隐私保持的决策树分类挖掘

李 光,王亚东,苏小红
(哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 隐私保持的数据挖掘是目前数据挖掘领域的重要研究方向之一,其首要研究内容是开发在不泄露隐私数据的前提下进行数据挖掘的方法.决策树是分类挖掘的一种重要方法,也是目前隐私保持的数据挖掘领域中少有的被深入研究了的分类方法.针对目前尚未对隐私保持的决策树挖掘方法进行系统总结的问题,本文对该领域进行综述.首先对问题背景进行介绍,随后介绍了该领域的研究现状,对现有方法进行了分类和总结,最后总结出该领域进一步研究的方向.

关键词: 数据挖掘; 隐私保持; 决策树
中图分类号: TP309 **文献标识码:** A **文章编号:** 0372-2112 (2010) 01-0204-09

Privacy Preserving Data Mining on Decision Tree

LI Guang, WANG Ya-dong, SU Xiao-hong
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*)

Abstract: Now, the privacy preserving data mining has become an important research problem. The chief research purpose of it is to find data mining methods which can work well without opening the privacy data. The decision tree is an important method for classification, and has been studied in depth for privacy preserving data mining. There was on good summarize for privacy preserving decision tree algorithm. To solve this problem, this paper gave a survey on the current privacy preserving decision tree data mining techniques. It first introduced the background of privacy preserving decision tree data mining; then introduced the studying status of this subject and gave a summarize; finally, it pointed out some future research directions.

Key words: data mining; privacy preserving; decision tree

1 引言

数据挖掘也被称为知识发现,旨在从海量数据中发现人们难以察觉却又感兴趣的知识.经过十几年的发展,数据挖掘已经成为研究界与工业界关注的热点之一,并且在实践中展现出了强大的能力.但是,数据挖掘也为数据的隐私保护带来了巨大挑战[1,2].如何在成功进行数据挖掘的同时保护隐私数据不被泄露成为了一个重要的问题.

为解决这一问题,提出了隐私保持的数据挖掘[3~5],目前它已成为数据挖掘研究的重要内容之一.传统的数据挖掘方法不考虑数据隐私保护的问题,假设所有数据都可以直接得到,这个假定在实际中是不现实的,出于隐私保护的需要,某些数据是不能公开的.隐私保持的数据挖掘克服了这一不足,它的首要研究问题是如何在不公开精确数据的前提下得到精确的数据挖掘结果.隐私保持的数据挖掘在医学,金融,国家安全,商业等存在大量隐私数据的领域有着广阔的应用前景.

分类是数据挖掘的重要内容,它通过学习已知类别的数据,来获得对未知类别的数据进行类别预测的能力.决策树是一种重要的分类方法,也是隐私保持分类挖掘的首选原型方法.目前,在隐私保持的数据挖掘研究中,该方法是少有的被深入研究了的分类方法.本文对隐私保持的决策树挖掘方法进行综述.

隐私保持的决策树挖掘研究主要还是针对传统决策树进行的,大都选用 ID3 类算法作为原型算法.这类算法的关键在于选取最佳的分类属性,一般通过计算并比较信息增益来实现,也有使用基尼指数(Gini index)的.隐私保持的决策树挖掘技术主要分为两类.一类是基于数据扰动的方法,另一类是基于安全多方计算的方法,本文将分别予以介绍.另外,决策树挖掘的一个新方向,惰性决策树(lazy decision-tree),也受到了研究者们的关注,本文也将介绍隐私保持的惰性决策树挖掘.

隐私保持的决策树挖掘方法的设计原则是精确、安

收稿日期:2008-07-15;修回日期:2009-09-24
基金项目:国家 863 高技术研究发展计划(No.2007AA02Z329,No.2007AA02Z302)

全、高效三者的协调与统一.精确是指最终能够得到准确的数据挖掘结果;安全是指对隐私数据提供了足够的保护;高效是指通信开销和时间复杂度小.这三方面的要求在现实中往往相互矛盾,要追求平衡和兼顾.

2　基于数据扰动的方法

在基于数据扰动的隐私保持的决策树挖掘方法中,真实数据不被公开,公开的是一组经过扰动的数据.这类方法一般具有可接受的时间复杂度和通信开销,因此主要关注精确和安全.即要求在扰动后的数据上构建的决策树分类精度高,又要求难以从扰动后数据得到真实数据的值.这两个要求往往是矛盾的.调和这两个相互矛盾的要求就成了此类方法研究的关键.

这类方法主要有加随机噪声,基于随机响应(randomized response)以及基于 K-匿名(K-Anonymity)这三种方法.另外也有研究者使用基于函数或奇异值分解的方法来进行扰动.

2.1　加随机噪声的方法

2000 年,Rakesh Agrawal 等人[6]提出用加随机噪声的方法来进行隐私保持的决策树挖掘.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,数据的使用者可以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.Rakesh Agrawal 等人[6]使用一种基于贝叶斯理论的迭代方法来估计真实数据 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.

Rakesh Agrawal 等人[6]的方法主要针对连续型属性.对于连续型属性,生成决策树的关键是选择一个属性 A 以及该属性的一个值 t,将样本集合 S 分裂为两个集合 S_1 和 S_2,分别包含 $A \leqslant t$ 和 $A > t$ 的样本,使得这种分裂带来的信息增益最大.如果知道了属性 A 的分布,就可以计算 $A \leqslant t$ 和 $A > t$ 的概率,又因为 S 的基数已知,因此就可以估算 S_1 和 S_2 的基数.将扰动后样本以属性 A 的值排序,属性 A 较小的 $|S_1|$ 个样本构成集合 S_1,剩余样本构成了 S_2.又因为分类属性不被认为是隐私数据,不被扰动,因此可以使用属性 A 的分布来估算这种分裂带来的信息增益,从而完成决策树的生成.实验表明,该方法可以得到精度很高的决策树.

2001 年,Dakshi Agrawal 等人[7]采用了同样的数据扰动方法,通过一个 EM 算法来重构 X 的分布,并给出了一种更为合理的基于信息论的隐私保护程度的度量指标.该 EM 算法收敛于 X 分布的极大似然估计,实验表明其精度高于基于贝叶斯的迭代算法.

2003 至 2005 年,H.Kargupta 等人[8,9]对于这种加随机噪声的方法的安全性提出了质疑.认为该方法不足以保护隐私数据,并基于随机矩阵理论,提出了一种从扰动后数据上估计真实数据的方法.2003 年,H Dutta 等人[10]将这种恢复真实数据的方法分别应用于经过随机扰动的连续型数据,购物篮数据以及图数据,实验表明,这种攻击方法具有广泛的适用性,在这三种数据上都相当有效.2006 年,Songtao Guo 等人[11]给出了该方法的误差上界,使得攻击者可以判断估计出的数据与真实数据的差距有多大.

2005 年,Zhengli Huang[12]对从加随机噪声的扰动数据上恢复真实数据的问题进行了深入研究.认为 H.Kargupta 等人[8,9]的方法成功的关键是充分的利用了各属性间的相关性.Zhengli Huang[12]提出了基于主成分分析以及基于贝叶斯理论的恢复真实数据的方法,并对各种方法进行了实验对比,发现当数据关联较为紧密时,H.Kargupta 等人[8,9]的方法效果较好.

2006 至 2008 年,Li Liu 等人[13,14]对原始的加随机噪声的方法进行了改进.提出了一种两段式方法.首先给真实数据加一个随机噪声,然后将加噪声数据的值域划分为多个区间,在加噪声数据落入的区间中随机选取一个数值作为最终的公开数据.区间的宽度可以由用户来控制.该方法通过引入更多的随机性来削弱数据的关联,从而抵抗了 H.Kargupta 等人[8,9]提出的攻击方法,并提供给用户更大的自主权,增强了方法的灵活性.Li Liu 等人[13,14]依然使用先恢复真实数据分布,然后利用该分布生成决策树的方法.

2009 年,Li Liu 等人[15]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 S_1 和 S_2,该方法的核心思想是对每个样本,计算它属于 S_1 和 S_2 的概率 p_1 和 p_2,并利用计算出的概率值来生成决策树,从而避免了估计真实数据的分布这一复杂的工作.Li Liu 等人[15]提出了两种生成决策树的方法,一是设立一个阈值,那些属于 S_1 的概率 p_1 大于该阈值的样本构成了 S_1,其余样本构成 S_2.二是借鉴了模糊数学的做法,样本分别以 p_1 和 p_2 的隶属度属于 S_1 和 S_2,在计算信息增益以及生成决策树时也要作相应调整.

加随机噪声的方法主要适用于连续型属性.它对某个样本的扰动不需要得到其他样本的信息,不需要可信任第三方来扰动数据.因此它可以应用在数据收集阶段,而且对于流数据也适用.尽管已经开发出了一些对于它的攻击方法,但也研究出了一些补救的手段.

2.2　基于随机响应的方法

2003 年,Wenliang Du 等人[16]提出了一种基于随机响应的数据扰动方法.该方法主要针对布尔型变量,其

他类型的变量在使用该方法时需要进行布尔化.在该方法中,设真实数据为 X,$X=0$ 或 1,设定参数 $\theta(0\leqslant\theta\leqslant1$ 且 $\theta\neq0.5)$,X 以 θ 的概率保持不变,以 $1-\theta$ 的概率取反,从而得到新的公开数据 Y.在该方法中,同一个样本中的不同属性要么都不变要么都取反,例如,一个有三个属性的样本"010",以 θ 的概率不变,仍为"010",以 $1-\theta$ 的概率取反,变为"101".Wenliang Du 等人[16]采用解方程组的方法来得到真实数据 X 的分布,并利用该分布得到决策树.另外还设计了利用扰动后测试集来计算决策树在真实测试集上分类正确率的方法,以解决决策树剪枝的问题.

2006 年,葛伟平等人[17]对该方法进行了拓展.使之不再局限于布尔型变量,可以适用于离散型变量,同时不再要求同一个样本中的不同属性采用相同的改变方式.设属性 A 在样本集 D 中一共有 m 个取值,分别为 $a_1,a_2,\cdots,a_m$.使用一个可逆的转移概率矩阵 P_A 来对 A 进行扰动.设

$$P_A=\begin{bmatrix}p_{11}&p_{12}&\cdots&p_{1m}\\p_{21}&p_{22}&\cdots&p_{2m}\\\vdots&\vdots&\cdots&\vdots\\p_{m1}&p_{m2}&\cdots&p_{mm}\end{bmatrix}$$

若 A 在真实数据中取值为 a_i,则它以 p_{ij} 的概率被扰动为 a_j.Wenliang Du 等人[16]的方法可以看成是这种方法的特例,相当于 $P_A=\begin{bmatrix}\theta&1-\theta\\1-\theta&\theta\end{bmatrix}$ 时的情况.

由于 P_A 是可逆的,使用解方程组的方法便可以求出真实数据分布进而生成决策树,可以使用最小描述长度准则进行剪枝.

2008 年,Piotr Andruszkiewicz[18,19]使用了葛伟平等人[17]的扰动方法,分别使用迭代过程和解方程组的方法来恢复真实数据的分布.Piotr Andruszkiewicz[18,19]不是直接使用重构出的真实数据的分布来生成决策树,而是利用重构出的真实数据的分布估计样本的真实值,然后再生成决策树.

基于随机响应的扰动方法适用于布尔型变量,经拓展后也可用于离散型变量.它在对某个样本进行扰动时也不需要其他样本的信息,不需要可信任第三方.该方法实际上在关联规则挖掘中应用更为广泛,因为很多隐私保持的关联规则挖掘研究都基于布尔型的购物篮数据,比较适合该方法.

2.3 基于 *K*-匿名的方法

K-匿名是一种常用的数据匿名化的方法[20],它还有一些改进型的变种,如 *L*-差异(L-Diversity)[21]等.*K*-匿名方法主要作用于准键(quasi-identifier),主要针对的是重标识攻击(re-identification).

在公开含有隐私信息的数据时,为了保护隐私,常将一些能够进行个人定位的属性隐藏不予公开.例如,在公开医疗档案数据时,将患者的姓名,身份证号等从数据库中删除,以防止将健康情况等隐私信息和某个个人联系起来.但仅仅隐藏这部分属性是不够的,某些属性的组合也能起到很好的定位效果,比如性别,邮编,受教育程度,年龄.可能单独的每个属性都不能很好的定位到某个个人,但它们组合起来却有很好的定位效果.这种属性组合就被称为准键,利用他们的组合来将隐私信息与某个个人联系起来,从而造成隐私泄露的攻击方法就被称为重标识攻击.

K-匿名方法通过对数据作一些扰动,使得数据库中出现的每个准键值都至少对应 K 个元组.这样在使用重标识攻击时,至少有 K 个元组是无法区分的,从而防止了这种攻击方法.传统的 *K*-匿名方法要求扰动造成的信息丢失最小.尽量使得扰动后数据和真实数据差别不大,以便在扰动后数据上直接进行数据挖掘.

一个新的研究趋势是针对数据的应用来进行 *K*-匿名,不再要求扰动造成的信息丢失最小,而是要求扰动对数据应用的影响最小.2007 年,Benjamin C M Fung 等人[22]使用一种自顶向下的方式来实现对于数据的 *K*-匿名处理.在处理过程中考虑了决策树的生成,使用启发式方法搜索合适的扰动方案,使得扰动过程对于 ID3 类算法的影响尽可能的小.最近,S Kisilevich 等人[23]先在真实数据上生成决策树,然后用这个决策树来指导 *K*-匿名处理.实验表明,这两种方法的效果都很好,在 *K*-匿名处理后的数据上可以生成精度很高的决策树.

K-匿名方法对于连续型和离散型数据都适用,在数据隐私保护方面是很受研究者们重视的一种方法.该方法最大的优点是它的灵活性,它并没有规定数据扰动的过程,而只是规定了扰动需要达到的目标,具体过程可以根据需要来选择.最近,研究者们开始利用这一点开发针对应用(例如决策树)的 *K* 匿名处理方法.

2.4 基于函数或奇异值分解的方法

针对特殊的问题环境和要求,研究者们还设计了其他一些数据扰动的方法.主要是基于函数或奇异值分解的方法.2005 年,Nan Zhang 等人[24]提出了一种基于奇异值分解的数据扰动方法.现实中,数据提供者希望数据的安全性好,而数据使用者希望数据的可用性高.Nan Zhang 等人[24]提出的方法具有协商机制,可以让数据提供者和使用者协商合适的扰动水平,只有在双方达成一致的情况下,数据传输才会进行,从而解决了这一矛盾.

该方法针对样本分两类的情况.首先对样本非类别属性进行编码,使得所有样本都成为等长的布尔值向量.将样本排列为矩阵,矩阵的每一行对应一个样

本.设数据使用者 P 已经得到了一部分数据,它们构成矩阵 T,T_0 和 T_1 分别表示不同类别的样本,令 $A_0 = T'_0T_0$,$A_1 = T'_1T_1$,其中,T'_0 和 T'_1 分别为 T_0 和 T_1 的转置,令 $A = A_0 - A_1 = T'_0T_0 - T'_1T_1$.对 A 进行分解,得:$A = V\Sigma V'$,其中 Σ 为由 A 的特征值 $\lambda_1,\lambda_2,\cdots,\lambda_n$($\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$)构成的对角阵,$V$ 是由 A 的特征向量构成的矩阵.设 V_k 为 A 的前 k 个特征向量构成的矩阵,即由最大的 k 个特征值对应的特征向量构成的矩阵.数据使用者 P 将 V_k 传给数据提供者 Q,Q 拥有样本 x,令 $x' = xV_kV'_k$,x' 是一个实数值向量,将 x' 进行布尔化便得到了扰动后数据 y,将 y 传给 P,P 拥有的样本集就成为了 $T \cup \{y\}$.反复进行这个过程,P 最终便得到了所有的样本.在该方法扰动后数据上可以直接使用 ID3 算法生成决策树.

该方法中 k 的大小决定了扰动水平.Nan Zhang 等人[24]给出了不同 k 值下扰动的误差估计,并据此提供了协商机制,使得数据提供者和数据使用者协商选取 k 值.2007 年,Shaofeng Bu 等人[25]给出了一种基于函数的数据扰动方法.它针对的问题环境是缺乏数据挖掘能力的数据拥有者 P 拥有数据 X.P 和数据挖掘服务供应商 Q 进行合作.P 向 Q 提供数据,由 Q 完成挖掘.要求 P 最终得到挖掘结果,Q 无法得到真实数据和挖掘结果.

该方法设计一个函数 f,用真实数据 X 对应的函数值 $f(X)$作为扰动后的数据 Y.P 将 $Y = f(X)$传递给 Q,Q 生成 Y 上的决策树 T',将其传给 P.由 P 将 T' 还原为 X 上的决策树 T.由于 Q 不知道 f 的具体形式,因此 Q 无法得到真实数据和最终的挖掘结果.

该方法采用反函数变换法来将扰动数据 Y 上的决策树 T' 还原为真实数据 X 上的决策树 T.具体来讲,若 T' 中一个非叶子节点对应属性 A,连接它和它的一个儿子节点的边对应属性值 a_i,A 对应的扰动函数为 f_A,其反函数为 g_A,则将这条边对应的属性值改为 $g_A(a_i)$.

函数 f 必须满足两个条件.首先,f 必须使得反函数变换法能够使用,其次,f 必须足够复杂,这样才能为隐私数据提供足够的保护.Shaofeng Bu 等人[25]将 f 设计为一个分段函数并给出了生成 f 的方法.

这两种方法是为了满足特殊的环境或要求而设计出的.数据扰动方法相当灵活,可以针对不同的需求和环境来设计.

3　基于安全多方计算的方法

安全多方计算主要研究一组互不信任的参与方之间保护隐私的协同计算问题,其基本要求是要确保输入的独立性,计算的正确性,同时不泄露输入值给参与计算的其他成员.

基于安全多方计算的隐私保持的决策树挖掘大多应用于分布式数据库.在该环境中,数据分布式存储在多个节点上,要求在每个节点都无法得到其他节点详细数据的前提下进行全局数据上的数据挖掘工作.在理想情况下,存在一个信任第三方,这时,只要各数据节点将各自的私有数据传递给信任第三方,由信任第三方来进行挖掘,并将最终结果传递给各数据节点就可以了.实际中往往不存在信任第三方时,经常使用基于安全多方计算的方法来交流必要的信息,隐私保持的完成挖掘.

这类方法使用基于安全多方计算的协议来交流挖掘算法需要的信息,使得挖掘算法可以在不直接交流样本的情况下运行.一般有很好的准确性,最终得到结果往往与不考虑隐私保护,直接在全局数据上进行挖掘是一样的.关注的重点是安全和效率.在挖掘的过程中要使得任何一个节点都无法得到其他节点的样本数据,而且要尽可能的减少算法的通信开销和计算时间.

分布式数据库分为两种类型,水平型和垂直型.水平型分布式数据库中,每个节点上仅包含一部分元组,但每个元组都是完整的,包含所有的属性.垂直型分布式数据库中每个节点上都包含所有元组,但每个元组都是不完整的,仅包含一部分属性.针对不同类型的分布式数据库,有不同的隐私保持决策树挖掘方法.

3.1　针对水平型分布式数据库的方法

在水平型分布式数据库中,数据节点拥有完整样本.这样完成 ID3 类算法的决策树挖掘需要解决的就有三个问题:首先如何选择最佳分类属性,在 ID3 类算法中归结为如何计算某个属性的信息增益;其次,如何计算样本集中出现最频繁的类别;最后,如何判断样本集是否拥有单一的类别.

2002 年,Yehuda Lindell 等人[26]针对两个节点的水平型分布式数据库,以 $ID3_\delta$ 算法(标准 ID3 算法的一个变种)为原型,设计了一种隐私保持的决策树生成算法.该方法使用不经意传输协议[27]、不经意多项式求值[28]以及两方安全计算协议[29]设计了 $x\ln x$ 协议,用该协议来计算某个属性的信息增益,然后使用两方安全计算协议确定信息增益最大的属性.在 $x\ln x$ 协议中,两个节点 A 和 B 分别输入 v_1 和 v_2,分别得到 w_1 和 w_2,满足 $w_1 + w_2 = (v_1 + v_2)\ln(v_1 + v_2) \bmod(|F|)$,其中 F 是一个域.该方法采用两方安全计算协议计算样本集中出现最频繁的类别.采用文献[28]和[30]的方法或使用两方安全计算协议来判断样本集是否拥有单一的类别.Yehuda Lindell 等人[26]的方法与不考虑隐私保护直接在全局数据上运行 $ID3_\delta$ 算法是等价的,训练结果一样.该方法在两个节点的水平型分布式数据库上具有可接受的计算复杂度和通信开销.算法运行过程中,任

何一个节点都得不到其他节点上样本的详细数据.

2002 年,Benny Pinkas[31]指出由于过高的时间复杂度和通信开销,该方法并不适用于多个节点的水平型分布式数据库.

2005 年,Ming-Jun Xiao 等人[32]针对多节点的水平型分布式数据库,提出了一种基于同态加密[33]的隐私保持的 ID3 算法.该方法使用同态加密对 $x\ln x$ 协议进行改造,并利用改造后的 $x\ln x$ 协议来计算信息增益.该方法还使用同态加密和置换函数设计了已知信息增益后计算最优分类属性的 FMP 协议.FMP 协议也用于计算某个样本集中出现最频繁的类别.该方法通过重复多次使用 Yehuda Lindell 等人[26]中的方法来判断样本集是否具有单一类别.Ming-Jun Xiao 等人[32]的方法成功地解决了在多个节点的水平型分布式数据库上隐私保持的训练决策树的问题.2006 年,Ming-Jun Xiao 等人[34]将这一思路应用于 *C*4.5 算法.针对多个节点的水平型分布式数据库,设计了一种基于同态加密的隐私保持 *C*4.5 算法.

2007 年,F Emekci 等人[35]以安全和协议为基础建立隐私保持的 ID3 算法.生成决策树可以归结为计算某些属性取确定值的样本的个数.在水平型分布式数据库中,对各节点上满足属性取值要求的样本的个数求安全和,便可以隐私保持的得到满足这些要求的总的样本个数,进而生成决策树.F Emekci 等人[35]就是利用这一原理,在安全和协议的基础上建立了隐私保持的 ID3 算法.

在安全和协议中,设一共有 n 个节点 $P_1, P_2, \cdots, P_n$,分别拥有输入 $v_1, v_2, \cdots, v_n$,协议运行后,要求每个节点除了自己的输入外,只能得到 $\sum_{i=1}^{n} v_i$,无法得到其他节点的输入.这是一个安全多方计算的经典问题,有很多解决方法.F Emekci 等人[35]设计了一种基于沙米尔分享(Shamir's secret sharing)[36]的安全和协议.从而解决了在多个节点的水平型分布式数据库上进行隐私保持的决策树挖掘的问题.

2008 年,Saeed Samet 等人[37]使用基于同态加密的方法设计了一种针对多个节点的水平型分布式数据库的隐私保持的 ID3 算法.该方法在选择最佳分类属性时,使用了基尼指数而不是最常见的信息增益.

3.2 针对垂直型分布式数据库的方法

在垂直型分布式数据库中,一般假定所有节点都知道样本的类别,这时寻找最佳分类属性就成为了决策树学习的关键,在 ID3 类算法中这又归结为计算某个属性的信息增益.

2002 年,Wenliang Du 等人[38]针对两个节点的垂直型分布式数据库设计了一种基于安全内积协议的隐私保持的 ID3 算法.

在安全内积协议中,两个节点 P_1, P_2 分别输入向量 v_1, v_2,设 $w = v_1 \cdot v_2$ 是 v_1 和 v_2 的内积.要求在运行协议后,每个节点都得到了 w,但不知道另一个节点的输入是多少.安全内积协议也是一个安全多方计算的经典问题.Wenliang Du 等人[38]使用了一种基于半信任第三方的安全内积协议.

正如前面指出的,生成决策树可以归结为计算若干属性取确定值的样本的个数.在垂直型分布式数据库中,这可以通过求向量内积来完成.设 A 为某些属性的集合,现在要计算在 A 上取指定值的样本的个数.设 B_1 和 B_2 分别表示节点 P_1 和 P_2 上的属性,令 $A_1 = B_1 \cap A, A_2 = B_2 \cap A$.假定一共有 n 个样本,P_1 和 P_2 分别在本地数据上计算 n 维向量 V_1 和 V_2,V_1 的第 i 个分量为 1 当且仅当第 i 个样本在 A_1 上取指定值,V_1 的其他分量都取 0.相应的,V_2 的第 i 个分量为 1 当且仅当第 i 个样本在 A_2 上取指定值,V_2 的其他分量也都取 0.这样,V_1 和 V_2 的内积便是在 A_1 和 A_2 上都取指定值,即在 A 上取指定值的样本的个数.这便是 Wenliang Du 等人[38]生成决策树的基本思想.

Wenliang Du 等人[38]还设计了一种基于安全内积协议的隐私保持的对未分类样本进行类别预测的方法.设 X 是一个没有标定类别的样本,X 垂直型分布式存储在两个节点 P_1 和 P_2 上,P_1 和 P_2 都知道决策树 T,使用该方法,P_1 和 P_2 可以用决策树 T 对 X 的类别做出预测,而且 P_1 和 P_2 都得不到对方拥有的关于 X 的数据.这种隐私保持的类别预测方法可以用于对决策树进行剪枝.Wenliang Du 等人[38]的方法仅适用于两个节点的垂直型分布式数据库,对多个节点的情况不适用.2007 年,Justin Zhan[39]提出了一种基于同态加密的隐私保持 ID3 算法,该方法适用于有多个节点的垂直型分布式数据库.

在同态加密中,使用 x 的密文 $E(x)$ 和 y 的密文 $E(y)$ 便可以计算 $x + y$ 的密文 $E(x + y)$.Justin Zhan[39]充分利用这一点,设计了在多个节点的垂直型分布式数据库上,计算某属性信息增益的密文的方法.这时计算出的只是信息增益的密文,无法直接比较大小.Justin Zhan[39]又基于同态加密设计了一个在仅知密文的情况下比较信息增益大小的方法.从而完成了决策树的生成,解决了在多个节点的垂直型分布式数据库上隐私保持的进行决策树挖掘的问题.

2009 年,Yanguang Shen 等人[40]针对两个节点的垂直型分布式数据库给出了一种隐私保持的 *C*4.5 算法.它基于安全内积协议和 $x\log(x)$ 协议来完成 *C*4.5 算法要求的信息增益率的计算,进而选择最佳分类属性并

生成决策树.

基于安全多方计算的方法普遍被认为具有较高的安全性,而且一般在准确性上也有保证.但它也有很多的不足.首先,因为要完成复杂的密码学操作,它往往需要较多的计算时间和通信开销,这使得它很难应用在大规模数据上.其次,为了提高算法的效率,往往对节点的行为做出限制,比如这类方法经常假定节点都是好奇而非恶意的,即节点都会遵守协议,不会相互串通,不会提供不真实的输入,只会从协议允许得到的信息上来推断敏感数据.这些假定在简化算法的同时也限制了算法的适用范围.再次,算法的灵活性不足.精确、安全、效率,这三个隐私保持数据挖掘的要求往往是相互矛盾的.基于数据扰动的方法一般可以通过参数的改变来调节扰动程度,从而在相互矛盾的要求中根据需要进行灵活的折中.而基于安全多方计算的方法缺乏这一机制.

4　隐私保持的惰性决策树挖掘

一般的决策树挖掘属于急切学习,对整个训练数据建立一般适用性模型,时间开销较大,而且由于一般适用性模型往往针对全局进行优化,因此对于某些特定样本不一定适用.惰性决策树是决策树研究的一个新方向,它将惰性学习引入决策树中,兼顾了急切学习与惰性学习的优点.

惰性决策树挖掘是在得到了待分类样本后进行的,它并不生成完整的决策树,仅生成决策树对待分类样本进行预测所使用的特定分枝,由这条分枝给出预测结果.惰性决策树已经引起了隐私保持的数据挖掘研究者们的关注.

2005 年,Pragnesh Jay Modi 等人[41]针对多个节点的垂直型分布式数据库,提出了一种隐私保持的惰性决策树挖掘算法.该算法中,用一个打分函数给尚未生长完全的决策树打分,原则是越简单的决策树分数越高.每个节点都拥有一个起初仅包含空树的决策树集合,每个节点都根据自己的本地数据来使用 ID3 算法逐层展开本地的分数最高的决策树,每展开一层都仅保留待分类数据所在的那个分枝.每次展开后各节点都要将结果广播给其他的节点,并要将其他节点传递来的决策树加入本地的决策树集合.当所有节点的分数最高的决策树都无法展开时,学习过程结束,每个节点使用本地分数最高的决策树分别给出待分类数据的分类预测.由于该算法中各个节点并不交流各自的私有数据,各个节点间的信息交流是通过相互传递决策树来完成的,因此对私有数据提供了保护.实验表明,该方法有相当高的预测精度.

2008 年,Geetha Jagannathan 等人[42]针对两个节点的水平型分布式数据库,提出了一种惰性决策树挖掘方法.该方法的原型算法是一种基于 ID3 算法的惰性决策树算法,该方法通过使用基于安全多方计算的协议使得该惰性决策树算法得以在保护隐私数据的前提下运行.最终所得的结果与在完整数据集上运行该惰性决策树算法所得结果完全一致.

隐私保持的惰性决策树的出现表明隐私保持数据挖掘的研究者们并没有局限于传统算法,而是积极地将隐私保持数据挖掘的思想应用于各种数据挖掘研究的新成果.

5　总结与展望

隐私保持的数据挖掘是目前数据挖掘领域的一个热点研究方向,其首要研究任务是开发在不共享精确数据的前提下,获取准确的数据挖掘结果的方法.在隐私保持的分类挖掘中,多选用决策树作为原型方法.目前,隐私保持的决策树挖掘主要针对传统决策树,主要有基于数据扰动和基于安全多方计算这两类方法.另外,作为决策树挖掘的一个新方向,惰性决策树也受到了研究者们的关注.

在基于数据扰动的方法中,真实数据被隐藏,公开的是一组扰动数据,通过处理扰动数据来得到真实数据上的分类器.常用的有三种扰动方法:加随机噪声的方法,基于随机响应的方法和基于 K-匿名的方法.另外还有一些针对特殊要求和问题设计的扰动方法,如基于函数或奇异值分解的方法.其中,加随机噪声的方法适用于连续型数据,该方法对某个样本的扰动不需要参考其他样本的信息,因此它可以应用在数据收集阶段,而且对于流数据也适用.基于随机响应的方法主要适用于布尔型变量,经拓展后也可用于离散型变量.它对某个样本进行扰动时也不需要其他样本的信息.基于 K-匿名的方法主要用于防止重标识攻击,它对于连续型和离散型数据都适用.而且由于只规定了扰动需要达到的目标,没有规定扰动的过程,因此具有很好的灵活性,扰动的具体过程可以根据需要来选择.

基于安全多方计算的方法大多应用于分布式数据库.普遍认为这类方法具有较高的安全性,而且一般准确性也有保证.但这类方法普遍效率不高,时间和通信开销都较大.而且,为了提高算法效率,经常对节点的行为做出限制,这些限制在简化算法的同时也限制了算法的适用范围.另外,算法的灵活性不足,缺乏在相互矛盾的要求中根据需要进行折中的机制.

通过对国内外已有工作的调研、分析和总结,本文认为隐私保持的决策树挖掘的后续研究应该从以下几个方向入手.

(1)目前,在加随机噪声的方法中,添加的噪声期望

一般为零,而且各个属性添加的噪声一般是相互独立的.现有的攻击方法大多针对这一特性进行攻击.因此可以考虑添加其他类型的噪声以对抗现有的攻击方法.

(2)基于随机响应的方法一般用解方程组的办法来恢复真实数据分布.在数据隐私保护程度较高时,待求解的方程组是病态的,导致数据可用性急剧下降.数学中对非适定问题已有广泛研究,得到了一些成果,例如正则化方法等.引入这些成果,应该可以在数据隐私保护程度较高时,提高数据的可用性.

(3)基于安全多方计算的方法在效率上比较成问题,难以应用于大规模数据.可以考虑将数据扰动方法与基于安全多方计算的方法结合起来,用数据扰动来代替某些密码学操作,以降低需要的计算时间和通信开销.

(4)现有的研究成果主要针对的是 ID3 算法,对于 *C*4.5 算法也有涉及.而决策树作为一个经典分类方法,并不局限于这两种算法.有必要将决策树学习中的各种研究成果,例如针对大样本集的 SLIQ 和 SPRINT 算法,都引入到隐私保持的决策树挖掘中来.

(5)数据挖掘的结果包含了真实数据的信息,有可能会造成隐私泄露.这有两重含义,一是某些挖掘结果本身就是隐私数据,是数据拥有者不愿意被旁人所知的;二是某些挖掘结果会有助于推断真实隐私数据.因此需要研究挖掘结果对数据隐私保护的影响.目前,这方面的研究在分类挖掘方面还很薄弱.2004 年,Murat Kantarcioglu 等人[43]做了一些探索性的工作,2008 年,Arik Friedman 等人[44]提出了一种不会破坏训练样本集 *K*-匿名性质的决策树挖掘算法.在这方面还需要做进一步的研究.

(6)现在还没有一个很好的安全性评价方法.现有的安全性评价,都称不上严格.应该借鉴密码学的做法,建立基于计算复杂性理论的安全评价标准.

(7)目前,隐私保持的决策树挖掘算法种类繁多且没有合适的工具软件.这种状况不利于在实际中推广和应用该项技术.因此有必要进行不同算法间的比较整合以及标准化的工作,进而进行工具软件的开发和应用.

参考文献:

[1] Herman T Tavani. Information privacy, data mining, and the internet[J]. Ethics and Information Technology, 1999, 1(2): 137 – 145.

[2] A Cavoukian. Data mining: staking a claim on your privacy [OL]. http://www.ipc.on.ca/english/Resources/Discussion-Papers/Discussion-Papers-Summary/?id=342. 1998-01-01/2009-09-22.

[3] Vassilios S Verykios, Elisa Bertino, Igor Nai Fovino, Loredana P Provenza, Yucel Saygin, Yannis Theodoridis. State-of-the-art in privacy preserving data mining[J]. ACM SIGMOD Record, 2004, 33(1): 50 – 57.

[4] 张锋,孙雪冬,常会友,赵淦森.两方参与的隐私保护协同过滤推荐研究[J].电子学报,2009,37(1):84 – 89.
ZHANG Feng, SUN Xue-dong, CHANG Hui-you, ZHAO Gansen. Research on privacy-preserving two-party collaborative filtering recommendation[J]. Acta Electronica Sinica, 2009, 37(1): 84 – 89. (in Chinese)

[5] Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza. A framework for evaluating privacy preserving data mining algorithms[J]. Data Mining and Knowledge Discovery, 2005, 11(2): 121 – 154.

[6] Rakesh Agrawal, Ramakrishnan Srikant. Privacy-preserving data mining[J]. ACM SIGMOD Record, 2000, 29(2): 439 – 450.

[7] Dakshi Agrawal, Charu C Aggarwal. On the design and quantification of privacy preserving data mining algorithms[A]. In Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems[C]. New York: ACM, 2001. 247 – 255.

[8] H Kargupta, S Datta, Q Wang, K Sivakumar. On the privacy preserving properties of random data perturbation techniques [A]. In Proceedings of the Third IEEE International Conference on Data Mining[C]. Washington, DC: IEEE Computer Society, 2003. 99 – 106.

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

[10] H Dutta, H Kargupta, S Datta, K Sivakumar. Analysis of privacy preserving random perturbation techniques: further explorations[A]. In Proceedings of the 2003 ACM Workshop on Privacy in the Electronic Society[C]. New York: ACM, 2003. 31 – 38.

[11] Songtao Guo, Xintao Wu. On the use of spectral filtering for privacy preserving data mining[A]. In Proceedings of the 2006 ACM Symposium Applied Computing[C]. New York: ACM, 2006. 622 – 626.

[12] Zhengli Huang, Wenliang Du, Biao Chen. Deriving private information from randomized data[A]. In Proceedings of the 2005 ACM SIGMOD[C]. New York: ACM, 2005. 37 – 48.

[13] Li Liu, M Kantarcioglu, B Thuraisingham. The applicability of the perturbation model-based privacy preserving data mining for real-world data[A]. In Proceedings of ICDMW'06[C]. Washington, DC: IEEE Computer Society, 2006. 507 – 512.

[14] Li Liu, M Kantarcioglu, B Thuraisingham. The applicability of the perturbation based privacy preserving data mining for real-world data[J]. Data & Knowledge Engineering, 2008, 65(1): 5 – 21.

[15] Li Liu, M Kantarcioglu, B Thuraisingham. Privacy preserving decision tree mining from perturbed data[A]. In Proceedings of the 42nd Hawaii International Conference on System Sciences[C]. Washington, DC: IEEE Computer Society, 2009. 1 – 10.

[16] Wenliang Du, Zhijun Zhan. Using randomized response techniques for privacy-preserving data mining[A]. In Proceedings of the Ninth ACM SIGKDD[C]. New York: ACM, 2003. 505 – 510.

[17] 葛伟平,汪卫,周皓峰,施伯乐. 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.
Ge Weiping, Wang Wei, Zhou Haofeng, Shi Baile. Privacy preserving classification mining[J]. Journal of Computer Research and Development, 2006, 43(1): 39 – 45. (in Chinese)

[18] Piotr Andruszkiewicz. Privacy preserving classification for continuous and nominal attributes[A]. In Proceedings of the 16th International Conference Intelligent Information Systems [C]. Zakopane, Poland: Polish Academy of Sciences, 2008. 451 – 460.

[19] Piotr Andruszkiewicz. Probability distribution reconstruction for nominal attributes in privacy preserving classification[A]. In Proceedings of the 2008 International Conference on Convergence and Hybrid Information Technology[C]. Washington, DC: IEEE Computer Society, 2008. 494 – 500.

[20] 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.
HAN Jian-min, CEN Ting-ting, YU Hui-qun. Research in microaggregation algorithms for k-anonymization[J]. Acta Electronica Sinica, 2008, 36(10): 2021 – 2029. (in Chinese)

[21] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, Muthuramakrishnan Venkitasubramaniam. L-diversity: privacy beyond k-anonymity[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 1 – 52.

[22] Benjamin C M Fung, Ke Wang, Philip S Yu. Anonymizing classification data for privacy preservation[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(5): 711 – 725.

[23] S Kisilevich, L Rokach, Y Elovici, B Shapira. A multi-dimensional suppression for k-anonymity[OL]. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=484034&isnumbe=4358933

[24] Nan Zhang, Shengquan Wang, Wei Zhao. A new scheme on privacy-preserving data classification[A]. In Proceedings of the Eleventh ACM SIGKDD[C]. New York: ACM, 2005. 374 – 383.

[25] Shaofeng Bu, Laks V S Lakshmanan, Raymond T Ng, Ganesh Ramesh. Preservation of patterns and input-output privacy [A]. In Proceedings of 2007 IEEE 23rd International Conference on Data Engineering[C]. Washington, DC: IEEE Computer Society, 2007. 696 – 705.

[26] Yehuda Lindell, Benny Pinkas. Privacy preserving data mining [J]. Journal of Cryptology, 2002, 15(3): 177 – 206.

[27] M Naor, B Pinkas. Efficient oblivious transfer protocols[A]. In Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms[C]. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2001. 448 – 457.

[28] M Naor, B Pinkas. Oblivious transfer and polynomial evaluation[A]. In Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing[C]. New York: ACM, 1999. 245 – 254.

[29] A C Yao. How to generate and exchange secrets[A]. In Proceedings of 27th FOCS[C]. Washington, DC: IEEE Computer Society, 1986. 162 – 167.

[30] R Fagin, M Naor, P Winkler. Comparing information without leaking it[J]. Communications of the ACM, 1996, 39(5): 77 – 85.

[31] Benny Pinkas. Cryptographic techniques for privacy-preserving data mining [J]. ACM SIGKDD Explorations Newsletter, 2002, 4(2): 12 – 19.

[32] Ming-Jun Xiao, Liu-Sheng Huang, Hong Shen, Yong-Long Luo. Privacy preserving id3 algorithm over horizontally partitioned data[A]. In Proceedings of PDCAT'05[C]. Washington, DC: IEEE Computer Society, 2005. 239 – 243.

[33] D Naccache, J Stern. A new cryptosystem based on higher residues[A]. In Proceedings of the 5th ACM Conference on Computer and Communications Security [C]. New York: ACM, 1998. 59 – 66.

[34] Ming-Jun Xiao, Kai Han, Liu-Sheng Huang, Jing-Yuan Li. Privacy preserving c4.5 algorithm over horizontally partitioned data[A]. In Proceedings of GCC'06[C]. Washington, DC: IEEE Computer Society, 2006. 78 – 85.

[35] F Emekci, O D Sahin, D Agrawal, A El Abbadi. Privacy preserving decision tree learning over multiple parties[J]. Data & Knowledge Engineering, 2007, 63(2): 348 – 361.

[36] A Shamir. How to share a secret[J]. Communications of the ACM, 1979, 22(11): 612 – 613.

[37] Saeed Samet, Ali Miri. Privacy preserving id3 using gini index over horizontally partitioned data[A]. In Proceedings of the 2008 IEEE/ACS International Conference on Computer Systems and Applications[C]. Washington, DC: IEEE Computer Society, 2008. 645 – 651.

[38] Wenliang Du, Zhijun Zhan. Building decision tree classifier on private data [A]. In Proceedings of the IEEE International Conference on Privacy, Security and Data Mining[C]. Darlinghurst, Australia: Australian Computer Society, 2002. 1 – 8.

[39] Justin Zhan. Using homomorphic encryption for privacy-preserving collaborative decision tree classification[A]. In Pro-

ceedings of the CIDM'07[C]. Washington, DC: IEEE Computer Society, 2007. 637 – 645.

[40] Yanguang Shen, Hui Shao, Li Yang. Privacy preserving c4.5 algorithm over vertically distributed datasets[A]. In Proceedings of the 2009 International Conference on Networks Security, Wireless Communications and Trusted Computing[C]. Washington, DC: IEEE Computer Society, 2009. 446 – 448.

[41] Pragnesh Jay Modi, Peter Woo Tae Kim. Classification of examples by multiple agents with private features[A]. In Proceedings of IAT'05[C]. Washington, DC: IEEE Computer Society, 2005. 223 – 229.

[42] Geetha Jagannathan, Rebecca N Wrigh. Privacy-preserving imputation of missing data[J]. Data & Knowledge Engineering, 2008, 65(1): 40 – 56.

[43] Murat Kantarcioglu, Jiashun Jin, Chris Clifton. When do data mining results violate privacy[A]. In Proceedings of the Tenth ACM SIGKDD[C]. New York: ACM, 2004. 223 – 229.

[44] Arik Friedman, Ran Wolff, Assaf Schuster. Providing k-anonymity in data mining[J]. The VLDB Journal, 2008, 17(4): 789 – 804.

作者简介:

李 光 男,1982 年 2 月生于陕西富平.分别于 2004 年和 2006 年在哈尔滨工业大学计算机科学与技术学院获工学学士和工学硕士学位.现为哈尔滨工业大学博士研究生,从事隐私保持数据挖掘、生物信息学等方面的有关研究.
E-mail: hit6006@126.com

王亚东 男,1964 年 6 月生于黑龙江.教授,博士生导师,哈尔滨工业大学计算机科学与技术学院副院长,黑龙江省生物医学信息技术与系统工程研究中心主任;国家"863"计划生物信息技术主题专家,中国人工智能学会理事.主要从事分布式人工智能、机器学习、知识工程、生物信息学等学术方向研究.主持完成国家自然科学基金、国家 863 计划项目、国际合作项目等 20 余项,获国家科技进步二等奖 1 项、在国内外重要期刊发表论文 50 余篇.
E-mail: ydwang@hit.edu.cn

苏小红 女,1966 年 10 月出生,辽宁海城人,博士,教授,博士生导师,中国计算机学会高级会员.主要研究方向:图像处理与目标识别,信息融合,软件缺陷检测,计算机图形学等.获黑龙江省科技进步二等奖 1 项、黑龙江省自然科学二等奖 1 项,在国内重要期刊和国际国内会议上发表论文 50 余篇. E-mail: sxh@hit.edu.cn

(上接第 198 页)

Wang Yong, Dou Qiang. An improved conflict detection algorithm of transactional memory[J]. Computer Engineering and Scieuce, 2008, 30(A1): 190 – 193. (in Chinese)

[5] Bloom B H. Space/time trade-offs in hash coding with allowable errors[J]. Communications of the ACM. 1970, 13(7): 422 – 426.

[6] Sanchez D. Design and implementation of signatures for transactional memory systems[J]. Technical Report CS-TR-2007 – 1611. 2007.

[7] Harris T, Fraser K. Language support for lightweight transactions[J]. ACM SIGPLAN Notices, 2003, 38(11): 388 – 402.

[8] 窦强,王勇.事务存储系统中 PGHB 冲突检测算法研究[J].计算机工程与科学,2010,32(1):120 – 122.
Dou Qiang, Wang Yong. The research on PGHB conflict detection algorithm in transactional memory systems[J]. Computer Engineering and Science, 2010, 32(1): 120 – 122. (in Chinese)

[9] D Tarjan, S Thoziyoor N P J. CACTI 4.0[R]. HP Labs2006.