

文章编号:1008-1542(2009)02-0087-05

## 基于决策树的知识表示模型及其应用

李萍<sup>1</sup>,李法朝<sup>1,2</sup>

(1. 河北科技大学理学院,河北石家庄 050018;2. 河北科技大学经济管理学院,河北石家庄 050018)

**摘要:**针对数据挖掘过程中的数据库精炼问题,在分析现行属性约简方法的特点和不足的基础上,结合决策树算法操作简单、分类速度快的特点,通过知识的规则化描述以及规则族之间的相似性比较,建立了一种基于决策树的属性约简方法(简记为BD-RED),讨论了规则族之间的相似性度量的可释化构建问题,给出了BD-RED的具体实施策略,并结合实例分析了BD-RED的性能。结果表明,BD-RED具有良好的结构特征和较强的可操作性,可以有效实现不同决策理念下的属性约简,适合不同类型的大规模数据库的属性约简。

**关键词:**决策树;数据挖掘;属性约简;规则;相似度量

中图分类号:O236

文献标识码:A

## Knowledge model based on decision tree and its application

LI Ping<sup>1</sup>, LI Fa-chao<sup>1,2</sup>

(1. College of Sciences, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China; 2. College of Economics and Management, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China)

**Abstract:** For the refinement of the database in data mining, by analyzing the characteristics and shortcomings of the current attribute reduction methods, and combining with it the features of simple operation and rapid classification of decision tree, the authors established a kind of attribute reduction method (BD-RED) based on decision tree using rule description of the knowledge and similarity measures between rules families. Further, we discussed the explanatory construction of similarity measure between rules families, gave the specific implementation strategy of BD-RED, and analyzed the performance through examples. The results show that BD-RED has a good structure and strong operability, and is an effective way to achieve attribute reduction under different consciousness, so it can be suitable to the large-scale attribute reduction.

**Key words:** decision tree; data mining; attribute reduction; rules; similarity

随着信息技术的发展,采用信息化手段来管理、监测和记录生产活动是社会发展的必然趋势,各行各业在不同的层面上积累了大量的数据资料。这些数据不仅可以理解为过去的活动记录,而且可以认为是某种试验结果的积累,当其积累到一定程度时,必然会反映出规律性的东西,因此,堆积如山的数据无异于一个巨大的宝库。如何发现隐藏在数据中的知识(简称为数据挖掘),并将之运用到社会活动中,是当今人工智能领域的一个重要研究内容。

分类是数据挖掘的一项核心任务,而分类的依据常常是人们所关心问题的某些方面的特征(属性)。由于数据库中的数据往往与给定的属性集中的某些属性的状态(即取值)无关或关联不大,直接采用给定的属

收稿日期:2008-12-26;修回日期:2009-03-09;责任编辑:陈书欣

基金项目:国家自然科学基金资助项目(70671034);河北省自然科学基金资助项目(F2006000346);河北省科技攻关项目(05547004D-2)

作者简介:李萍(1979-),女,河北石家庄人,硕士研究生,主要从事数据挖掘、机器学习等方面的研究。

性集来挖掘知识将增大数据挖掘的难度,特别是对于巨型数据库而言,可能会导致相关数据挖掘算法的失效,因而,如何精炼数据挖掘的属性集(称之为属性约简),是数据挖掘的一个关键环节。针对该问题,学者们进行了有益的探讨,取得了许多成果。例如:文献[1]—文献[3]讨论了基于正区域的属性约简方法;文献[4]—文献[6]讨论了基于信息熵的属性约简方法;文献[7]—文献[9]讨论了基于辨识矩阵的属性约简方法。虽然这些算法均具有良好的理论基础,但它们的空间复杂度和时间复杂度均较高,不能有效处理大型数据库的属性约简问题。

基于上面的分析,笔者结合决策树算法操作简单、分类速度快的特点,通过将知识库抽象为规则族及规则族之间的相似性比较,建立了一种基于决策树的属性约简方法(简记为BD-RED)。具体工作如下:1)建立了规则的形式化描述模式;2)从结构化的角度讨论了规则族之间的相似性度量的构建问题;3)给出了BD-RED的具体实施原则;4)结合具体实例分析了BD-RED的特征和性能。

## 1 决策树算法概述

决策树算法是一种贪婪型的递归方法,是当今机器学习领域的一个重要工具。它起源于HUNT等人提出的概念学习系统(concept learning system,简记为CLS),目前的决策树算法大都是对CLS算法的改进或由CLS衍生而来。其代表性的工作有:1986年,QUINLAN以信息增益作为启发策略,提出了著名的ID3<sup>[10]</sup>算法以及在ID3算法的基础上,演化出的ID4和ID5等;1993年,QUINLAN提出了适应于连续属性的决策树归纳包C4.5<sup>[11]</sup>;著名的决策树方法还有CART<sup>[12]</sup>,SLIQ<sup>[13]</sup>,SPRINT<sup>[14]</sup>和PUBLIC<sup>[15]</sup>等。在这些算法中,ID3是最为常用的,其执行过程如下。

步骤1 选取所有条件属性中信息增益最大的属性作为根节点。

步骤2 1)如果节点上的每个分支中所包含的示例具有相同的决策属性值,则产生叶子节点;2)如果节点的每个分支所包含的示例具有基本相同的决策属性值,即分类误差在允许的范围之内,则产生叶子节点;3)如果在节点及节点的上层,所有条件属性都被使用过,则产生叶子节点。

步骤3 否则,选取该节点及其上层未用过的具有最大信息增益的属性作为扩展属性,并按照扩展属性的取值对当前节点进行分支(划分);重复步骤2和步骤3,直至不能再进行分支为止。

步骤4 将产生的树转换成产生式规则。

注1 属性A关于示例集T的信息增益为  $\text{Gain}(A, T) = \text{infor}(T) - \sum_{i=1}^s \frac{|T_i|}{|T|} \times \text{infor}(T_i)$ 。其中,  $|T|$  表示集合T中的元素个数,  $T_1, T_2, \dots, T_s$  是按属性A的不同的取值所形成的对示例集T的划分;  $\text{infor}(T)$  表示T的信息熵,即  $\text{infor}(T) = - \sum_{j=1}^m \frac{|T_j^d|}{|T|} \times \log_2 \frac{|T_j^d|}{|T|}$ ,其中,  $T_1^d, T_2^d, \dots, T_m^d$  是按决策属性的不同取值所形成的对示例集T的划分。

## 2 规则型知识的结构化表示模型

笔者用E表示示例样本的全体,  $C = \{C_1, C_2, \dots, C_s\}$  表示条件属性, D表示决策属性,并称(E, C, D)为基本知识库。其中,  $C_i$  的取值范围为  $V(C_i) = \{c_{i1}, c_{i2}, \dots, c_{im_i}\}$ ,  $i = 1, 2, \dots, s$ ; D的取值范围为  $V(D) = \{d_1, d_2, \dots, d_n\}$ 。规则是一种最为常用的表达知识的方式,其基本形式如下:若  $C_1 = c_1, C_2 = c_2, \dots, C_s = c_s$ , 则  $D = d$ 。由于实际中的数据信息常常存在不同形式的不完备性(如噪音、缺值等),因而从中所归纳出来的知识往往具有不同程度的不确定性,若采用  $(0, 1)$  来刻画其相应的可靠(真实)程度,那么,规则R可通俗地表示为

$$R = (c_1, c_2, \dots, c_s, d, \alpha) \quad (1)$$

式中:  $c_s$  表示条件属性  $C_s$  的取值;  $d$  表示决策属性 D 的取值;  $\alpha$  表示知识 R 的可靠程度。

注2 实际中某些规则常与部分条件属性的取值无关。例:若  $C_1 = c_1, C_2 = c_2$ , 则  $D = d$ 。为了用式(1)的形式描述该规则,约定  $C_3, C_4, \dots, C_s$  的取值均为  $\alpha$ , 且  $\alpha$  仅是一种代号,表示与取值无关。

注3 规则的可靠度  $\alpha$  是反映规则与示例库中的数据信息的吻合程度的一种数量指标,  $\alpha = 1$  表示完全吻合,  $\alpha < 1$  表示部分吻合,  $\alpha$  越大,表明规则的真实性越强。

注 4 在实际问题中,应结合数据库的特点及对规则知识的用途来适当限定 的取值。

### 3 基于决策树的属性约简方法(BD-RED)

#### 3.1 BD-RED 的结构

属性约简是简化数据库的一个重要措施,其含义是在保持决策能力或结构特征基本不变的前提下,删除不相关或不重要的部分数据信息,其本质是属性子集的选取问题。利用知识的一般描述模式(1)可知,若  $(E, C, D)$  与  $(E, C - \{C_i\}, D)$  具有某种意识下的等同性(即以示例库  $E$  为基础,分别利用属性集  $C$  和  $C - \{C_i\}$  所得到的知识具有某种等同性),那么属性  $C_i$  在该意识下即可认为是多余的,依此类推即可得到属性集  $C$  的某种最简约简  $B$ 。若将知识通俗理解为规则,并采用决策树作为提取规则的方法,那么即可建立一种基于决策树的属性约简方法(简记为 BD-RED),其实施步骤如下(其中,  $(E, C, D)$  表示利用  $(E, C, D)$  所得到的规则全体)。

步骤 1 输入示例集  $E$  和属性集  $A = C - D$ 。

步骤 2 选择决策树的启发策略  $P$ 、规则族之间的相似度量  $M$ 、属性的重要性衡量准则以及属性的约简强度(即属性的多余标准)  $(0, 1]$ 。

步骤 3 利用启发策略  $P$  确定基本规则族  $R = (E, C, D)$ 。

步骤 4 计算  $C$  中各条件属性的重要性,并按重要程度由小到大排序为  $C_1^*, C_2^*, \dots, C_t^*$ 。

步骤 5 从 1 到  $t$  依次执行下面的操作:利用启发策略  $P$  确定规则族  $B = (E, C - \{C_i^*\}, D)$ ,若不存在  $i \in \{1, 2, \dots, t\}$  使得  $M(R, B) < \alpha$ , 则令  $B = A$ , 并转步骤 6; 否则,考虑第 1 个满足  $M(R, B) < \alpha$  的  $C_k^*$ , 并将属性集  $A$  更新为  $A - \{C_k^*\}$ , 转步骤 4。

步骤 6 输出  $B$ 。

显然,步骤 1 到步骤 6 仅仅给出了一个基本操作流程,而步骤 2 中各标准是进一步体现 BD-RED 的约简意识的参数。在决策树的启发策略和属性的重要性衡量准则选取方面,现行文献已进行了许多的相关讨论,下面着重考虑规则族之间的相似度量问题。

#### 3.2 规则族之间的相似度量的构建准则

由式(1)知,2 个规则之间的相似性问题可以转化为 2 个向量的各分量的状态相似性问题。由于式(1)中的  $\alpha$  表示规则与相应的属性取值无关,因而对  $\alpha$  的处理在一定程度上决定了规则之间的相似问题。若用  $x_i$  表示第  $i$  个条件属性的状态差异(约定相同时为 1,不同时为 0),  $y$  表示决策属性的状态差异(约定相同时为 1,不同时为 0),  $z$  表示可靠度的差别,并约定 0 和 1 分别表示规则之间 2 种极端相似状态(0 表示绝对的不相似,1 表示完全的相似),那么规则之间的相似性度量问题即可抽象为一个从  $\{0, 1\}^n \times \{0, 1\} \times \{0, 1\}$  到  $[0, 1]$  的函数  $f(x_1, x_2, \dots, x_n, y, z)$  且满足:

条件 1 (绝对不同性):  $f(x_1, x_2, \dots, x_n, 0, z) = 0$ ;

条件 2 (绝对同性):  $f(1, 1, \dots, 1, 1, 0) = 1$ ;

条件 3 (相对同性):  $f(1, 1, \dots, 1, 1, z)$  关于  $z$  单调不增。

其中,  $f(x_1, x_2, \dots, x_n, 0, z) = 0$  表示当决策属性的状态不同时,无论条件属性取什么值以及规则的可靠度如何,相应的规则是绝对不相似的;  $f(1, 1, \dots, 1, 1, 0) = 1$  表示当属性的取值状态和可靠度相同时,相应的规则是绝对相似的;  $f(1, 1, \dots, 1, 1, z)$  关于  $z$  单调不增表示当删除某些条件属性(即相应的取值为 0)后所得的规则与原规则相似时(即决策属性的取值相同),其相似程度随可靠度的差别的增大而减小。称满足上述条件 1—条件 3 的函数  $f(x_1, x_2, \dots, x_n, y, z)$  为规则的相似函数(简称为相似函数)。不难验证,对  $\forall \alpha \in (0, 1)$ ,

$0 < w_i \leq 1$ , 且  $\sum_{i=1}^n w_i = 1$ , 函数  $f_1(x_1, x_2, \dots, x_n, y, z) = y(1 - z) \sum_{i=1}^n w_i x_i$  是相似函数。

按照以上讨论,对于规则  $R = (c_1, c_2, \dots, c_s, d, \alpha)$  和  $R' = (c_1, c_2, \dots, c_s, d', \alpha')$  以及相似函数  $f(x_1, x_2, \dots, x_n, y, z)$ , 规定:对属性的取值  $a$  和  $b$ , 当  $a = b$  时,  $a \cdot b = 1$ ; 当  $a \neq b$  时,  $a \cdot b = 0$ , 则

$$S(R, R') = f(c_1 \cdot c_1, c_2 \cdot c_2, \dots, c_s \cdot c_s, d \cdot d', \alpha - \alpha') \quad (2)$$

为  $R$  与  $R'$  的一种相似度量,且满足  $S(R, R) = S(R', R')$ ,  $S(R, R) \in [0, 1]$ , 即  $S$  是规则之间一种相似关系。

根据式(2),对规则族  $R^{(1)} = \{R_1^{(1)}, R_2^{(1)}, \dots, R_s^{(1)}\}$  和  $R^{(2)} = \{R_1^{(2)}, R_2^{(2)}, \dots, R_t^{(2)}\}$ , 记

$$S(\mathbf{R}^{(1)}, \mathbf{R}^{(2)}) = (S(R_i^{(1)}, R_j^{(2)}))_{s \times t}, \tag{3}$$

则矩阵  $S(\mathbf{R}^{(1)}, \mathbf{R}^{(2)})$  从整体上描述了  $\mathbf{R}^{(1)}$  与  $\mathbf{R}^{(2)}$  中各规则之间的相似特征;而

$$a_i = \max\{S(R_i^{(1)}, R_1^{(2)}), S(R_i^{(1)}, R_2^{(2)}), \dots, S(R_i^{(1)}, R_t^{(2)})\} \tag{4}$$

描述了  $R_i^{(1)}$  (近似) 属于  $\mathbf{R}^{(2)}$  的程度,  $i = 1, 2, \dots, s$ ;

$$b_j = \max\{S(R_1^{(1)}, R_j^{(2)}), S(R_2^{(1)}, R_j^{(2)}), \dots, S(R_s^{(1)}, R_j^{(2)})\} \tag{5}$$

描述了  $R_j^{(2)}$  (近似) 属于  $\mathbf{R}^{(1)}$  的程度,  $j = 1, 2, \dots, t$ ; 从而

$$M(\mathbf{R}^{(1)} \cong \mathbf{R}^{(2)}) = \frac{1}{2} [\frac{1}{s} \sum_{i=1}^s a_i + \frac{1}{t} \sum_{j=1}^t b_j], \tag{6}$$

从平均的角度描述了  $\mathbf{R}^{(2)}$  与  $\mathbf{R}^{(1)}$  (近似) 等同的程度, 即式(6)可以作为 BD-RED 中的相似性度量。

表 1 数据训练集

Tab. 1 Data training sets

#### 4 实例分析

本部分结合实例来考察 BD-RED 的有效性。表 1 给出了影响夏天天气舒适度的一些指标的相关数据, 试确定影响天气舒适性的简化指标集。若将穿衣指数、温度、湿度、风力视为条件属性, 舒适性视为决策属性, 则该问题的本质是一种属性约简。下面利用 BD-RED 来确定约简结果, 其参数设置如下: 选择 ID3 算法作为规则族的获取策略; 条件属性的信息增益作为属性的排序原则; 式(6)作为规则族之间的相似度量, 即约简强度;  $f(x_1, x_2, \dots, x_n, y, z) = y(1 - z) \sum_{i=1}^n (x_i/n)$  作为相似函数(其中,  $n$  表示条件属性的个数); 并视 等同于各属性的各种取值; 节点的纯度作为规则的可靠度。

由于在实际中数据库中的数据往往带有不同程度的主观或客观上的不确定性, 因而在数据库的简化过程中, 可以存在适当的偏差(即节点纯度和规则族的相似程度均可以适当小于 1)。按照上述参数设置, 表 2 给出了针对 3 种不同约简强度的约简结果。

1) 试验 1 表明, 在保证知识绝对可靠(即节点纯度为 1)和决策能力不变(即规则族的相似程度为 1)的前提下, 各条件属性均是不可缺少的。

2) 试验 2 表明, 在保证知识绝对可靠(即节点纯度为 1)和决策能力基本不变(即规则族的相似程度不小于 0.95)的前提下, 温度和风力是多余属性。

3) 试验 3 表明, 在保证知识基本可靠(即节点纯度不小于 0.71)和决策能力基本不变(即规则族的相似程度不小于 0.83)的前提下, 温度、湿度和风力均是多余属性。

注 5 如果选用不同的参数, 如相似函数的选取, 对相似强度的不同描述, 和对 的不同处理等, 可以得到不同程度的约简结果。

注 6 对于规则族的获取策略的选取, 需根据数据类型特点及数据库大小, 选择合适的决策树算法。

#### 5 结 语

利用决策树算法操作简单、分类速度快的特点, 在建立了规则知识的规范化描述基础上, 提出了一种基于决策树的属性约简方法(简记为 BD-RED), 给出了 BD-RED 的实施策略, 并结合实例进行了分析。结果表明: 1) BD-RED 可以针对不同的数据库, 结合问题的特点及不同的决策理念来选择约简参数; 2) BD-RED 以决策树为基础, 复杂度较低; 3) BD-RED 具有良好的结构特征和较强的可操作性, 可以有效地实现不同决

穿衣指数	温度	湿度	风力	舒适性
较多	很高	很大	没有	N
较多	很高	很大	很大	N
较多	很高	很大	中等	N
正常	很高	很大	没有	P
正常	很高	很大	中等	P
很多	适中	很大	没有	N
很多	适中	很大	中等	N
很多	很高	正常	没有	P
很多	很高	正常	很大	N
较多	适中	很大	没有	N
较多	适中	很大	中等	N
很多	适中	正常	没有	N
很多	适中	正常	中等	N
较多	适中	正常	中等	P
较多	适中	正常	很大	P
正常	适中	很大	很大	P
正常	适中	很大	中等	P
正常	很高	正常	没有	P
很多	适中	很大	很大	N
正常	很高	正常	中等	P

表2 实验结果

Tab.2 Results of testing

试验编号	参数设计		约简结果	约简后的知识库					
	节点纯度	约简强度		穿衣指数	温度	湿度	风力	舒适性	可靠度
试验 1	1	1	穿衣指数, 温度,湿度, 风力	很多	适中			N	1
				很多	很高		很大	N	1
				很多	很高		没有	P	1
				较多		很大		N	1
				较多		正常		P	1
试验 2	1	0.95	穿衣指数, 湿度	很多		很大		N	1
				较多		很大		N	1
				较多		正常		P	1
				正常				P	1
试验 3	0.71	0.83	穿衣指数	很多				N	0.86
				较多				P	0.71
				正常				P	1

策理念下的属性约简,适合不同类型的大规模数据库的约简。但是作为规则族间相似度量的核心任务,如何选取或者建立最优的相似函数,使约简达到最好的效果,还需要进一步的研究。

#### 参考文献:

- [1] HU X H, CERCONE N. Learning in relational databases: A rough set approach[J]. International Journal of Computational Intelligence, 1995, 11(2): 323-338.
- [2] GUAN J W, BELL D A. Rough computational methods for information systems[J]. Artificial Intelligences, 1998, 105(122): 77-103.
- [3] ZHAO S Y, ERIC C C. Tsang on fuzzy approximation operators in attribute reduction with fuzzy rough sets[J]. Information Science, 2008, 178(16): 3163-3176.
- [4] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1998, 36(6): 681-684.
- [5] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 959-965.
- [6] 曹付元, 梁吉业, 钱宇华. 基于信息熵的决策表约简[J]. 计算机应用, 2005, 25(11): 2630-2631.
- [7] 杨明. 一种基于改进差别矩阵的属性约简增量式更新算法[J]. 计算机学报, 2007, 30(5): 815-822.
- [8] SKOWRON A, RAUSZER C, SLOWINSKI R. Intelligent Decision Support Handbook of Applications and Advances of the Rough Sets Theory[M]. Dordrecht: Kluwer Academic Publishers, 1992.
- [9] WANG Jue, WANG Ju. Reduction algorithm based on discernibility matrix the ordered attributes method[J]. Journal of Computer Science & Technology, 2001, 16(16): 489-504.
- [10] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.
- [11] QUINLAN J R. Programs for Machine Learning[M]. San Mateo: Morgan Kaufmann Publish, 1993.
- [12] BREIMAN L, FRIEDMAN J, OLSEN R, et al. Classification and Regression Trees[M]. New York: Chapman & Hall, 1984.
- [13] MEHTA M, AGRAWAL R, RISSANEN J. SLIQ: A fast scalable classifier for data mining[A]. Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology[C]. London: Springer-Verlag, 1996. 18-32.
- [14] SHAFER J, AGRAWAL R, MEHTA M. SPRINT: A scalable parallel classifier for data mining[A]. In Proceedings of the 22nd International Conference on Very Large Databases (VLDB)[C]. San Francisco: Morgan Kaufmann Publishers Lnc, 1996. 544-555.
- [15] RASTOGI R, SHIM K. PUBLIC: A decision tree classifier that integrates building and pruning[A]. In Proceedings of the 24th International Conference on Very Large Databases (VLDB)[C]. San Francisco: Morgan Kaufmann Publishers Lnc, 1998. 404-415.

# 欢迎赐稿！欢迎订阅！