分布式决策树挖掘的隐私保护研究

方炜炜[1,2] 胡 健[1,3] 杨炳儒[1] 周长胜[2]
(北京科技大学信息工程学院 北京 100083)[1] (北京信息科技大学 北京 100192)[2]
(江西理工大学信息工程学院 赣州 341000)[3]

摘 要 数据挖掘中的隐私保护是试图在不精确访问原始数据值的前提下,挖掘出准确的模式与规则。围绕分布式决策树挖掘的隐私保护问题展开研究,提出一种基于同态加密技术的决策树挖掘算法,使各参与方在不共享其隐私信息的前提下达到集中式挖掘的效果。理论分析和实验结果表明,该算法具有很好的隐私性、准确性和适用性。

关键词 隐私保护,数据挖掘,决策树,同态加密

Research of Privacy-preserving in Distributed Decision-tree Mining

FANG Wei-wei[1,2] HU Jian[1,3] YANG Bing-ru[1] ZHOU Chang-sheng[2]
(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)[1]
(Beijing Information Science and Technology University, Beijing 100192, China)[2]
(School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China)[3]

Abstract Privacy-preserving data mining is discovering accurate patterns and rules without precise access to the original data. This paper focused on privacy-preserving research in the situation of distributed decision-tree mining, and presented a decision-tree mining algorithm based on homomorphic encryption technology, which can get accurate mining effect in the premise of no sharing of private information among mining participators. Theoretical analysis and experiment results show that this algorithm can provide good capability of privacy-preserving, accuracy and efficiency.

Keywords Privacy-preserving, Data mining, Decision-tree, Homomorphic encryption

1 引言

近几十年来,随着数据采集和存储技术的进步,庞大的数据库日益增多。人们面临的已经不是信息缺失,而是从数据量巨大的资料中选择性地收集认为有用的信息。数据挖掘跨越数据库、人工智能、机器学习、统计分析等多个学科,通过对数据归纳、分析和推理,从中发掘出潜在的模式,获取未知的、可信的并具有价值的知识。然而,越来越多的数据挖掘需要多方数据合作,例如不同银行希望从共享数据中得到欺诈行为的模式;政府机构需要和航天部门合作,挖掘恐怖行为的踪迹。在这类情况下,数据常常分布在不同的地点,从属于不同的组织,政府、企业和其他一些机构在进行协同工作完成全局性的数据挖掘时,往往希望在不共享自己存储的隐私数据的前提下,获取共同挖掘的规则结果。因而,隐私保护数据挖掘(PPDM, privacy-preserving data mining)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,以实现隐私数据的保护和基于统计的模式抽取两者兼得为其最终目标。

隐私保护数据挖掘的基本思想在于对原始数据或者挖掘方法进行某种改进,使得各参与节点在无法获取其他节点信息及敏感中间计算结果的同时,获取某些可共享的挖掘规则和隐含知识。自1999年Rakesh Agrawal在KDD99上做了一场精彩的主题演讲,将隐私保护数据挖掘作为未来的研究重点之一后,国内外学者对此进行了研究,提出各种隐私保护技术,其基本策略主要分为数据干扰和查询限制两种。数据干扰策略就是首先通过数据变换、数据中添加噪音等方法对原始数据进行干扰,然后再针对经过干扰的数据进行挖掘,得到所需的模式和规则;查询限制策略则是通过数据隐藏、数据抽样和数据划分等方式,避免数据挖掘者拥有完整的原始数据,而后再利用贝叶斯等概率统计方法得到所需的挖掘结果。但是,这两种策略本身都存在一些固有的缺陷。在采用数据干扰策略的方法中,所有经过干扰的数据均与真实的原始数据直接相关;而在采用查询限制策略的方法中,所有提供的数据又都是真实的原始数据,这些都会降低方法对隐私数据的保护程度。

为了提高对隐私数据的保护程度,本文提出了一种基于同态加密的决策树挖掘方法(DMBHE, Decision-tree Mining Based on Homomorphic Encryption),使得在分布式决策树挖掘的过程中,在各参与方无需共享自己的隐私数据的前提下,通过计算器端直接在加密数据上计算加密的全局统计信息,

到稿日期:2008-05-13 本文受国家自然科学基金重点项目(60675030),国家自然科学基金重点项目(69835001),北京市教委科技计划面上项目KM200811232013,08年北京信息科技大学科研基金项目资助。

方炜炜 博士研究生,讲师,研究方向为隐私数据挖掘,E-mail:liveinbetter@163.com;**胡 健** 副教授,研究方向为数据挖掘;**杨炳儒** 教授,博士生导师,研究方向为推理机制与知识发现;**周长胜** 副教授,研究方向为数据挖掘。

半可信第三方挖掘者在解密后的全局统计信息上进行决策树构建，从而实现了原始信息的隐私保护。最后通过实验数据进行验证，表明 DMBHE 算法可达到集中式数据挖掘的效果，具有很好的隐私保护性、挖掘准确性。

2 问题与架构

分布式环境中，与传统的集中式数据挖掘不同，隐私保护数据挖掘需要考虑两方面：一方面，各数据拥有方希望自己持有的本地信息不被其他数据拥有方获取；另一方面，他们又希望通过合作挖掘共享全局的隐含规则。决策树分类是数据挖掘中的一种常用方法，以其挖掘结果易理解、精度高和鲁棒性好而著称。本文主要研究在多个水平划分(horizontally partitioned)数据库(即数据集按记录分布在不同站点)的联合样本集上实现隐私保护的决策树挖掘。

2.1 问题描述

设有 3 个不同的机构 A,B,C，分别拥有属性类别一致的数据集 RS_a,RS_b,RS_c。三方机构均希望基于联合数据集 $S=RS_a \cup RS_b \cup RS_c$ 构造决策树，用于数据分类。而在构造决策树的同时，必须保证任何一方均不会泄露自己的与最终挖掘结果无关的信息。隐私保护的决策树挖掘，就是在不精确访问原始数据集的条件下，尽可能准确地构造出分类决策树。

通常，决策树的构造过程是采用自顶向下进行递归的分治方式构造。即从“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 S，使用统计分裂规则(如信息增益、增益率和 Gini 指标)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 S 排列到适当的分支(即样例的该属性值对应的分支下)；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。

可见，决策树算法的技术难点也就是选择一个最佳分裂属性作为分支节点。利用一个好的取值来产生分支，不但可以加快决策树的生长，而且最重要的是，产生的决策树结构好，可以找到较好的规则信息。本文采用 Gini 指标对属性进行二元划分，选择 Gini 值最小的属性作为最佳分裂属性，构造分类决策树。

假设 A,B,C 三数据库中的记录数分别为 m_a,m_b,m_c，则属性 T 的 Gini 值为 $Gini_T(S)^{[1]}=\frac{m_a}{m}Gini_{Ta}(RS_a)+\frac{m_b}{m}Gini_{Tb}(RS_b)+\frac{m_c}{m}Gini_{Tc}(RS_c)$，其中 $Gini_{Ta}(RS_a)$ 表示属性 T 在数据库 A 中的 Gini 值。因为 Gini 指标考虑每个属性的二元划分，如果属性 T 的二元分裂将数据集 RS_a 分成两类 RS_{a1} 和 RS_{a2}，则给定该划分 RS_a 的 Gini 指标为 $Gini_{Ta}(RS_a)^{[1]}=\frac{RS_{a1}}{RS_a}Gini(RS_{a1})+\frac{RS_{a2}}{RS_a}Gini(RS_{a2})$。

2.2 总体架构

本文基于同态加密技术，提出一种可实现隐私保护的分布式决策树挖掘方法 DMBHE，其总体架构如图 1 所示。图中 HE 代表同态加密 HE(Homomorphic Encryption)。假设该架构模型满足以下条件：

1)所有数据记录以水平划分形式存储在各站点，且各站点之间不进行任何通信；

2)计算器只负责按照固定公式执行计算，对于各站点数据库信息、参加计算的数据及结果的含义均一无所知；

3)挖掘服务器负责数据挖掘，并将最终结果传输给各站点，而对于各站点的数据库信息一无所知。

在每次产生最佳分裂属性时，分为 3 步完成：第 1 步，各水平划分数据库在本地计算局部统计信息，并利用同态加密技术进行加密，然后传输给计算器；第 2 步，计算器获得加密信息后，按照固定公式求和，并传输给半可信第三方挖掘者；第 3 步，半可信第三方挖掘者依据同态解密算法进行解密，选择全局 Gini 值最小的属性作为最佳分裂属性，然后将其结果发送给各节点。产生当前最佳分裂属性后，各节点将本地训练集 RS_i 排列到适当的分支(即样例的该属性值对应的分支下)，再重复该过程。

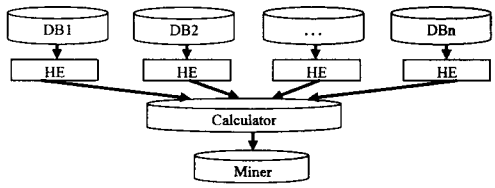

图 1 分布式决策树隐私保护挖掘的总体架构图

3 同态加密技术

同态加密 HE(Homomorphic Encryption)最初由 Rivest 等人于 1978 年提出，是一种允许直接对密文进行操作的加密变换技术。但是由于其对已知明文攻击是不安全的，后来由 Domingo 在文献[2]中做了进一步的改进。HE 技术最早用于对统计数据进行加密，由算法的同态性保证了用户可以对敏感数据进行操作但又不泄露数据信息。

定义 1 假设 E_{k1} 和 D_{k2} 分别代表加密、解密函数，明文数据空间中的元素是有限集合 $\{M_1,M_2,\cdots,M_n\}$，α,β 代表运算符。若 $\alpha(E_{k1}(M_1),E_{k1}(M_2),\cdots,E_{k1}(M_n))=\beta(D_{k2}(M_1),D_{k2}(M_2),\cdots,D_{k2}(M_n))$ 成立，则称函数族 $(E_{k1},D_{k2},\alpha,\beta)$ 为一个同态加密。

本文所要设计的同态加密算法，必须满足如下等式成立：

$D(E(m_a\times Gini_{Ta}(RS_a))+E(m_b\times Gini_{Tb}(RS_b))+E(m_c\times Gini_{Tc}(RS_c)))=m_a\times Gini_{Ta}(RS_a)+m_b\times Gini_{Tb}(RS_b)+m_c\times Gini_{Tc}(RS_c)$

即对于明文空间的任意 x,y,z，等式 $D(E(x)+E(y)+E(z))=x+y+z$ 均成立。

同态加密算法：

(1)选择两个大于 0 的安全素数 p,q，并计算 $n=p\times q$。

(2)对于明文空间的任意值 x，其加密值 $y=E(x)$，$E(x)$ 定义如下：

$y=E(x)=\mathrm{mod}((x+p),n)$，其中 mod 是求模运算。

同态解密算法：

对于给定的 $y=E(x)$，利用 p 作为密钥进行解密，$x=D(y)=\mathrm{mod}(y,p)$。证明基于上述同态加解密算法，对于明文空间的任意 x,y,z，等式 $D(E(x)+E(y)+E(z))=x+y+z$ 均成立。

证明：$D(E(x)+E(y)+E(z))=D((\mathrm{mod}(x+p),n)+(\mathrm{mod}(y+p),n)+(\mathrm{mod}(z+p),n))$

$=D(\mathrm{mod}(x+y+z+3p,n))=\mathrm{mod}(\mathrm{mod}(x+y+z+3p,n),p)$

$=\mathrm{mod}((\mathrm{mod}(x+y+z),n)+\mathrm{mod}(3p,n)),p)$

$=\mathrm{mod}((\mathrm{mod}(x+y+z),n),p)+\mathrm{mod}(\mathrm{mod}(3p,n),p)$

因为 $n=p\times q$，所以 $\mathrm{mod}(\mathrm{mod}(3p,n),p)=0$；

因为由解密算法可知 $D(E(x))=\mathrm{mod}((\mathrm{mod}(x+p),n),p)=x$，

所以 $\mathrm{mod}((\mathrm{mod}(x+y+z),n),p)=x+y+z$；

所以 $D(E(x)+E(y)+E(z))=x+y+z$ 成立。

4 DMBHE 算法描述

DMBHE 算法分为 3 个部分：客户端（各数据拥有方 RS_a，RS_b，RS_c）算法、计算器端和服务器端（半可信第三方挖掘者）算法。具体策略如下：

算法 1 DMBHE 客户端算法

输入：本地训练样例 RS_a（RS_b 或 RS_c）；

输出：与分支节点关联的加密各属性局部统计信息；

1）扫描训练样例 RS_a，寻找与当前分支节点相关联的样例 S_a 及属性集 $T_a=\{T_{a1},T_{a2},\cdots,T_{am}\}$；

2）根据 $Gini_{Tak}(S_a)=\frac{|S_{a1}|}{|S_a|}Gini(S_{a1})+\frac{|S_{a2}|}{|S_a|}Gini(S_{a2})$ 求解所有关联属性 T_{ak}（$n\geqslant k\geqslant 1$）的 Gini 值，其中 $Gini(S)=1-\sum_{i=1}^{m}p_i^2$（$p_i$ 是 S 中元组属于 C_i 类的概率，并用 $|C_i|/|S|$ 估计，对 m 个类计算和）；

3）求解局部信息数组 $T_a=\{Ta_1,Ta_2,\cdots,Ta_n\}$，其中 $T_{ak}=m_a\times Gini_{Tak}(S_a)$（$n\geqslant k\geqslant 1$）；

4）对数组 T_a 中各元素进行同态加密，形成加密局部信息数组 $T_a'=\{Ta_1',Ta_2',\cdots,Ta_n'\}$；

5）将加密局部信息数组传输给计算器；

6）接受服务器端传回的最佳分裂属性 T_k；

7）若传回的属性 T_k 不为空，创建树节点 t。

算法 2 DMBHE 计算器端算法

输入：A,B,C 三客户端发送的加密局部信息数组 T_a'，T_b'，T_c'；

输出：三数组各元素之和；

1）接受 A,B,C 三客户端发送的加密局部信息数组 T_a'，T_b'，T_c'；

2）求三数组各元素之和 $T'=\{T_1',T_2',\cdots,T_n'\}$，其中 $T_k'=T_{ak}'+T_{bk}'+T_{ck}'$（$n\geqslant k\geqslant 1$）；

3）将数组 $T'=\{T_1',T_2',\cdots,T_n'\}$ 传输给服务器（挖掘者）端。

算法 3 DMBHE 服务器（挖掘者）端算法

输入：数组 $T'=\{T_1',T_2',\cdots,T_n'\}$；

输出：最佳分裂属性 T_k；

1）接受计算器端信息数组 $T'=\{T_1',T_2',\cdots,T_n'\}$，其中 T_k'（$n\geqslant k\geqslant 1$）代表 $E(m_a\times Gini_{Ta}(RS_a))+E(m_b\times Gini_{Tb}(RS_b))+E(m_c\times Gini_{Tc}(RS_c))$；

2）利用密钥 p 运行同态解密算法，求解出明文 $T=\{T_1,T_2,\cdots,T_n\}$，其中 T_k（$n\geqslant k\geqslant 1$）代表 $m_a\times Gini_{Ta}(RS_a)+m_b\times Gini_{Tb}(RS_b)+m_c\times Gini_{Tc}(RS_c)$；

3）对所有关联属性 T_k（$n\geqslant k\geqslant 1$）进行排序，依据 Gini 定义，查找其值最小的属性 T_k；

4）将最佳分裂属性 T_k 发送给客户端。

在本算法中，各站点发送的是加密后的局部统计信息，保证了原始信息不泄露；计算器端接受的是加密数据，对于各站点数据库原始信息一无所知，由于不知道密钥 p，对于所计算的数据含义也一无所知；服务器端（挖掘者）收到的信息是加密后的全局统计结果，通过密钥 p，获取到解密后的全局统计结果，但对于任何局部信息或原始信息均无法获取，从而整个架构算法保证了原始信息的隐私性。

5 实验

为了验证本算法的可行性，我们在采用 VC++自行研制的知识发现软件系统 ICCKDSS（运行界面如图 2 所示）上实现上述算法。该系统运行硬件平台是主频 600Mz 的 Pentium 兼容的 PC机、64MB 内存、200MB 以上的可用磁盘空间、SVGA 监视器；软件平台是 WIN2000 Sever 操作系统、Microsoft SQL Server 数据库系统、C/S 结构的运行模式。

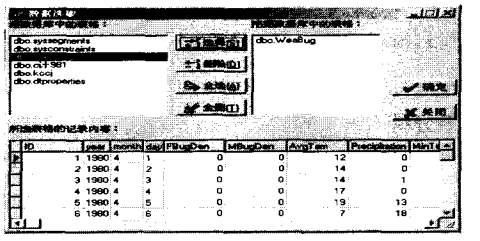

图 2 ICCKDSS 系统运行界面

我们对 UCI 机器学习数据集[10]上的 5 个数据集进行两组实验，一组用来测试 DMBHE 算法和传统决策树算法的挖掘效果比较，另一组用来测试 DMBHE 算法隐私信息的保护效果（因为某些敏感规则可倒推出部分原始信息），其实验结果如图 3 和图 4 所示。DMBHE 算法在最差的情况下所挖掘出的规则数能达到传统算法挖掘规则数的 85%，其隐私保护程度能达到保护所有隐私信息的 83%。

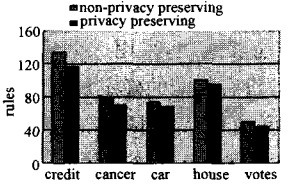

图 3 挖掘规则数比较

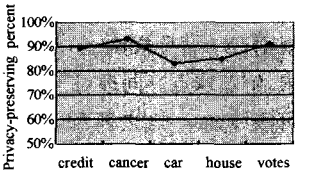

图 4 隐私信息保护效果

结束语 本文提出了一种可实现隐私保护的分布式决策树挖掘算法，本算法的核心思想是由本地站点生成加密的局部统计信息，计算器端只需进行简单求和，基于加解密算法同态，半可信第三方挖掘者解密后即可获得真实的全局统计信息。在整个过程中，任何参与一方均无法获取原始或真实的局部信息。通过理论分析及实验证明，本算法能在原始隐私信息不泄露的前提下，达到集中式决策树挖掘的效果。在以后的研究工作中，我们将其方法扩展到其他的挖掘任务中，如关联规则、聚类和异常点检测等等，并致力于研究出更高效的隐私保护算法。

参 考 文 献

[1] 韩家炜. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2001

[2] Domingo-Ferrer J, Herrera-Joancomarti J. A new privacy homomorphism and applications [J]. Information Processing Letters,

1996,60(5):277-282

[3] Xiang Guang-li, Chen Xin-meng. A Method of Homomorphic Encryption [J]. Wuhan University Journal of Natural Sciences, 2006,11(1):181-184

[4] Verykios V, Bertino E. State-of-the-art in Privacy-preserving Data Mining. SIGMOD,2004,33 (1)

[5] Rizvi S J, Haritsa J R. Maintaining data privacy in association rule mining [A]// Proceedings of the 28th International Conference on Very Large Databases. Hong Kong,2002:682-693

[6] Agrawal R, Srikant R. Privacy-preserving data mining [A]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. United States,2000:439-450

[7] Agrawal D, Aggarwal C. On the design and quantification of privacy preserving data mining algorithms [A]// Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Santa Barbara, United States: ACM Press,2001:247-255

[8] Cliffton C, Kantarcioglu M, Vaidya J. Tools for privacy preserving distributed data mining [J]. ACM SIGKDD Explorations Newsletter,2004,4(2):28-34

[9] Yang Bingru. Knowledge Discovery Theory Based on Inner Mechanism: Construction, Realization and Application. Elliott & Fitzpatrick Inc. USA,2004

[10] UCI Machine Learning Repository. http://www.ics.uci.edu/~mlearn/

[11] Pinkas B. Cryptographic techniques for privacy-preserving data mining [J]. ACM SIGKDD Explorations Newsletter, 2006, 4 (2):12-19

(上接第 234 页)

r_7: $f(x,\text{Speed\_radio}) \leqslant 1.08 \wedge f(x,\text{Dist\_radio}) \leqslant 1.02 \rightarrow x$ is success // 由 $\overline{B}(X_1)$ 中的 x_8 所支持,$C(r_8)=0.75$。

3.3 结果分析

为验证算法执行效果,用改进过截球策略的球队程序 Njust-2 与原球队程序 Njust-1 进行 6 场比赛。其中原球队 Njust-1 仅采用单一的二分法进行截球:在球员距球的距离小于一定阈值时用二分法求取截球点,执行截球动作,但不考虑该动作的成功率如何。选取两球队 6 场比赛中的两个样本数据进行实验分析。样本数据如表 2 所列。

表 2 截球样本属性表

样本	属性				
	BH_ang	Speed_radio	Dist_radio	Ang_radio	Stamina
X_1	21.9	0.54	0.13	0.79	2000
X_2	35.2	0.67	0.98	1.50	3500

对于样本 X_1,根据上文所提规则匹配流程,先查找确定规则,X_1 满足规则 r_1,说明截球能够成功,采用二分法建模求,解截球点。同理对于样本 X_2,没有匹配的确定规则,但有可能规则 r_7 与之匹配,且规则可信度大于阈值 δ(δ 取值 0.60),满足截球条件,采用二分法建模,求解截球点。

仿真结束后,记录下与截球相关信息和对比赛结果的影响,对比数据如表 3 所列。

表 3 实验对比结果

	平均截球成功率	进球总数
原代码	42.3%	7
改进后	58.9%	13

从表 3 中可以看出,采用传统二分法进行截球的球队,截球成功率在 42.3%,进球总数为 7 个;而采用本文所述方法改进截球策略后,对于截球成功率较高的场景才执行截球动作,使得其截球成功率在 58.9%,进球总数为 13 个。可见采用本文所提方法后,球员的截球成功率及球队的进球数得到了明显提高。

结束语 本文针对 RoboCup 仿真组比赛中智能体的核心技能之一——截球的决策问题,提出了一种基于优势关系粗糙集的截球策略,并在仿真实验中进行了决策分析。本方法将解析法和经验法相结合,不仅考虑了噪声对运动模型的干扰,而且针对与截球相关训练数据的特点,采用基于优势关系的约简算法和规则提取算法提取出更有现实意义的决策规则,进而实现更合理的决策。在仿真平台上的比赛实验中取得了较好的效果,证明了本方法的有效性与可行性。

参 考 文 献

[1] Hirokik, Minoru A, Yasuo K, et al. RoboCup: a challenge problem for AI and robotics[A]. Hirokik RoboCup-97: Robot Soccer World Cup I[C]. Berlin: Springer,1998:1-19

[2] 李实,陈江,孙增圻. 清华机器人足球队的结构设计与实现[J]. 清华大学学报:自然科学版,2001,41 (7):94-97

[3] 高隽. 人工神经网络原理及仿真实例[M]. 北京:机械工业出版社,2003:76-79

[4] 杨增光,李龙澍. 决策树学习在 Robocup 仿真球队中的应用研究[J]. 系统仿真学报,2004,16(4):653-656

[5] 刘扬,王浩,等. 一种基于支持向量回归方法在 RoboCup 中的应用 [J]. 合肥工业大学学报:自然科学版,2007,30(10):1258-1264

[6] Greco S, Matarazzo B, Slowinski R. Rough approximation by dominance relations [J]. International Journal of Intelligent Systems,2002,17:153-171

[7] Greco S, Matarazzo B, Slowinski R. Rough sets theory for multicriteria decision analysis [J]. European Journal of Operational Research,2002,129:1-47

[8] Yang X B, Yang J Y, Wu C, et al. Dominance-based rough set approach and knowledge reductions in incomplete ordered information system[J]. Information Sciences, 2008, 178(4): 1219-1234

[9] Chen M, Foroughi E, et al. Users Manual: RoboCup Soccer Server for Soccer Server Version 7.07 and later,2002. http://sserver.sourceforge.net/

[10] 何光渝. VisualC++常用数值算法集[M]. 北京:科学出版社,2002:528-530

[11] 清华大学毕业设计论文[OL]. 北京:http://www1.cs.columbia.edu/~criver/robocup/criver_thu_thesis.pdf