DOI:10.16451/j.cnki.issn1003-6059.2010.06.018

基于隐私保护的决策树模型*

方炜炜[1,2] 杨炳儒[2] 杨 君[2] 周长胜[1]

[1](北京信息科技大学 计算中心 北京 100192)
[2](北京科技大学 信息工程学院 北京 100083)

摘 要 在分布式环境下,实现隐私保护的数据挖掘,已成为该领域的研究热点.文中着重研究在垂直分布数据中,实现隐私保护的决策树分类模型.该模型创建新型的隐私保护决策树,即由在茫然半诚实方存储的全局决策表和各站点存储的局部决策树组成,并结合索引数组和秘密数据比较协议,实现在不泄漏原始信息的前提下决策树的生成和分类.经过理论分析和实验验证,证明该模型具有较好的安全性、准确性和适用性.

关键词 隐私保护数据挖掘(PPDM),决策树,垂直分布
中图法分类号 TP 181

Decision-Tree Model Research Based on Privacy-Preserving

FANG Wei-Wei[1,2], YANG Bing-Ru[2], YANG Jun[2], ZHOU Chang-Sheng[1]
[1](*Computer Center*, *Beijing Information Science and Technology University*, *Beijing* 100192)
[2](*Information Engineering School*, *University of Science and Technology Beijing*, *Beijing* 100083)

ABSTRACT

How to realize privacy-preserving data mining becomes a research hotspot in a distributed environment. A model is proposed to realize privacy-preserving decision-tree classifying when data are vertically partitioned. In this model, a privacy-preserving decision-tree is proposed, which is composed of Global-Table stored by an obvious semi-honest partner and several local decision-trees stored by different sites. By using indexed array and private data comparison protocol, decision-tree generation and classification can be realized without uncovering the original information. Theoretical analysis and experimental results demonstrate the proposed model provides good capabilities of privacies preserving, accuracy and efficiency.

Key Words Privacy-Preserving Data Mining (PPDM), Decision-Tree, Vertical Distribution

1 引 言

近年来,随着信息产业广泛而深入的发展,面向不同应用领域的数据激增,数据挖掘 Data Mining (DM) 理论及方法逐渐成为研究热点,致力于从海量的数据中挖掘出有效的、新颖的、潜在可用的和容

* 国家自然科学基金资助项目(No. 60875029)
收稿日期: 2009 - 07 - 27; 修回日期: 2010 - 03 - 01
作者简介 **方炜炜**,女,1979 年生,博士研究生,主要研究方向为数据挖掘. E-mail: Liveinbetter@163.com. **杨炳儒**,男,1943 年生,教授,博士生导师,主要研究方向为推理机制与知识发现. **杨君**,男,1970 年生,博士研究生,副教授,主要研究方向为多关系数据挖掘. **周长胜**,男,1961 年生,博士,副教授,主要研究方向为数据流挖掘.

易理解的知识模式[1]. 然而在数据挖掘的过程中存在着隐私安全问题. 一旦原始资料中的敏感隐私信息泄漏给外界,会对企业或个人的隐私和信息安全构成威胁. 在这种情况下,如何在保证个人隐私的前提下进行数据挖掘,成为一个急需解决的问题. 面向隐私保护的数据挖掘(Privacy-Preserving Data Mining, PPDM) 技术应运而生,成为数据挖掘领域中一个极其重要而富有挑战性的课题,以“隐私数据及敏感规则的保护“和“隐含规则知识的准确挖掘“两者兼得为其最终目标.

近年来国内外已有不少学者对该课题展开研究. 2000 年,R. Agrawal 和 Y. Lindell 分别提出采用数据扰动(Data Perturbation) 技术和安全多方计算(Secure Multiparty Computation) 技术来解决隐私保护问题,为后续学者的研究奠定了基础. 数据扰动技术有两种: 1) 通过向原始数据添加“噪音”[2-4],然后基于贝叶斯、期望最大化算法(Expectation Maximization, EM) 等方法重构模型,消除噪音影响; 2) 通过数据变换,如文献[5]中提出特征值分解和文献[6]中提出离散余弦变换矩阵等方法,使得挖掘者在新构建数据模型上进行数据挖掘任务. 2006 年,Kamalika 在文献[7]中指出该类方法在应用中的众多缺陷,如添加噪音而造成挖掘精度的损失,通过新构建数据模型可推导部分原始隐私信息等,因而近年来该类技术发展缓慢. 安全多方计算技术,即通过将密码学领域中的研究成果应用于数据挖掘中的隐私保护. 2002 年,B. Pinkas[8] 证明不同种类的数据挖掘问题都可以转化为安全的多方计算. 2004 年,Clifton[9] 提出可通过将安全求和、安全求并、安全交集大小以及标量积方法应用于分布式数据挖掘. 2007 年,Justin[10] 提出可通过将同态加密、数字信封技术应用于分布式挖掘,实现各参与方在不泄漏本地隐私信息的前提下,挖掘出可共享规则.

从数据挖掘任务的角度来看,目前国内外 PPDM 技术的研究成果大多是关于关联规则[2,11-12]、聚类[4,6,10] 和朴素贝叶斯分类[13] 等,而少数几篇针对决策树构建展开研究的文献[14]、[15]是讨论在水平分布数据的情况. 本文将紧紧围绕垂直分布数据情况下决策树分类的隐私保护技术展开研究. 创建新型的决策树. 该决策树由可公开的全局决策表和存储在各参与方不公开的局部决策树组成,其中全局决策表不显示任何属性名、值信息,最大化地实现了隐私信息保护. 将新决策树构造、ID 索引数组集合与秘密数据比较协议相结合,实现决策树生成和分类两个过程的隐私保护,并对模型安全性、计算和通信复杂度进行详细分析.

2　相关知识概述

决策树分类是数据挖掘应用中一种常用的技术,因其挖掘结果容易理解、精度高和鲁棒性好而著称. 其构造过程是从“选择最佳分裂属性作为根节点被测试”开始,基于训练样例集 D,使用统计分裂规则来确定分类能力强的属性作为根节点的测试. 然后为根节点属性的每个可能值产生一个分支,并把训练样例集 D 排列到适当的分支. 再重复该过程,用每个分支节点关联的训练样例来选取在该点被测试的最佳分裂属性. 直至条件属性全部判断完毕或剩余样例集的决策属性值相同而停止树的增长.

定义 1(垂直分布数据集)　数据库 D 中包含有 m 条数据记录,每条数据记录均由 n 个条件属性 $R = R_1, R_2, \cdots, R_n$ 和1个决策属性 C 组成. 现将数据库 D 分布式存储在 k 个站点 $P_1, P_2, \cdots, P_k$,每个站点拥有条件属性集 $R_{P_i} = R_t \cup \cdots \cup R_s$,($1 \leqslant t, s \leqslant n$, $1 \leqslant P_i \leqslant k$) ,如满足

1) $D_{P_1} \cup D_{P_2} \cup \cdots \cup D_{P_k} = D$;

2) $\forall i \neq j, R_{P_i} \cap R_{P_j} = \varnothing$,

则称为垂直分布数据集.

在决策树的生成过程中,本地选择最佳分裂属性是关键步骤之一,常采用信息论中的信息增益、增益率和 Gini 指标作为衡量标准. 本文将采用信息增益作为评价指标.

定义 2(信息熵)　对于数据集 D, $V_c = \{C_1, C_2, \cdots, C_m\}$ 是决策属性 C 的值域集, p_i 是任意样本属于 C_i 的概率,则用于识别 D 中元组的类标号所需要的平均信息量称为 D 的信息熵:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i).$$

定义 3(信息增益)　如果依据条件属性 R 的 v 个不同观测值$\{r_1, r_2, \cdots, r_v\}$ 可将 D 中的元组划分为 v 个子集$\{D_1, D_2, \cdots, D_v\}$,则基于按 R 划分对 D 的元组分类所需期望信息:

$$InfoR(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j),$$

信息增益即为原信息需求(仅基于类比例) 与新的信息需求(对 R 划分后得到的) 之间的差,表示为

$$Gain(R) = Info(D) - Info_R(D).$$

3　隐私保护的决策树

本文中所研究的基于隐私保护的决策树模型,

对其垂直分布数据集做如下假设:

1) 每个站点的记录集均含有决策属性 C 和记录编号 ID;

2) 每个站点 $D_{P_i}(1 \leqslant P_i \leqslant k)$ 均含有 m 条记录集.

各站点在选出本地最佳分裂属性后,需要相互比较,为避免传输和比较过程中的原始信息泄漏,我们采用秘密比较协议.

定义 4(秘密比较协议) Alice 输入秘密数 i, Bob 输入秘密数 j, Alice 和 Bob 安全地计算 GT 函数,当 $GT(i,j) = 1$ 时,$i > j$ 成立;否则 $GT(i,j) = 0$, $i \leqslant j$.

图灵奖获得者姚启智教授曾提出的百万富翁协议(The Millionaire Protocol)[16],其通信复杂度是待比较数据的二进制位数的指数次幂,只适合两个很少的秘密数据进行比较. Cachin 为提高效率,引入一个茫然(Oblivious)半诚实方,提出 ϕ-隐藏假设[17],本模型依据基于 ϕ-隐藏假设和同态公钥加密体制的 ϕ-HA 协议[18]求解全局最佳分裂属性.

定义 5(ϕ-隐藏假设) 设 m 为一个不知其因数的合数,$\phi(m)$ 为欧拉函数,P_k 为 k 比特的素数所成之集,R_k 为合数 m 所成之集,其中 $m = pq$, p 与 q 为 k 比特的素数且其中一个为安全素数,即形如 $2s+1$ 且 s 为素数,另一个为准安全素数,即形如 $2tr+1$ 且 t 与 r 为素数. 我们说 $m \in R_k$ 隐藏了素数 t,若 $t \in P_k$ 使 $t \mid \phi(m)$.

由上定义可断言[14]:若随机数 $m \in R_k$,隐藏了素数 $p_1, p_2 \in P_k$ 是独立地随机选取的素数,则 (m, p_1) 与 (m, p_2) 是计算不可区分的.

定义 6(同态公钥加密体制) 设 S 为一公钥密码体制,k 为其安全参数,X 为信息空间. $E_k: \{0,1\}^k \times X \to C$ 为公开加密函数,$E_k(u,x) \in C$, $u \in \{0,1\}^k$ 为一随机串,$x \in X$ 为一消息,C 为密文空间,$D_k: C \to X$ 为保密的解密函数,并假设 X 为 Abel 加群,C 为 Abel 加群,对两个任意的信息 $x_1, x_2 \in X$ 及随机串 $u_1, u_2 \in \{0,1\}^k$,若 $E_k(u_1, x_1)$ 与 $E_k(u_2, x_2)$ 是计算不可区分的,则称公钥密码体制 S 是语义安全的;若还存在 $u \in \{0,1\}^k$,使

$$E_k(u, x_1 + x_2) = E_k(u_1, x_1) \times E_k(u_2, x_2),$$

则称 S 是语义安全的同态公钥加密体制.

3.1 模型思想

设原始数据集 D(见表 1)中包含有记录标号 ID、条件属性集{*Outlook*, *T.* , *Humid*, *Windy*}和决策属性 *play*. 现垂直划分存储在 A 和 B 站点,其中站点 A 包含属性集为{*ID*, *Outlook*, *T.* , *play*},站点 B 包含属性集为{*ID*, *humid*, *windy*, *play*}.

表 1 数据库 D

Table1 Database D

ID	*Outlook*	*T.*	*Humid*	*Windy*	*Play*
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rain	Mild	High	False	Yes
5	Rain	Cool	Normal	False	Yes
6	Rain	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Hot	Normal	False	Yes
10	Rain	Mild	Normal	False	Yes
11	Sunny	Hot	Normal	True	No
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rain	Mild	High	True	No

为避免各数据拥有方属性名和值的泄漏,首先我们对传统的决策树构造形式进行改进,由全局决策表 *Global-Table*、局部决策树 *Local-Tree*(A/B)构成,并均初始化为空. 然后引入茫然半诚实方 T(用于秘密数据比较和存储全局决策表 *Global-Table*),设初始 *ID* 索引数组集包含所有记录的编号,根据如下步骤构造出决策树.

step 1 各站点产生本地最佳分裂属性. 各参与站点利用信息增益 $Gain(R)$ 作为最佳分裂属性选择度量,计算出本地 *ID* 索引数组集所包含的记录集的最佳分裂属性 R_A, R_B.

step 2 秘密比较全局最佳分裂属性值,更新全局决策表. 参与方保密数据 R_A、R_B 和茫然半诚实方 T 均形式化为概率多项式时间图灵机并通过安全信道交互,用来计算的函数

$$\pounds: R_A \times R_B \times D_T \to \{0,1\} \times \{0,1\} \times D_T,$$

即

$$\pounds(a,b,\varepsilon) = \begin{cases} (0,0,\varepsilon), & a = b \\ (1,0,\varepsilon), & a > b \\ (0,1,\varepsilon), & a < b \end{cases}$$

其中,R_A, R_B 表示为其二进制形式,如

$$R_A = \sum_{i=0}^{l} a_i 2^i, \ a_i \in \{0,1\}.$$

A、B 方的输入 a、b 均为 l 比特的正数,T 的输入输出仅为空字母 ε,除 a 与 b 之间的大小或相等关系外,A、B 均不能从自己的输入与输出 $\pounds(a,b,\varepsilon)$ 中得到关于对方的输入任何信息. 通过 ϕ-HA 协

议[17-18] 比较 R_A、R_B 大小确定最佳分裂属性,并将最佳分类属性站点名添加至全局决策表 *Global-Table*.

step 3　更新局部决策树、*ID* 索引数组集合. 最佳分裂属性站点方将最佳分裂属性名添加至本地局部决策树 *Local-Tree*(*A*/*B*) ,并根据最佳分裂属性值分布情况产生本地的索引数组集合,记录同一属性值的 *ID* 号. 如 *outlook* 为最佳分裂属性时,站点 *A* 生成

$$AR_1[1] = \{1,2,8,9,11\},$$

$$AR_1[2] = \{3,7,12,13\},$$

$$AR_1[3] = \{4,5,6,10,14\}.$$

step 4　最佳分裂属性站点发送信息. 最佳分裂属性站点将属性值个数发送给 *T* 更新全局决策表 *Global-Table*,*ID* 索引数组信息发送给其它参与方.

step 5　判断是否停止决策树生成.

step 5. 1　判断全局决策表 *Global-Table* 元素个数是否达到条件属性总数 *v*,如是则停止构造决策树,否则继续 step 5. 2.

step 5. 2　判断 ID 索引数组所指向的记录集的决策属性值是否相同,是则产生叶子节点(如 $AR_1[2]$ 产生 *path* 3) ,否则继续 step 5. 3.

step 5. 3　转向 step 1,各参与方根据新 *ID* 索引数组所指向的记录集继续执行.

最终可生成图 1 所示的全局决策表 *Global-Table* 和图 2 所示的本地局部决策树 *Local-Tree*(*A*/*B*) .

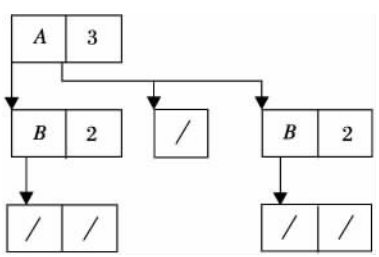


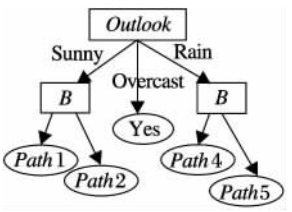
图 1　全局决策表

Fig. 1　Global-Table

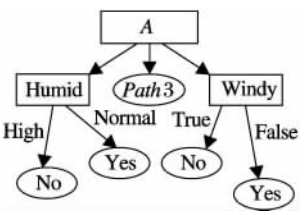
使用该决策树对校验样本进行分类时,我们可以求解路径交集来提高分类效率和实现隐私保护. 举例如下,假设有记录

c = (*Outlook* = *sunny*, *T.* = *cool*, *Humid* = *normal*, *Windy* = *false*) ,

数据拥有方 *A* 根据本地存储的 *Local-Tree*(*A*) 和条件 *Outlook* = *sunny*, *T.* = *cool* 判断该样本可能路径是 *path* 1, *path* 2. 同样,数据拥有方 *B* 根据 *Local-Tree*(*B*) 和条件 *Humid* = *normal*, *Windy* = *false* 判断该样本可能路径是 *path* 2, *path* 5. 那么由 *A*、*B* 两站点判断路径交集则是 *path* 2,再通过 *B* 站点即可判断该记录的决策属性为 *play* = *yes*.



(a) *Local-Tree*(*A*)



(b) *Local-Tree*(*B*)

图 2　局部决策树

Fig. 2　Local-Tree

3. 2　算法描述

根据上述思想,设计算法如下.

输入　*D*(*A*/*B*) ,数据库 *A*(或 *B*) 的训练集; *C*,决策属性; *v*,条件属性的总数; 全局决策表 *Global-Table* = ϕ

输出　*Global-Table*,可公开的全局决策表;

Local-Tree(*A*/*B*) ,*A*(或 *B*) 局部决策树

步骤(*A*、*B* 方各自在本地运行)

if *num*(*Global-Table*) = *v* then

　Local-Tree(*A*/*B*) 添加叶子节点,标记 *D*(*A*/*B*) 绝大多数类 *c*;

else if *D*(*A*/*B*) 拥有相同类 *c* then

　Local-Tree(*A*/*B*) 添加 *c* 类叶子节点;

else

　{ 求 *D*(*A*/*B*) 中本地的最佳分裂属性 *a*(或 *b*) ;

　调用 ϕ-HA 协议比较 $Info_a$(*DA*) 和 $Info_b$(*DB*) ,获全局最佳分裂属性 *t*;

　Global-Table 数组添加元素(*A*/*B*) *m*;

if 本地站点拥有最佳属性 *t* then

　根据 *t* 属性值分布,划分训练集{ $D(t_1)$,$D(t_2)$,…, $D(t_m)$ } ,建立相应 *ID* 索引数组集合,发送给对方;

　调用 *PPID*($D(t_i)$,*C*) ; }

end if

4　性能分析

设数据库 *D* 的记录数为 *m*,垂直分布于 *k* 个站

点 P_1, P_2, …, P_k, 数据库 D 的条件属性 R 有 n 个, 每个属性的可能观测值的最大数目为 v, 决策树的总节点数为 s, 局部最佳分裂属性信息增益值二进制位长为 r. 现从计算复杂度、通信复杂度和模型安全性 3 个方面分析如下.

4.1 计算和通信复杂度

对于站点 D_{P_i}, 判断 *Global-Table* 数组元素个数是否等于 v 的时间复杂度是 $O(1)$, 如果相等, 查找绝大多数类 c 的时间复杂度 $O(m)$. 判断 $D(A/B)$ 拥有相同类 c 的时间复杂度 $O(m)$. 划分训练集的时间复杂度 $O(v \times m)$. 各站点求解最佳分裂属性的时间复杂度的总和为 $O(n \times m)$. 因此在整个运行过程中, 计算数据的时间复杂度为

$$O(1 \times k \times s) + O(m \times k \times s) + O(m \times k \times s) + O(v \times k \times m \times s) + O(n \times m \times s).$$

运行过程中, 利用 ϕ-*HA* 协议比较节点最佳分裂属性的通信复杂度为 $O(r)$. 节点最佳属性拥有方划分 *ID* 索引数组后需要给 $k-1$ 个站点发送 *ID* 信息, 因为站点产生的索引集上限为 m, 其通信复杂度为 $O(k \times v \times m \times s)$.

4.2 模型安全性

在决策树的生成过程中, 可公开的全局决策表仅存储站点名和分支个数, 站点间传输的是 *ID* 编号, 因而各参与方拥有的属性名和个体属性值信息均得到保护. 通过使用 ϕ-*HA* 协议比较全局最佳分裂属性时, 因为 $(A(a), B(b), T(\varepsilon))$ 与 $\pounds(a, b, \varepsilon)$ 计算不可区分, 保证该协议的正确性和安全性, 其证明请参阅文献 [17]、[18], 所以各参与方均不能通过输入与输出获取对方信息增益值, 茫然半诚实方对于参与方的输入也是一无所知, 保证节点信息增益值的不泄漏.

在决策树的分类过程中, 通过路径交集查找决策属性值, 既可提高执行效率, 又可避免根据最终结果回溯原始记录信息.

4.3 实验验证

本实验环境是 5 台微机构成 100MB 局域网, 处理器 Intel Pentium M 1.4GHz, 内存 1GB, 硬盘 80GB, 操作系统 Windows Sever 2000, 开发环境 JBuilder 2006 Enterprise, 编程语言 Java. 我们采用 UCI 机器学习数据集储存库中 Iris, Soybean 和 Zoo 数据集作为测试数据集, 如表 2 所示.

实验中我们采用 2/3 的数据集作为训练样本, 余下 1/3 的数据集作为测试样本. 图 3 给出基于 ϕ-HA 协议的分布式 PPDM 隐私保护决策树挖掘与传统的 ID 3 决策树挖掘的分类结果准确率比较.

表 2 测试数据集

Table 2 Testing data set

数据集	样本个数	属性维数
Iris	101	16
Soybean	47	35
Zoo	150	50

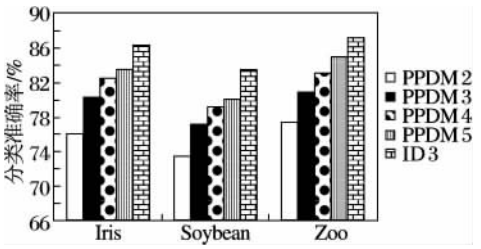


图 3 PPDM 与 ID 3 分类结果准确率对比

Fig. 3 PPDM and ID 3 classifying results

我们针对 3 个数据集, 分别采用集中式传统 ID 3 方法挖掘和 2 ~ 5 个参与方的分布式 PPDM 隐私保护方法挖掘. 从实验结果可以看出, 对于样本数量最多的 Zoo 数据集, 采用传统 ID 3 挖掘的准确率为 87.2%, 采用 5 个参与方分布式 PPDM 隐私保护挖掘的准确率为 85%, 而样本个数偏少但属性维数较多的 Soybean 数据集其结果分别为 83.5% 和 80%, 说明样本个数越多, 其分布式 PPDM 隐私保护挖掘结果准确率越接近于传统集中式 ID 3 挖掘结果准确率. 另一方面, 对于同一测试数据集采用分布式 PPDM 隐私保护挖掘时, 其挖掘准确率随着参与方个数的增多而提高, 但其执行时间也会有所增加, 如图 4 所示.

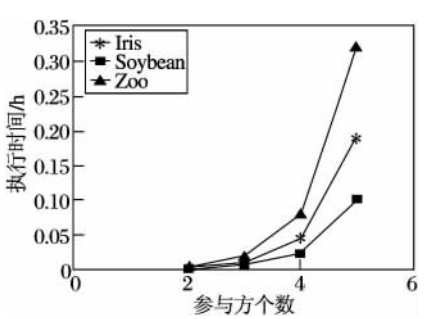


图 4 样本参与个数改变时执行时间对比

Fig. 4 Running time comparison when number of samples change

从图 4 的结果可以看出, 由于 3 个数据集样本的数量级小, 在只有 2 个参与方分布式 PPDM 挖掘时执行时间差不多, 但当随着参与方的增多, 由于相

互之间的通信时间增长,其执行时间都相应的增加.

5　结 束 语

我们就基于隐私保护的决策树构建及分类模型展开深入研究,创建一种不泄漏属性名和值信息的全局决策表和局部决策树. 通过采用秘密数据比较协议实现全局最佳属性比较的信息无泄漏,通过 ID 索引数组传递待挖掘子集,通过求解路径交集避免根据校验样本回溯站点原始信息,就模型的计算、通信复杂度和安全性进行分析讨论,并通过实验验证模型方法具备较好的安全性、准确性和适用性.

在未来的工作中,我们将继续研究如何基于多方安全计算协议设计关于不泄漏各站点路径进行交集的求解,并进一步提高挖掘算法的运行效率,扩充其解决问题的范围.

参 考 文 献

[1] Han Jiawei. Data Mining: Definition and Technology. Beijing, China: Mechanism Industry Publish, 2001 (in Chinese)
(韩家炜. 数据挖掘: 概念与技术. 北京: 机械工业出版社, 2001)

[2] Rizvi S J, Haritsa J R. Maintaining Data Privacy in Association Rule Mining // Proc of the 28th International Conference on Very Large Databases. Hongkong, China, 2002: 682 – 693

[3] Agrawal R, Srikant R. Privacy-Preserving Data Mining. ACM SIGMOD Record, 2000, 29(2) : 439 – 450

[4] Vaidya J, Clifton C. Privacy-Preserving *K*-Means Clustering over Vertically Partitioned Data // Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2003: 206 – 215

[5] Li Feng, Li Shenghong, Li Jianhua. An SVD-Based Advanced Data Perturbation Method for Privacy-Preserving Data Mining. Journal of Shanghai Jiaotong University, 2009, 43(3) : 427 – 431 (in Chinese)
(李 锋, 李生红, 李建华. 一种基于特征值分解的数据挖掘隐私保护扰乱增强方法. 上海交通大学学报, 2009, 43(3) : 427 – 431)

[6] Zhang Guorong, Yin Jian. Privacy Data Preserving Method Based on Discrete Cosine Transform Matrix. Computer Engineering, 2009, 35 (2) : 157 – 162 (in Chinese)
(张国荣, 印 鉴. 基于离散余弦变换矩阵的隐私数据保护方法. 计算机工程, 2009, 35(2) : 157 – 162)

[7] Chaudhuri K, Mishra N. When Random Sampling Preserves Privacy // Proc of the 26th Annual International Cryptology Conference. Santa Barbara, USA, 2006: 198 – 213

[8] Pinkas B. Cryptographic Techniques for Privacy-Preserving Data Mining. ACM SIGKDD Explorations Newsletter, 2002, 4(2) : 12 – 19

[9] Cliffton C, Kantarcioglu M, Vaidya J, *et al*. Tools for Privacy Preserving Distributed Data Mining. ACM SIGKDD Explorations Newsletter, 2004, 4(2) : 28 – 34

[10] Zhan J. Using Cryptography for Privacy Protection in Data Mining Systems // Proc of the 1st WICI International Workshop on Web Intelligence Meets Brain Informatics. Beijing, China, 2007: 494 – 513

[11] Kantarcioglous M, Clifton C. Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. IEEE Trans on Knowledge and Data Engineering, 2004, 16(9) : 1026 – 1037

[12] Vaidya J, Clifton C. Privacy, Preserving Association Rule Mining in Vertically Partitioned Data // Proc of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Canada, 2002: 639 – 644

[13] Zhang Peng, Tan Shiwei. Privacy Preserving Naïve Bayes Classification. Chinese Journal of Computers, 2007, 30(8) : 1267 – 1272 (in Chinese)
(张 鹏, 唐世谓. 朴素贝叶斯分类中的隐私保护方法研究. 计算机学报, 2007, 30(8) : 1267 – 1272)

[14] Emekci F, Sahin O D, Agrawal D, *et al*. Privacy Preserving Decision Tree Learning over Multiple Parties. Data & Knowledge Engineering, 2007, 63(2) : 348 – 361

[15] Agrawal D, Aggarwal C. On The Design and Quantification of Privacy Preserving Data Mining Algorithms // Proc of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Santa Barbara, USA, 2001: 247 – 255

[16] Yao A C. Protocols for Secure Computations // Proc of the 23rd Annual IEEE Symposium on Foundations of Computer Science. Chicago, USA, 1982: 160 – 164

[17] Cachin C. Efficient Private Bidding and Auctions with an Oblivious Third Party // Proc of the 6th ACM Conference on Computer and Communication Security. Singapore, Singapore, 1999: 120 – 127

[18] Qin Jing, Zhang Zhenfeng, Feng Dengguo, *et al*. A Protocol of Specific Secure Two-Party Computation. Journal of China Institute of Communications, 2004, 25(1) : 35 – 42 (in Chinese)
(秦 静, 张振峰, 冯登国, 等. 一个特殊的安全双方计算协议. 通信学报, 2004, 25(1) : 35 – 42)