未知，故用 $\bar{\hat{x}}=\frac{\bar{y}}{\overline{A}}$ 来代替. 于是得到 x 的方差估计值

$$\hat{S}^2=\hat{Var}(x)=\frac{\hat{Var}(y)-\sigma_A^2\cdot\bar{\hat{x}}^2-\sigma_B^2}{\sigma_A^2+\overline{A}^2}$$

$$=\frac{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2-\frac{\sigma_A^2\cdot\bar{y}^2}{\overline{A}^2}-\sigma_B^2}{\sigma_A^2+\overline{A}^2} \tag{3}$$

这样，就可以利用集合 A 中元素的均值 $\overline{A}$ 和方差 σ_A^2 以及集合 B 中元素的方差 σ_B^2，并通过发布数据集 D' 中属性 At 的取值 $Y=\{y_1,y_2,\cdots,y_n\}$，来根据式(2)和式(3)，计算出 $\bar{\hat{x}}$ 和 $\hat{S}^2$ 的取值. 然后，以此作为原始数据集 D 中属性 At 的取值 $X=\{x_1,x_2,\cdots,x_n\}$ 的分布特征 $N(\bar{\hat{x}},\hat{S}^2)$，进而实施分类挖掘.

5　隐私保护的朴素贝叶斯分类算法

本节将在一般的朴素贝叶斯分类算法的基础上,利用基于 ERRPH 的枚举类型数据特征重构方法和基于 TRR 的数值类型数据特征重构方法,给出通过发布数据集 D'来对未知样本 Z 进行分类的算法.

朴素贝叶斯作为一种简单实用的分类方法,将预测未知样本 Z 属于具有最高后验概率(条件 Z)的类.即将未知的样本分配给类 C_i,当且仅当 $P(C_i|Z)\geq P(C_j|Z)$,其中 $1\leq j\leq m$ 且 $j\neq i$. 这样,我们最大化 $P(C_i|Z)$ 的值,而使得 $P(C_i|Z)$ 最大的类 C_i 则称为最大后验假定. 根据贝叶斯定理,有 $P(C_i|Z)=\frac{P(Z|C_i)P(C_i)}{P(Z)}$. 由于 $P(Z)$ 对于所有类为常数,故只需让 $P(Z|C_i)P(C_i)$ 的取值最大即可. 于是,朴素贝叶斯分类算法的核心内容就是要计算和比较 $P(Z|C_i)P(C_i)$ 的值,其中 $i=1,2,\cdots,m$.

在经过 ERRPH 方法和 TRR 方法随机处理所得到的发布数据集 D'中,因为类标号是没有进行处理的,所以类的先验概率仍然可以用 $P(C_i)=\frac{s_i}{s}$ 计算,其中 s_i 是类 C_i 中的训练样本数,而 $s=|D|$ 是训练样本总数. 在属性相互独立的条件下,$P(Z|C_i)=\prod_{k=1}^{n}P(z_k|C_i)$. 但 $P(z_1|C_i),P(z_2|C_i),\cdots,P(z_n|C_i)$ 的值无法通过训练样本直接计算得到,而需要借助 ERRPH(对枚举类型数据)和 TRR(对数值类型数据)的特征重构方法.

(1) 如果 A_k 是枚举类型的属性,则设 $P'(z_k|C_i)=\frac{s'_{ik}}{s_i}$,其中 s'_{ik} 是在发布数据集 D'中属性 A_k 的取值为 x_k,且属于类 C_i 的训练样本数,而 s_i 是属于类 C_i 中的训练样本数. 根据式(1)可知

$$P(z_k|C_i)=\frac{P'(x_k|C_i)-p_2}{p_1}=\frac{\frac{s'_{ik}}{s_i}-p_2}{p_1} \tag{4}$$

(2) 如果 A_k 是数值类型的属性,则设发布数据集 D'中属于类 C_i 的训练样本数为 N_i,分别记作 y_1,y_2,$\cdots$,y_{Ni},它们的均值为 $\bar{y}_{ik}$,而样本 y_j 在属性 A_k 上的取值为 y_{jk}. 分别根据式(2)和式(3),我们可以计算得到

$$\hat{\mu}_i=\frac{1}{N_i\cdot\bar{A}}\cdot\sum_{j=1}^{N_i}y_{jk},$$

$$\hat{\sigma}_i^2=\frac{\frac{1}{N_i-1}\sum_{j=1}^{N_i}(y_{jk}-\bar{y}_{ik})^2-\frac{\sigma_A^2\cdot\bar{y}_{ik}^2}{\bar{A}^2}-\sigma_B^2}{\sigma_A^2+\bar{A}^2}.$$

于是可知

$$P(z_k|C_i)=g(z_k,\hat{\mu}_i,\hat{\sigma}_i)=\frac{1}{\sqrt{2\pi}\hat{\sigma}_i}e^{-\frac{(z_k-\hat{\mu}_i)^2}{2\hat{\sigma}_i^2}} \tag{5}$$

这样,$P(Z|C_i)P(C_i)$就可以通过式(4)和式(5)计算得到. 对未知样本 Z 分类时,就可以使用经过 ERRPH 方法和 TRR 方法随机处理过的发布数据集 D',来对每个类 C_i,计算 $P(Z|C_i)P(C_i)$ 的值. 并将未知样本 Z 分到类 C_i,当且仅当 $P(C_i|Z)\geq P(C_j|Z)$,其中 $1\leq j\leq m$ 且 $j\neq i$.

6　分析与评价

本节将从隐私性、准确性、高效性和适用性这 4 个方面,对本文提出的朴素贝叶斯分类中基于 ERRPH 和 TRR 的隐私保护方法进行详细的分析和评价,并与原有的隐私数据决策树构建(Building Decision Tree on Private Data,DTPD)方法以及基于随机化数据的决策树分类(Decision Tree Classification over Randomized Data,DTCRD)方法进行对比.

6.1　隐私性

数据挖掘中隐私保护方法研究的出发点和最终目标,是在合理保护隐私数据的前提下,进行数据挖掘和知识发现,寻找其中潜在的模式和规则. 本文提出的朴素贝叶斯分类中的隐私保护方法是基于 KD[3] 架构设计的,分两个阶段进行. 而隐私性则主要是在第一个阶段,即利用 ERRPH 方法和 TRR 方法进行随机处理的过程中实现的.

在处理枚举类型数据的时候,使用的是 ERRPH 方法. 数据使用者得到 $y=r(x)$,而且 y 和 x 的取值范围是相同的,都是 $dom(A)=\{a_1,a_2,\cdots,a_k\}$. 对于给定的 y 值来讲,x 可以是 $dom(A)$ 中的任意一个取值,因而无法通过 y 值来还原 x 的取值,也就是说隐私信息 x 得到了保护. 在发布数据集中,比例为 $1-p_0$ 的数据被隐藏,只有比例为 p_0 的真实数据混杂在大量的伪造数据中被提交给了数据使用者. 并且由于每条记录所对应的随机化参数都是未知的,因而这些真实数据是很难被识别出来的. 即使是在随机化参数被泄露的情况下,也至多暴露比例为 p_0 的真实数据. 这样,ERRPH 方法既克服了基于数据干扰策略的处理方法中,所有变化后的数据均与

真实的原始数据直接相关的不足，又消除了基于查询限制策略的处理方法中，所有提供的都是真实原始数据的缺陷，从而实现了对隐私信息的很好保护.

为了比较不同方法对隐私信息的保护程度，我们提出了一个隐私破坏系数 $Breach$ 的指标. 它是指原始数据被重构或者预测出来的概率，其定义如下：

$$Breach = P_{\text{真实数据的比例}} \times P_{\text{真实数据被识别出的概率}} + P_{\text{非真实数据的比例}} \times P_{\text{非真实数据被识别出的概率}} \times P_{\text{非真实数据被还原的概率}}.$$

于是，ERRPH 方法的隐私破坏系数为 $Breach = p_0 \cdot \frac{p_0}{p_0 + p_i}$. 当除 p_0 之外的其余 k 个随机化参数均相等，即 $p_1 = p_2 = \cdots = p_k = \frac{1-p_0}{k}$ 时，$Breach = \frac{p_0^2}{p_0 + \frac{1-p_0}{k}} = \frac{k \cdot p_0^2}{(k-1)p_0 + 1}$. 如果 $k=2$，则 ERRPH 方法退化成 RRPH 方法，此时 $Breach = \frac{2p_0^2}{p_0+1}$. 因为 DTPD 方法使用的是基于沃纳模型的随机化回答方法，所以根据文献[8]中的分析结果可知：当 $0 < p < \frac{1}{\sqrt{2}}$ 时，该方法比 DTPD 方法具有更高的隐私保护程度.

在处理数值类型数据的时候，使用的是 TRR 方法，这是基于统计学中的转换模型构建的. 数据使用者得到的是 $y = r(x) = ax + b$，其中 a 和 b 是分别从集合 A 与集合 B 中随机选取的元素. 由于我们只知道集合 A 与集合 B 中的元素分别服从给定均值和方差的高斯分布，但并不知道它们具体可能的取值，因而无法通过 y 值来还原 x 的取值，也就是说隐私信息 x 得到了保护.

为了能够评价类似方法的隐私保护程度，我们引入了一个隐私破坏区间宽度 $BreachWidth$ 的指标. 它的定义是：如果原始数据 x 落到区间 $[x_1, x_2]$ 上的概率为 c，即 $P(x_1 \le x \le x_2) = c$，则称区间 $[x_1, x_2]$ 是置信度为 c 的隐私破坏区间，而该区间的宽度 $(x_2 - x_1)$ 就定义了置信度为 c 的隐私破坏区间宽度，即

$$BreachWidth = x_2 - x_1.$$

由于 X 服从高斯分布 $N(\mu, \sigma^2)$，所以 TRR 方法在不同置信度条件下的隐私破坏区间宽度如表 2 所示. 可以看出，TRR 方法的隐私破坏区间宽度只和 X 所服从的分布的标准差有关，并且随着置信度的增加而不断变大. 另外，TRR 方法的隐私破坏区间宽度与 DTCRD 方法中使用高斯分布进行干扰（这也是隐私保护程度最高的一种方式）时的隐私破坏区间宽度是一致的，而比使用其它干扰方式（包括离散化和均值分布干扰）进行隐私信息保护时的效果要好.

表 2 不同置信度的隐私破坏区间宽度

置信度(c)/%	隐私破坏区间宽度($BreachWidth$)
30	$0.78 \cdot \sigma$
50	$1.34 \cdot \sigma$
70	$2.08 \cdot \sigma$
90	$3.28 \cdot \sigma$
95	$3.92 \cdot \sigma$
97	$4.34 \cdot \sigma$
99	$5.16 \cdot \sigma$
99.9	$6.58 \cdot \sigma$
99.99	$7.8 \cdot \sigma$

6.2 准确性

保护好数据的隐私，是数据挖掘中隐私保护方法的最基本要求，而我们的最终目标是要通过挖掘来获取真实、可用的知识与规则.

对于 ERRPH 方法，由式(1)得到的估计量 $\hat{\pi}_i = \frac{\lambda_i - p_i}{p_0}$ 是 π_j 的无偏估计量；方差 $Var(\hat{\pi}_i) = \frac{1}{p_0^2} Var(\lambda_i)$.

设数据集中 D 中的样本总数为 N，则

$$Var(\lambda_i) = \frac{\lambda_i(1-\lambda_i)}{N} = \frac{p_i(1-p_i) + (p_0 - 2p_0p_i)\pi_i - p_0^2\pi_0^2}{N}.$$

所以，

$$Var(\hat{\pi}_i) = \frac{p_i(1-p_i) + p_0(1-p_0-2p_i)\pi_i}{Np_0^2} + \frac{\pi_i(1-\pi_i)}{N}.$$

忽略抽样误差，

$$Var(\hat{\pi}_i) = \frac{p_i(1-p_i) + p_0(1-p_0-2p_i)\pi_i}{Np_0^2}.$$

当 $p_1 = p_2 = \cdots = p_k = \frac{1-p_0}{k}$ 时，$Var(\hat{\pi}_i) = \frac{1-p_0}{Nk^2p_0^2}[k - 1 + p_0 + k(k-2)p_0\pi_i]$. 如果 $k=2$，则 $Var(\hat{\pi}) = \frac{(1-p)(1+p)}{4Np^2}$.

对于数值类型的数据，使用的是 TRR 方法. 为了进行朴素贝叶斯分类，假设 X 符合高斯分布 $N(\mu_x, \sigma_x^2)$.

我们知道，$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 是 μ_x 的无偏估计量. 但 x_i 是未知的，因而用 $\hat{x}_i$ 代替 x_i，得到估计量 $\bar{\hat{x}} =$

$\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i$,也是 μ_x的无偏估计量. 于是,

$$Var(\bar{\hat{x}})=Var\left(\frac{1}{n}\sum_{i=1}^{n}\frac{y_i-\bar{B}}{\bar{A}}\right)$$

$$=\frac{\sum_{i=1}^{n}Var(y_i)}{n^2\cdot\bar{A}^2}=\frac{\sigma_A^2\cdot\overline{x^2}+\sigma_B^2}{n\cdot\bar{A}^2}.$$

因为其中的 x 未知,所以要计算 $Var(\bar{\hat{x}})$的估计量 $\hat{Var}(\bar{\hat{x}})=\frac{1}{n^2}\sum_{i=1}^{n}\frac{\sigma_A^2\cdot\hat{x}_i^2+\sigma_B^2}{\bar{A}^2+\sigma_A^2}$. 由 $E[\hat{Var}(\bar{\hat{x}})]=\frac{1}{n}\cdot\frac{\sigma_A^2\cdot E(\hat{x}^2)+\sigma_B^2}{\bar{A}^2+\sigma_A^2}=Var(\bar{\hat{x}})$,知 $\hat{Var}(\bar{\hat{x}})$是 $Var(\bar{\hat{x}})$的无偏估计量. 与此同时,由于$\hat{Var}(\bar{\hat{x}})$中不包含 $\bar{B}$. 因此可以设计 $\bar{B}=0$. 此时,$\hat{x}_i=\frac{y_i}{\bar{A}}$,进而得到$\bar{\hat{x}}=\frac{\sum_{i=1}^{n}y_i}{n\cdot\bar{A}}$,

$$\hat{Var}(\bar{\hat{x}})=\frac{1}{n}\frac{\sigma_A^2\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\bar{A}}\right)^2+\sigma_B^2}{\bar{A}^2+\sigma_A^2}=\frac{\sigma_A^2\frac{1}{n}\sum_{i=1}^{n}y_i^2+\bar{A}^2\sigma_B^2}{n\bar{A}^2(\bar{A}^2+\sigma_A^2)}.$$

另一方面,

$$E(\hat{S}^2)=E\left[\frac{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}-\frac{\sigma_A^2\bar{y}^2}{\bar{A}^2}-\sigma_B^2}{\sigma_A^2+\bar{A}^2}\right]$$

$$=E\left[\frac{\hat{Var}(y)-\sigma_A^2\left(\frac{\bar{y}}{\bar{A}}\right)^2-\sigma_B^2}{\sigma_A^2+\bar{A}^2}\right]$$

$$=Var(x)+\frac{\sigma_A^2\bar{x}^2-\sigma_A^2E^2(\bar{\hat{x}})}{\sigma_A^2+\bar{A}^2}=Var(x).$$

也就是说,$\hat{S}^2$是 $Var(x)$的无偏估计量.

我们记 $\beta=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2-\frac{\sigma_A^2\cdot\bar{y}^2}{\bar{A}^2}$,则 $E(\beta)=\frac{n\overline{y^2}}{n-1}-\left(\frac{n}{n-1}+\frac{\sigma_A^2}{\bar{A}^2}\right)\bar{y}^2$. 于是,

$$Var(\hat{S}^2)=\frac{Var(\beta)}{(\sigma_A^2+\bar{A}^2)^2}=\frac{E(\beta^2)-E^2(\beta)}{(\sigma_A^2+\bar{A}^2)^2}$$

$$=\frac{n[\overline{y^4}+(n-1)\overline{y^2}-n\cdot\overline{y^2}^2]}{(n-1)^2(\sigma_A^2+\bar{A}^2)^2}.$$

可以看出,在 TRR 方法中,原始数据方差估计值的误差会随着集合 A 中元素均值和方差的增大而增加. 在下一节中,我们将通过实验,对于不同方法的挖掘结果准确性进行详细的分析比较.

6.3　高效性

与普通数据挖掘方法相比,隐私保护的数据挖掘方法面对着更加复杂的要求. 本文提出的朴素贝叶斯分类中基于 ERRPH 和 TRR 的隐私保护方法,在对隐私信息进行合理保护和对未知样本做出较准确分类的同时,并没有带来过多的计算复杂度增长,其时间和空间代价与一般的朴素贝叶斯分类方法成线性关系.

原有的 DTPD 方法和 DTCRD 方法都是决策树构建中的隐私保护方法. 对于 DTPD 方法来讲,其计算的复杂度与一般的决策树分类方法成线性关系;而对于 DTCRD 方法来讲,会由于在分布重构过程中所进行的大量迭代计算,而使得计算的复杂度大幅度增加.

6.4　适用性

本文提出的朴素贝叶斯分类中的隐私数据保护方法,还具有非常好的适用性. 具体体现在数据挖掘算法的适用性、隐私数据类型的适用性和数据分布状况的适用性这三个方面.

首先,从数据挖掘算法的角度来看,本文提出的朴素贝叶斯分类中的隐私保护方法,是在 KD^3 通用架构与流程的基础上设计的,将整个方法分成了相对独立的两个阶段,所以当我们希望将其扩展,用于决策树构建等其它的分类算法、关联规则挖掘等其它的数据挖掘算法时,第一阶段的 ERRPH 方法和 TRR 方法可以直接使用,只需要在第二阶段的挖掘过程中,利用相应的特征重构方法就可以了.

其次,从数据类型的角度来看,现有数据挖掘中隐私保护方法所使用的随机处理方法,针对的都是数值类型或者布尔类型的数据,都不适用于枚举类型的数据. 例如,DTPD 方法是分类挖掘中布尔类型隐私数据的保护方法;而 DTCRD 方法则是分类挖掘中数值类型隐私数据的保护方法. 而本文提出的 ERRPH 数据处理和特种重构方法,能够很好地处理枚举类型的数据. 而且,通过 ERRPH 方法和 TRR 方法的配合使用,实现了能够同时支持枚举类型和数值类型的完整的隐私保护的朴素贝叶斯分类算法.

另外,现有的朴素贝叶斯分类中的隐私保护方法,只能适用在基于水平划分或者垂直划分的分布式数据库中,而本文提出的基于 ERRPH 和 TRR 的隐私保护方法,能够同时适用于集中式数据库和分布式数据库. 无论原始数据属于一个还是多个数据所有者,都可以将经过 ERRPH 方法和 TRR 方法随机处理后的发布数据集交给数据使用者,并利于相应的特征重构方法,对未知样本实施朴素贝叶斯分类.

综上所述,本文提出的朴素贝叶斯分类中基于

ERRPH 和 TRR 的隐私保护方法,在隐私性、准确性、高效性和适用性等方面都取得了良好的效果.下一节将通过实验对相应的分析结果做进一步的印证.

7 实验分析

在本节中,我们将通过一组实验,来检验本文提出的朴素贝叶斯分类中基于 ERRPH 和 TRR 的隐私保护(Privacy Preserving Naive Bayes,PPNB)方法.一方面,我们会将 PPNB 方法与原始的朴素贝叶斯(Naive Bayes,NB)分类方法,对同一批未知样本进行分类的结果进行对照;另一方面,我们还会将 PPNB 方法与 DTPD 方法,在进行隐私保护分类挖掘时结果的准确性进行对比,并说明数据隐私性和挖掘结果准确性与随机化参数之间的关系.

7.1 实验方法

我们在实验中将使用一组模拟的数据集和一组真实的数据集.模拟数据集是一组银行卡客户信息,包括客户的年龄、学历、职业、婚姻状况、收入水平、持卡数量、刷卡频率、刷卡额度、存款额度、贷款额度等描述属性以及客户价值(高、中、低)的分类属性.我们随机抽取了 50000 个样本作为训练数据集,另取了 5000 个样本作为测试数据集.真实数据集来自某刑侦系统的犯罪嫌疑人信息,总计 33389 条记录,我们选取前 30000 条记录作为训练数据集,其余 3389 条记录作为测试数据集,并将根据犯罪嫌疑人的属性信息,预测其参与案件的类型.

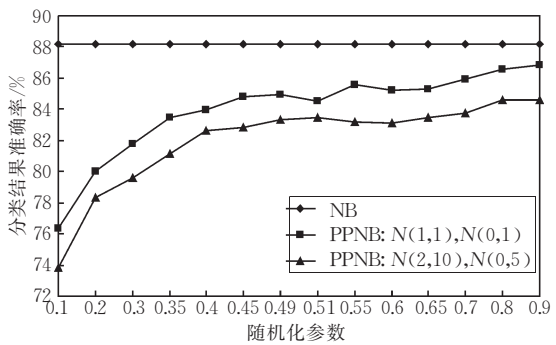
针对这两个数据集,我们将首先利用 ERRPH 方法和 TRR 方法对它们进行随机处理.对于枚举类型的属性,将选取不同的随机化参数 $p_0=0.1$, 0.2, 0.3, 0.35, 0.4, 0.45, 0.49, 0.51, 0.55, 0.6, 0.65, 0.7, 0.8, 0.9; $p_1=p_2=\cdots=p_k=\dfrac{1-p_0}{k}$,使用 ERRPH 方法对原始数据进行随机处理.对于数值类型的属性,我们则将选取满足给定高斯分布的随机数 a 和 b,来对原始数据集进行随机处理.其中,a 分别服从高斯分布 $N(1,1)$ 和高斯分布 $N(2,10)$;b 则分别服从标准正态分布 $N(0,1)$ 和高斯分布 $N(0,5)$.

接下来,我们就可以使用本文提出的 PPNB 方法,根据经过随机处理的训练数据集,来对测试数据集中样本的类标号进行预测,并与它们实际的类标号相对照,计算出分类结果的准确率.为了有所参照,我们还将利用一般的朴素贝叶斯分类方法,来根据原始的训练数据集对测试数据集中的样本标号进行预测,并计算出分类结果的准确率.

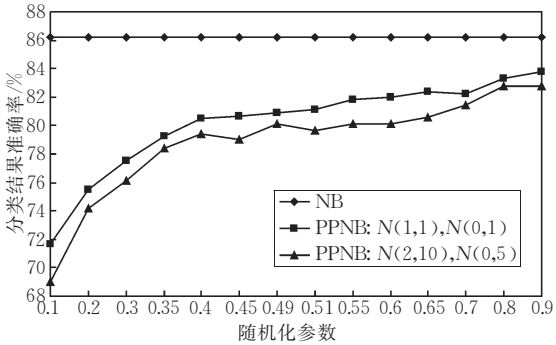
最后,为了与现有分类挖掘中的隐私保护方法 DTPD 相比较,我们将针对模拟数据集,分别选取一个分界点将所有属性都转化成了布尔属性.然后,选取不同的随机化参数 $p_0=p=0.1$, 0.2, 0.3, 0.35, 0.4, 0.45, 0.49, 0.51, 0.55, 0.6, 0.65, 0.7, 0.8, 0.9; $p_1=p_2=\dfrac{1-p_0}{2}$,分别利用 ERRPH(此时已经退化为 RRPH)方法和 DTPT 方法先对原始数据进行随机处理;再利用经过处理的训练数据集,对测试集中的样本进行分类,并比较两种方法的分类结果准确性.

7.2 实验结果

图 2 中分别给出了 PPNB 方法在不同随机化参数取值的情况下,针对(a)银行卡客户数据和(b)刑侦数据的分类结果准确率.其中,横轴表示针对枚举类型属性的 ERRPH 方法中 p_0 的值,纵轴表示分类结果准确率,两条则折线表示针对数值类型属性的 TRR 方法中不同的两组参数.另外,我们还与普通的 NB 方法针对原始数据集的分类结果准确率进行了比较.



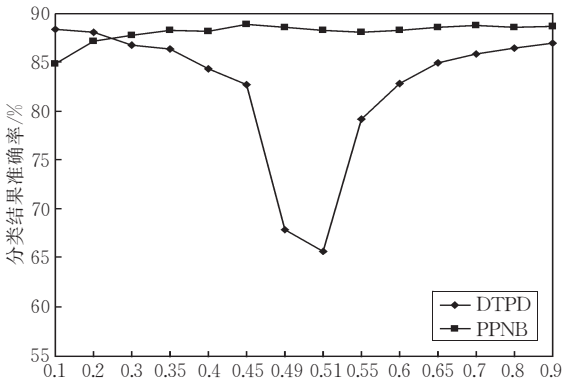
(a) 银行卡系统客户数据



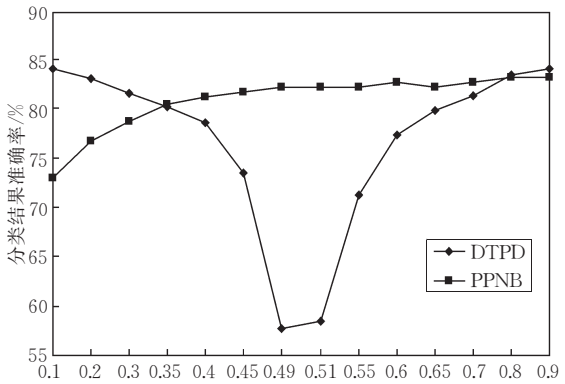
(b) 刑侦系统犯罪嫌疑人数据

图 2 PPNB 方法的分类结果准确率

图 3(a)中给出了 PPNB 方法和 DTPD 方法,对银行卡客户数据的分类结果准确率随参数 p 变化的情况;图 3(b)中则给出了这两种方法,对犯罪嫌疑人数据的分类结果准确率随参数 p 变化的情况.



(a) 银行卡系统客户数据



(b) 刑侦系统犯罪嫌疑人数据

图 3　PPNB 方法和 DTPD 方法的分类结果准确率比较

7.3　结果分析

首先,从图 2 的结果可以看出:PPNB 方法的分类结果准确率会随着 ERRPH 方法中随机化参数 p_0(即真实数据所占的比例)取值的增大而不断提高;还将随着 TRR 方法中参数集合方差的增大而降低.

然后,从图 3 的结果可以看出:DTPD 方法的分类结果准确率变化比较大,当 p 接近 0 或 1 时,挖掘结果比较准确;但此时的隐私破坏系数接近 1,方法对隐私的保护程度很差.在 p 从 0 或 1 逐渐接近 0.5 的过程中,隐私破坏系数会逐渐减小,隐私保护的程度不断提高,但挖掘结果的准确率将显著下降.而本文提出的 PPNB 方法,分类结果的准确率变化相对平稳,随着 p 值,也就是真实数据所占的比例,从 0 增加到 1,隐私破坏系数也从 0 增长到 1,方法对隐私的保护程度不断下降,而挖掘结果准确率不断提高.

再从图 3(a)的详细结果还可以看出:当 p 的取值较小时,DTPD 方法的准确率比 PPNB 方法高;而当 p 的取值超过 0.3 以后,PPNB 方法的准确率就要高于 DTPD 方法了,特别是当 p 值接近 0.5 时,准确率更是相差了很多.这些实验结果与第 6 节中关于隐私性和准确性方面的论证是完全一致的.

图 3(b)的结果和图 3(a)的结果相比,整体效果基本一致,只是两种方法的分类结果准确率都有不同程度下降.其中,PPNB 方法的分类结果正确率下降的更加明显,在图 3(a)中,当 $p\geq 0.3$ 时就比 DTPD 方法具有更高的分类结果准确率,而在图 3(b)中只有 p 在区间[0.4,0.7]上时才具有更高的分类结果准确率.这是因为朴素贝叶斯分类的方法需要类条件独立的假定,模拟数据集是这样构造的,但真实的数据集中属性间是存在相互依赖关系的.尽管如此,只要参数选择合理,PPNB 方法仍然可以取得较好的分类效果.

在理论分析和实验结果的基础上,权衡数据的隐私性和挖掘结果的准确性,我们建议在区间[0.35,0.6]上选取随机化参数 p 的值,在分类挖掘中使用 PPNB 方法进行隐私信息的保护.

8　总　结

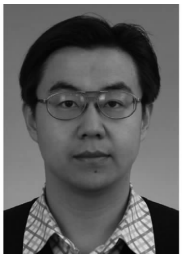
在本文中,我们紧紧围绕分类挖掘中的隐私保护问题展开研究,给出了一种基于数据处理和特征重构的朴素贝叶斯分类中的隐私保护方法.我们根据隐私数据类型的不同,分别提出了一种针对枚举类型的 ERRPH 数据处理和特征重构方法和一种针对数值类型的 TRR 数据处理和特征重构方法;并在此基础上实现了一个完整的隐私保护的朴素贝叶斯分类算法.

我们还通过理论分析和实验,说明了在朴素贝叶斯分类中基于 ERRPH 和 TRR 的隐私保护方法中,随机化参数的选择与数据隐私性和挖掘结果准确性之间的关系与相互影响,并就方法的运行效率和适用范围进行了详尽分析.分析和实验的结果均表明,该方法具有很好的隐私性、准确性、高效性和适用性.

参　考　文　献

[1] Verykios V S, Bertino E, Fovino I N, Provenza L P, Saygin Y, Theodoridis Y. State-of-the-art in privacy preserving da-

ta mining. SIGMOD Record, 2004, 33(1): 50-57
[2] Agrawal R, Srikant R. Privacy-preserving data mining//Proceedings of the 2000 ACM SIGMOD Conference on Management of Data. Dallas, Texas, USA, 2000: 439-450
[3] Evfimievski A. Randomization in privacy preserving data mining. SIGKDD Explorations, 2002, 4(2): 43-48
[4] Agrawal D, Aggarwal C C. On the design and quantification of privacy preserving data mining algorithms//Proceedings of the 20th ACM Symposium on Principles of Database Systems. Santa Barbara, California, USA, 2001: 247-255
[5] Du W L, Zhan Z J. Using randomized response techniques for privacy-preserving data mining//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington DC, USA, 2003: 505-510
[6] Rizvi S J, Haritsa J R. Maintaining data privacy in association rule mining//Proceedings of the 28th International Conference on Very Large Data Bases. Hong Kong, China, 2002: 682-693
[7] Johnsten T, Raghavan V V. A methodology for hiding knowledge in databases//Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining. Maebashi City, Japan, 2002: 9-17
[8] 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)
[9] Du W L, Zhan Z J. Building decision tree classifier on private data//Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining. Maebashi City, Japan, 2002: 1-8
[10] Pinkas B. Cryptographic techniques for privacy preserving data mining. SIGKDD Explorations, 2002, 4(2): 12-19
[11] Lindell Y, Pinkas B. Privacy preserving data mining//Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology. Santa Barbara, California, USA, 2000: 36-54
[12] Kantarcoglu M, Vaidya J. Privacy preserving Naive Bayes classifier for horizontally partitioned data//Proceedings of the IEEE ICDM Workshop on Privacy Preserving Data Mining. Melbourne, FL, USA, 2003: 3-9
[13] Vaidya J, Clifton C. Privacy preserving Naive Bayes classifier for vertically partitioned data//Proceedings of the 4th SIAM International Conference on Data Mining. Lake Buena Vista, Florida, USA, 2004: 522-526
[14] Clifton C, Kantarcioglu M, Vaidya J, Lin Z, Zhu M Y. Tools for privacy preserving distributed data mining. SIGKDD Explorations, 2002, 4(2): 28-34
[15] Zhang P, Tong Y H, Tang S W, Yang D Q. KD^3 scheme for privacy preserving data mining//Proceedings of the 4th IEEE International Conference on Intelligence and Security Informatics. San Diego, CA, USA, 2006: 659-661



ZHANG Peng, born in 1978, Ph.D.. His research interests include date warehousing, date mining, and OLAP.

TANG Shi-Wei, born in 1939, professor, Ph.D. supervisor. His research interests include data model, database system, data warehousing, and data mining.

Background

This work is supported by the National Natural Science Foundation of China under grant No. 60403041, and the Dissertation Foundation of Beijing Municipal Science and Technology Commission under grant No. ZZ6027.

Nowadays, there is growing concern with the privacy implications of data mining. How to solve the privacy preserving problems during the mining process has become one of the most important topics in data mining. Data mining has an essential property that the patterns from large amounts of data usually depend on the aggregate and statistical data, but not the individual data records. Then, privacy preserving data mining that is to discovery accurate patterns without precise access to the original data has become a novel research direction.

In this paper, the authors present a privacy preserving Naive Bayes classification approach based on data randomization and feature reconstruction. An ERRPH method and a TRR method are respectively presented for enumerated data and numerical data. Then, a privacy preserving Naive Bayes classification algorithm is implemented based on those methods. Theoretical analyses and experimental results show that it can provide better privacy, accuracy, efficiency, and applicability.

The research group has been working on privacy preserving date mining since 2004, and issued about ten papers including the scheme, the workflow, and the evaluation measures for privacy preserving date mining, privacy preserving association rule mining approaches, privacy preserving classification approaches, and the effect of correlation of attributes for privacy preservation.