隐私保护的分布式决策树分类算法的研究 *

申艳光, 邵 慧, 张永强
(河北工程大学 信息与电气工程学院, 河北 邯郸 056038)

摘 要: 针对分布式决策树构造过程中的隐私保护问题,引入安全多方计算方法设计了可以保护隐私的分布式 C4.5 决策树分类算法。该算法适用于数据集垂直分布和水平分布两种情况,同时提出了一种新的隐私保护程度的度量方法。实验结果证明设计的隐私保护分布式决策树分类算法不仅很好地保护了原始数据不泄露,同时保持了较高的分类精度。

关键词: 分布式数据挖掘; 隐私保护; 安全多方计算; C4.5 决策树算法; 垂直分布; 水平分布
中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2010)08-3070-03
doi:10.3969/j.issn.1001-3695.2010.08.069

Research on privacy preserving distributed decision-tree classification algorithm

SHEN Yan-guang, SHAO Hui, ZHANG Yong-qiang
(*School of Information Science & Electrical Engineering, Hebei University of Engineering, Handan Hebei* 056038, *China*)

Abstract: To solve the problem of privacy preserving of the distributed decision-tree building process, this paper introduced the secure multi-party computation method and designed an algorithm of privacy preserving distributed C4.5 algorithm, which was applicable to the vertically and horizontally distributed dataset, and also proposed a new computation method of the degree of privacy protection. Experimental results demonstrate that the privacy preserving algorithm can well protect the original data from revealing, and keep high classification correct accuracy.

Key words: distributed data mining; privacy preserving; secure multi-party computation; C4.5 decision-tree algorithm; vertically distributed; horizontally distributed

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的数据集中提取隐含的、事先未知的、但又潜在有用的信息和知识的过程。与此同时,数据挖掘也面临着许多问题的挑战。其中,数据挖掘的隐私和信息安全问题尤其得到关注。2000 年 Agrawal 等人[1]和 Lindell 等人[2]首次提出了隐私保护的数据挖掘概念,引起了学术界和产业界的广泛关注,并吸引了一批学者致力于该领域的研究。从此,隐私保护的数据挖掘成为了研究重点。近年来,随着 WWW 的普及应用,Internet 成为最大的分布式数据源,在数据挖掘过程中,经常需要来自不同站点数据库中的数据,如不同银行希望从共享数据中得到欺诈行为的模式;政府机构需要与航天部门合作,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,这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精确数据的前提下,获取准确的挖掘规则结果。因此,隐私保护的分布式数据挖掘成为了一个全新的研究方向。

本文主要考虑数据挖掘者为两方的情况,数据集水平分布时,应用安全多方和协议和安全 $x\ln(x)$ 协议构造具有隐私保护效果的决策树分类器;数据集垂直分布时,应用标量积协议和安全 $x\ln(x)$ 协议构造具有隐私保护效果的决策树分类器。

1 问题描述

本文假设数据集 S 垂直(水平)分布于 A 和 B 两方,分别为 S_a 和 S_b,并且数据集 $S=S_a\cup S_b$,则:

a)如果数据集 S 垂直划分:则 A 和 B 两方有相同的记录数,用 N 代表总的记录数;每一条记录被分为两部分,S_a 包含一部分属性,S_b 则包含另一部分属性,用 n 表示所有属性个数;A 和 B 共享属性集及类标号属性,用 m 表示类的个数。

b)如果数据集 S 水平划分:则 A 和 B 两方拥有不同的记录数,但每条记录是完整的,A 和 B 共享属性集以及类标号属性,用 m 表示类的个数。

2 隐私保护的分布式决策树算法研究

2.1 隐私保护的 PPC4.5 决策树算法

C4.5 算法由 Quinlan J R 在 1993 年提出,是 ID3 算法的改进。C4.5 算法在 ID3 的基础上增加了对连续型属性和属性值空缺情况的处理,对树剪枝也有了较成熟的方法。与 ID3 不同,C4.5 采用基于信息增益比例(GainRatio)的方法选择测试属性。本文以 C4.5 算法[3]为基础,设计了保护隐私的 PPC4.5 $(S,\hat{A}\_attribute)$ 算法,文中假设 $\hat{A}$ 表示当前节点,$\hat{A}$_attribute 表示当前测试属性集,$S=S_a\cup S_b$ 表示当前数据库。

Begin

(1)创建根节点 T;

垂直划分:A 计算 S_a 中每一属性的信息增益比例,B 计算 S_b 中每

收稿日期: 2010-01-01; **修回日期**: 2010-02-29 **基金项目**: 河北省教育厅科学研究计划项目(2009421);河北省自然科学基金资助项目(F2008000752)

作者简介: 申艳光(1970-),女,河北邯郸人,教授,硕士,主要研究方向为数据挖掘和信息安全(shenyanguang@yahoo.com);邵慧(1983-),女,山东泰安人,硕士,主要研究方向为数据挖掘;张永强(1966-),男,教授,主要研究方向为软件工程.

一属性的信息增益比例,用信息增益比例最大的节点做根节点;

水平划分:A、B 需要合作才能计算信息增益比例。

(2)if S 都属于同一类 C,则返回 T 为叶节点,标记为类 C;

垂直划分:由于 A、B 双方共同分享类标号,要确定 S 是否属于同一类 C,只需考察 $S_a(S_b)$ 记录是否属于同一类 $C_a(C_b)$,若都属于同一类,则返回叶节点,用 $C_a(C_b)$ 标记。

水平划分:分别考察 S_a 中的记录是否属于同一类 C_a,S_b 中的记录是否属于同一类 C_b。若都属于同一类,再考察 C_a 是否等于 C_b。运用 Yao 的安全两方比较协议[4]可以确定 C_a 是否等于 C_b。若 $C_a = C_b$,则返回叶节点,用 C_a 或 C_b 标记。

(3)if $\hat{A}$_attribute 为空 OR S 中所剩的样本数少于某给定值,则返回 T 为叶节点,用 S 中最频繁的类标记;

垂直划分:由于 A、B 两方共同分享所有属性的名称,则 $\hat{A}$_attribute 是否为空它们都可知。当 $\hat{A}$_attribute 为空时,需要找出 S 中最频繁的类 C_i,又因为 A、B 共享类标号,所以只需扫描 $S_a(S_b)$ 一个数据库统计其最频繁的类即可。

水平划分:A 统计 S_a 中每个类的记录数 $|S_a(C_i)|$,B 统计 S_b 中每个类的记录数 $|S_b(C_i)|$。运用 Yao 的安全两方比较协议,A 输入 $(|S_a(C_1)|,\cdots,|S_a(C_m)|)$,$B$ 输入 $(|S_b(C_1)|,\cdots,|S_b(C_m)|)$,计算输出为 $i = \max\{(|S_a(C_i)| + |S_b(C_i)|)\}$。

(4)创建队列 Q,根节点 T 入队列 Q;

(5)while 队列不为空

(6){

(7)从队列中取出第一个节点 $\hat{A}$;

(8)for each $\hat{A}$_attribute 中的属性计算信息增益比例 GainRatio;

(9)T 的最佳分裂属性 test_attribute = $\hat{A}$_attribute 中具有最高信息增益比例的属性;

(10)if 分裂属性为连续型 then 找到该属性的分割阈值;

(11)用信息增益比例最大的那个属性把节点 $\hat{A}$ 分为 k 个子集 $\hat{A_1}$,$\hat{A_2}$,$\cdots$,$\hat{A_k}$;

(12)for each 由节点 $\hat{A}$ 长出的新叶节点 $\hat{A}${

if 该叶节点对应的样本子集只有惟一的一种决策类别,则将该叶节点标记为该类别;

else 将该节点入队列 Q,在该叶节点上执行 PPC4.5(S',$\hat{A}$_ attribute)对它继续分裂;

(13)　　}

(14)}

(15)计算每个节点的分类错误,进行树剪枝。

end

2.2　节点最佳分裂属性的确定方法

为了确定每一节点的最佳分裂属性,需要计算每一节点所有属性的信息增益比例,将值最大的属性作为节点的分裂属性,数据集垂直划分和水平划分的最佳分裂属性的确定方法不同,分开来讨论。

2.2.1　数据集垂直划分的最佳分裂属性的确定方法

设 S 代表当前节点的数据集,R 表示计算当前节点属性的信息增益比例所需要的数据集,会出现两种情况:a)R 中的属性和当前测试属性集 $\hat{A}$_attribute 属于一个数据集,则其中任意一方均可以单独地计算属性的信息增益比例;b)R 中的属性和当前测试属性集 $\hat{A}$_attribute 不属于同一个数据集,此时一方需要联合另一方的数据才能计算信息增益比例。下面主要讨论情况 b)的信息增益比例计算方法。

把 R 分为两个子集 R_a 和 R_b,设 R_a 是从 S_a 获取的数据集,R_b 是从 S_b 获取的数据集。E_a 表示仅与 S_a 中有关的属性组成的逻辑表达式,E_b 表示仅与 S_b 中有关的属性组成的逻辑表达式。A 扫描 S_a 生成一个 N 维向量 V_a,如果第 i 条记录满足 E_a,则 $V_a(i) = 1$,否则 $V_a(i) = 0$。A 可以自己计算出向量 V_a 的值,同理,B 也可以自己计算出向量 V_b 的值。$V(i) = V_a(i) \wedge V_b(i)$ 表示同时满足条件 E_a 和 E_b 的记录。

设 V_i 是一个 N 维向量,如果第 t 条记录属于类 i 则 $V_i(t) = 1$,否则 $V_i(t) = 0$。A 和 B 都可以独自计算出向量 V_i 的值。标量积 $V_aV_b = \sum_{i=1}^{N} V_a(i) * V_b(i)$ 表示同时满足条件 E_a 和 E_b 的非零项的记录个数;$\hat{P_i} = V_a \cdot (V_b \wedge V_i) = (V_a \wedge V_i) \cdot V_b$ 表示数据集 S 中属于类 i 的记录数。

1)计算 $E(S,\hat{A}\text{_attribute})$

$$E(S,\hat{A}\text{_attribute}) = -\sum_{j=1}^{v}\frac{s_{1j} + s_{2j} + \cdots + s_{mj}}{s}$$

$$I(s_{1j},s_{2j},\cdots,s_{mj}) = -\sum_{j=1}^{v} P_j I(s_{1j},s_{2j},\cdots,s_{mj})$$

其中:①$P_j = \frac{V_a \cdot V_b}{|S|}$,$|S| = \sum_{i=1}^{m}\hat{P_i}$;

②$I(S_{1j},S_{2j},\cdots,S_{mj}) = -\sum_{i=1}^{m} P_{ij}\log_2(p_{ij}) = -\sum_{i=1}^{m}\frac{\hat{P_i}}{V_a \cdot V_b}\log_2(\frac{\hat{P_i}}{V_a \cdot V_b})$。

2)计算 $\text{Gain}(S,\hat{A}\text{_attribute})$

$$\text{Gain}(S,\hat{A}\text{_attribute}) = I(S_1,S_2,\cdots,S_m) - E(S,\hat{A}\text{_attribute})$$

其中 $I(s_1,s_2,\cdots,s_m)$ 不用联合,一方可以单独完成这部分的计算。将 1)计算出的熵的值代入此式即可求出信息增益的值。

3)计算 $\text{GainRatio}(S,\hat{A}\text{_attribute})$

$$\text{GainRatio}(S,\hat{A}\text{_attribute}) = \frac{\text{Gain}(S,\hat{A}\text{_attribute})}{\text{Split}I(S,\hat{A}\text{_attribute})}$$

其中 $\text{Split}I(S,\hat{A}\text{_attribute}) = -\sum_{j=1}^{v} P_j\log_2(p_j)$ 且 $P_j = \frac{V_a \cdot V_b}{|S|}$,$|S| = \sum_{i=1}^{m}\hat{P_i}$。

将上式计算出的信息增益的值代入此式即可求出信息增益比例的值。重复上面的计算过程直到求出此节点所有属性的信息增益比例值,选取值最大的属性分裂节点,循环直到决策树构造完成。

2.2.2　数据集水平划分的最佳分裂属性的确定方法

假设 $|S_a(C_i)|$ 表示当前节点取自 S_a 的属于类 i 的记录数,$|S_b(C_i)|$ 表示当前节点取自 S_b 的属于类 i 的记录数;$|S_a(\text{attr}_j,C_i)|$ 表示当前节点取自 S_a 的属性值等于 j 的分支中属于类 i 的记录数,$|S_b(\text{attr}_j,C_i)|$ 表示当前节点取自 S_b 的属性值等于 j 的分支中属于类 i 的记录数。

1)计算期望信息 $I(s_1,s_2,\cdots,s_m)$

$$I(s_1,s_2,\cdots,s_m) = -\sum_{i=1}^{m} p_i\log^2(p_i)$$

其中 $p_i = \frac{\text{Sum}(|S_a(C_i)|,|S_b(C_i)|)}{\text{Sum}(\sum_{i=1}^{m}|S_a(C_i)|,\sum_{i=1}^{m}|S_b(C_i)|)}$。

2)计算熵 $E(S,\hat{A}\text{_attribute})$

$$E(S,\hat{A}\text{_attribute}) = -\sum_{j=1}^{v}\frac{s_{1j} + s_{2j} + \cdots + s_{mj}}{s}$$

$$I(s_{1j},s_{2j},\cdots,s_{mj}) = -\sum_{j=1}^{v} P_j I(s_{1j},s_{2j},\cdots,s_{mj})$$

其中:①$P_j = \frac{\text{Sum}(\sum_{i=1}^{m}|S_a(\text{attr}_j,C_i)|,\sum_{i=1}^{m}|S_b(\text{attr}_j,C_i)|)}{\text{Sum}(\sum_{i=1}^{m}|S_a(C_i)|,\sum_{i=1}^{m}|S_b(C_i)|)}$;

② $I(s_{1j}, s_{2j}, \cdots, s_{mj}) = -\sum_{i=1}^{m} p_{ij}\log_2(p_{ij})$　$p_{ij} = \dfrac{\mathrm{Sum}(|S_a(\mathrm{attr}, C_i)|, |S_b(\mathrm{attr}_j, C_i)|)}{\mathrm{Sum}(\sum_{i=1}^{m}|S_a(\mathrm{attr}_j, C_i)|, \sum_{i=1}^{m}|S_b(\mathrm{attr}_j, C_i)|)}$;

3)计算信息增益 Gain(S,$\hat{A}$_ attribute)

$$\mathrm{Gain}(S,\hat{A}\_\mathrm{attribute}) = I(s_1, s_2, \cdots, s_m) = -E(S,\hat{A}\_\mathrm{attribute})$$

将上面两个公式计算得出的值代入此公式即可求得信息增益的值

4)计算信息增益比例 GainRatio(S,$\hat{A}$_attribute)

$$\mathrm{GainRatio}(A) = \frac{\mathrm{Gain}(S,\hat{A}\_\mathrm{attribute})}{\mathrm{Split}I(S,\hat{A}\_\mathrm{attribute})}$$

其中:$\mathrm{Split}I(S,\hat{A}\_\mathrm{attribute}) = -\sum_{j=1}^{v} p_j\log_2(p_j)$

且 $P_j = \dfrac{\mathrm{Sum}(\sum_{i=1}^{m}|S_a(\mathrm{attr}_j, C_i)|, \sum_{i=1}^{m}|S_b(\mathrm{attr}_j, C_i)|)}{\mathrm{Sum}(\sum_{i=1}^{m}|S_a(C_i)|, \sum_{i=1}^{m}|S_b(C_i)|)}$

数据集垂直分布时,应用标量积协议[5]即可求出的 $V_a \cdot V_b$ 值,并且 $V_a \cdot V_b$ 的结果被分为两部分 x_a 和 x_b,即 $V_a \cdot V_b = x_a + x_b$,$A$ 保管 x_a,B 保管 x_b,这样可保证 A 无法知道 V_b 的内容,B 无法知道 V_a 的内容,从而保护了各自的隐私。应用 $x\ln(x)$ 协议[6]可得 $\ln(x_a + x_b) = u_a + u_b$,$u_a$、$u_b$ 分别由 A、B 保管,所以 $(x_a + x_b)\ln(x_a + x_b) = (x_a + x_b)(u_a + u_b) = x_a u_a + x_b u_b + x_b u_a + x_{aub}$。其中 $x_b u_a$ 的计算结果分为两部分 y_a 和 y_b,分别由 A、B 各自保管,同理,$x_a u_b$ 的计算结果也分为两部分 w_a 和 w_b,分别由 A、B 各自保管。A 可以计算 $Z_a = x_a u_a + y_a + w_a$,$B$ 可以计算 $Z_b = x_b u_b + y_b + w_b$,则 $(x_a + x_b)\ln(x_a + x_b) = Z_a + Z_b$,结果仍是分为两部分 Z_a 和 Z_b,分别由 A、B 各自保管,从而保护了各自的隐私。同理,数据集水平分布时,先应用安全和协议[7],然后再应用 $x\ln(x)$ 协议即可实现双方的隐私保护。

3　隐私保护程度的度量方法

隐私保护程度是指防止原始数据或其隐含的知识被推导出的程度。现有的隐私度量一般用“披露风险”(disclosure risk)[8]来描述,披露风险表示攻击者根据所发布的数据和其他背景知识(background knowledge),可能披露隐私的概率。披露风险只给出了隐私度量的定性说明,没有给出一个定量计算方法。因此,本文提出了一种决策树分类挖掘隐私保护程度的度量方法:

隐私保护程度 = PPC4.5 算法获得的分类规则数 * $W_{\mathrm{ppc4.5}}$/C4.5 获得的分类规则数 * $W_{\mathrm{c4.5}}$

其中:将决策树的叶子节点数定义为获得的分类规则数;分类正确率指正确分类的记录数占总的记录数的比率,那么权重 $W_{\mathrm{ppc4.5}}$ 指 PPC4.5 算法的分类正确率,权重 $W_{\mathrm{c4.5}}$ 指 C4.5 算法的分类正确率。

4　实验结果分析

本文在 Weka 平台上封装实现了隐私保护的 PPC4.5 算法,如图 1 所示。为了验证算法的可行性,这里选取了 Weka 自带的 2 个数据集(soybean 和 iris)和 UCI 机器学习数据库中[9]的 3 个数据集(adult、car 和 credit)进行了实验验证。实验结果如图 2 ~ 4 所示。

图 2 是 PPC4.5 和 C4.5 算法的分类正确率的对比。

图1　PPC4.5在Weka中的实现

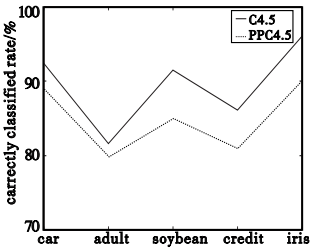

图2　分类正确率对比

图 3 是对于 5 个不同的数据集 PPC4.5 和 C4.5 算法获得的规则数的比较。

图 4 显示了 PPC4.5 算法对不同数据集的隐私保护程度,根据隐私保护程度的计算公式和图 2、3 的数据可以得到图 4。

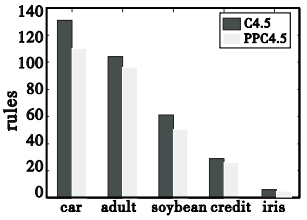

图3　规则数比较

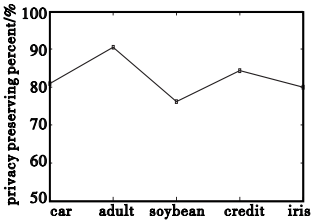

图4　隐私保护程度

本实验采用 10 次交叉验证法验证生成决策树的分类精度。将本文提出的隐私保护分布式决策树分类算法的分类准确率与传统算法对比可以看出,该算法的分类准确程度较传统算法有所下降,但是下降在 6% 之内,仍然可以接受,隐私保护程度在最差的情况下也可以达到 76% 以上。由实验结果可以看出,本文提出的算法较好地保护了原始数据不泄露,同时保持了较高的分类精度。

5　结束语

现有的一些数据挖掘算法并没有考虑到数据隐私问题,本文将密码学中的安全多方计算方法与数据挖掘方法相结合设计实现了具有隐私保护功能的分布式 C4.5 决策树分类算法,详细介绍了最佳分裂节点的确定方法和信息增益比例的计算方法,同时提出了一种决策树挖掘隐私保护程度的度量方法。本文只考虑了对隐私数据的隐藏,但未涉及规则的隐藏,并且未考虑隐私保护的个性化需求,对规则的隐私保护以及个性化隐私保护将是未来进一步研究的内容。

参考文献:

[1] AGRAWAL R, SRIKANT R. Privacy-preserving data mining[C]// Proc of ACM SIGMOD on Management of Data. 2000:439-450.

[2] LINDELL Y, PINKAS B. Privacy preserving data mining [J]. Journal of Cryptology, 2002(15):177-206.

[3] QUINLAN R J. C4.5: Programs for machine learning[M]. San Mateo, CA: Morgan Kaufmann Publisher, 1993.

[4] YAO A C. Protocols for secure computations[C]//Proc of the 23rd Annual IEEE Symposium on Foundations of Computer Science. 1982.

[5] DU Wen-liang, ZHAN Zhi-jun. Building decision tree classifier on private data[C]//Proc of IEEE International Conference on Data Mining Workshop on Privacy, Security and Data Mining. Maebashi City: Australian Computer Society, Inc, 2002.

[6] PINKAS B, LABS H P. Cryptographic techniques for privacy preserving data mining [J]. SIGKDD Explorations, 2002, 4(2):12-19.

[7] 杨林,张霞萍,白治国,等. 基于SMC协议的分布式聚类分析隐私保护的研究[J]. 计算机工程与设计,2008,29(21):5424-5426.

[8] 周水庚,李丰,陶宇飞,等. 面向数据库应用的隐私保护研究综述[J]. 计算机学报, 2009, 32(5):847-861.

[9] UCI. Machine learning repository[EB/OL]. http://archive.ics.uci.edu/ml/datasets.html.