文章编号:1673-9469(2008)01-0108-03

基于粗糙集的决策树规则提取算法

陈建辉,陈 贞
(莆田学院 电子信息工程学系,福建 莆田 351100)

摘要:针对ID3算法用信息增益作为在各级非叶节点上选择属性的标准的局限性,结合统计学独立检验思想,给出一种新的属性依赖性和重要性定义,以新的属性重要性为启发式信息设计决策树规则提取算法。实例分析的结果表明,该算法能提取更为简洁有效的决策规则。

关键词:粗糙集;属性依赖性;决策树;规则提取
中图分类号:TP311.13 **文献标识码**:A

An rules extraction algorithm of decision tree based on rough set theory

CHEN Jian-hui, CHEN Zhen
(Electronic & Information Engineering Department, Putian University, Putian 351100, China)

Abstract: A new attribute dependency and significance were defined with the independent check theory of the statistical aiming at the disadvantages of the standard for choosing the attributes of the branch nodes with the information gain in the ID3 algorithm. An algorithm for rules extraction of decision tree was designed. The new attribute significance was used as the heuristic information in which. The experiment and comparison show that the algorithm provides more precise and simple decision tree.

Key words: rough set; attribute dependency; decision tree; rules extraction

粗糙集理论是一种处理模糊和不精确性问题的新型数学工具,其主要思想就是在保持分类能力不变的前提下,通过知识约简找到核值,导出问题的决策或分类规则。由于粗糙集理论具备很强的定性分析能力,因此已被广泛地应用于机器学习、故障诊断、决策分析、过程控制、模式识别、数据挖掘等领域,并取得了成功。决策规则提取是粗糙集理论中最重要的环节之一。国内外众多学者之所以热衷于研究更快、更好的属性约简算法,其最终目的也是为了决策规则的更简单、更准确,而决策规则的性能直接决定了信息系统的识别率[1]。目前,很多粗糙集理论中的决策规则提取算法都是基于决策树学习的。ID3算法及其衍生C4.5算法被公认是决策树学习中最优秀的算法,而目前基于决策树学习的算法多数是针对ID3和C4.5做进一步的改进研究。ID3算法用信息增益作为在各级非叶节点上选择属性的标准,获得样本集最大的类别信息,但该方法对具有许多输出的检验有严重的偏差。本文利用粗糙集理论,给出了一种新的启发式函数,这种启发式函数不仅考虑了属性的重要性,还兼顾了各属性值的个数,能解决对具有许多输出的检验有严重偏差的问题,一般能得到很好的效果。

1 属性的依赖性及重要性度量

通过信息系统发现知识,主要是用属性来表达知识的分类,各种属性在表达知识分类中的作用是不同的。有些属性是绝对不必要的,去掉这些属性并不影响知识发现;有些属性是绝对必要的,去掉这些属性必然会影响到知识发现;有些属性是相对必要的,它与所有绝对必要属性搭配起来才不影响知识发现。因此,信息系统的属性关于知识发现的重要性应该有一个度量指标[2]。

定义1是结合统计学独立检验思想,给出一种新的依赖性度量,这种依赖性度量实质上是一种包含度[3]。在统计学中,许多调查数据往往用

收稿日期:2007-12-27
作者简介:陈建辉(1974-),男,福建莆田人,硕士,讲师,从事粗糙集理论及应用和数据挖掘方面的研究。

一个表格形式来表示,如果想考察表中的两个因素是否独立,可以用 x^2 检验来解决。x^2 的值越小,说明两因素越独立,反之两因素越相关联,即这两属性的依赖性越强。在属性约简中,也要考虑属性之间的依赖关系,所以可以从这想到定义属性的依赖性度量。

定义 1 五元组 $S = < U, C, D, f, V >$ 是一个决策表,其中 U 为论域,C 为条件属性集,$|U| = M, U/ind(D) = D^* = \{Y_1, Y_2, \cdots, Y_m\}, U/ind(C) = C^* = \{X_1, X_2, \cdots, X_n\}$,定义知识 D 相对于知识 C 的依赖度为

$$\gamma(D^* | C^*) = \frac{1}{M(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(q_{ji} - \frac{|X_i||Y_j|}{M})^2}{\frac{|X_i||Y_j|}{M}}$$

其中 q_{ij} 是知识 C 的第 i 个等价类在知识 D 的第 j 个等价类中的元素的个数,$|X_i|$ 表示知识 C 的第 i 个等价类中所含样本的个数,$|Y_j|$ 表示知识 D 的第 j 个等价类中所含样本的个数。当 $\gamma(D^* | C^*) = 1$ 时,称知识 D 完全依赖于知识 C;当 $0 < k = \gamma(D^* | C^*) < 1$ 时,称知识 D 是 k 度依赖于知识 C;当 $\gamma(D^* | C^*) = 0$ 时,称知识 D 完全独立于知识 C。

定义 2 五元组 $S = < U, C, D, f, V >$ 是一个决策表,其中 U 为论域,C 为条件属性集,D 为决策属性,属性子集 $C' \subseteq C$ 关于 D 的重要性定义为

$$\sigma_{CD}(C') = \gamma(D^* | C^*) - \gamma(D^* | (C - C')^*)$$

当 $C' = \{a\}$ 时,属性 $a \in C$ 关于 D 的重要性定义为

$$\sigma_{CD}(a) = \gamma(D^* | C^*) - \gamma(D^* | (C - \{a\})^*)$$

$\sigma_{CD}(a)$ 的值越大,说明该属性越重要。

2 算法基本思想及步骤

在 ID3 的决策树的每个节点上使用信息增益度量选择测试属性,选择具有最高信息增益的属性作为当前节点的测试属性,该属性使得对结果划分中的样本分类所需的信息量最小。但这种启发式方法往往并不是最优的,即决策树的节点不是最少的。因其偏向于选择取值较多的属性,而属性值较多的属性并不总是最优的属性。本文针对以上问题提出如下的改进方案:在构造决策树的过程上,改进各级非叶节点属性的选择标准,以属性重要性与属性值的个数之比作为启发式信息,避免 ID3 算法中子树重复和某些属性被多次选择的缺点,便于得到更优的决策树。

输入:一个决策表 $S = < U, C, D, f, V >$,U 是论域,C 是条件属性集合,$D = \{d\}$ 是决策属性集合。

输出:最小决策树 T。

步骤 1 计算各条件属性集 a_i 的重要性 $\sigma_{CD}(a_i)$ 与各条件属性集 a_i 值个数 n_{a_i} 之比 $k = \frac{\sigma_{CD}(a_i)}{n_{a_i}}$。

步骤 2 选取一个 k 值最大的属性作为节点(若所有的 k 均为 0,则选择一个属性值个数最少的属性作为节点)。

步骤 3 对所选属性的每个属性值,创建一个分支,并据此划分样本,若某分支的所有样本都在同一个决策类,则该节点成为叶节点,并用该类标记。

步骤 4 对每个非叶节点,递归地形成每个划分上的样本决策树,一旦一个属性出现在一个节点,就不必考虑该节点的任何后代,直到没有剩余属性。

3 实例

表 1 给出了决策表 $S = < U, C, D, f, V >$,其中 $U = \{1, 2, \cdots, 14\}$,$C = \{a_1, a_2, a_3, a_4, a_5\}$,$D = \{d\}$。

首先计算各属性的重要性。从表 1 可知:$M = |U| = 14, m = 2, |Y_1| = 5, |Y_2| = 9$,表 1 为相容决策表,所以 $\gamma(D^* | C^*) = 1$。

$\sigma_{CD}(a_1) = \gamma(D^* | C^*) - \gamma(D^* | (C - \{a_1\})^*) =$

表 1 决策表

Tab.1 Decision table

论域 U		1	2	3	4	5	6	7	8	9	10	11	12	13	14
条件属性集 C	a_1	1	1	2	3	3	3	2	1	1	3	1	2	2	3
	a_2	1	1	1	2	3	3	3	2	3	2	2	2	1	2
	a_3	1	1	1	1	0	0	0	1	0	0	0	1	0	1
	a_4	0	1	0	0	0	1	1	0	0	0	1	1	0	1
决策属性 d		0	0	1	1	1	0	1	0	1	1	1	1	1	0

$$\frac{1}{14}\left(\frac{(1-\frac{10}{14})^2}{\frac{10}{14}}+\frac{(1-\frac{18}{14})^2}{\frac{18}{14}}+\frac{(1-\frac{5}{14})^2}{\frac{5}{14}}+\frac{(0-\frac{9}{14})^2}{\frac{9}{14}}+\right.$$
$$\frac{(1-\frac{10}{14})^2}{\frac{10}{14}}+\frac{(1-\frac{18}{14})^2}{\frac{18}{14}}+\frac{(0-\frac{10}{14})^2}{\frac{10}{14}}+\frac{(2-\frac{18}{14})^2}{\frac{18}{14}}+$$
$$\frac{(1-\frac{10}{14})^2}{\frac{10}{14}}+\frac{(1-\frac{18}{14})^2}{\frac{18}{14}}+\frac{(0-\frac{5}{14})^2}{\frac{5}{14}}+\frac{(1-\frac{9}{14})^2}{\frac{9}{14}}+$$
$$\frac{(0-\frac{5}{14})^2}{\frac{5}{14}}+\frac{(1-\frac{9}{14})^2}{\frac{9}{14}}+\frac{(1-\frac{10}{14})^2}{\frac{10}{14}}+\frac{(1-\frac{18}{14})^2}{\frac{18}{14}}+$$
$$\left.\frac{(0-\frac{5}{14})^2}{\frac{5}{14}}+\frac{(1-\frac{9}{14})^2}{\frac{9}{14}}\right)=0.622$$

同理求得 $\sigma_{CD}(a_2)=0,\sigma_{CD}(a_3)=0,\sigma_{CD}(a_4)=0.311$

显然 $\frac{\sigma_{CD}(a_1)}{3}>\frac{\sigma_{CD}(a_4)}{2}>\frac{\sigma_{CD}(a_2)}{3}=\frac{\sigma_{CD}(a_3)}{2}$，所以选择 a_1 为初始节点，对于 a_1 的每个属性值构造一个分支，对于属性值 2 对应的都是相容决策规则，所以划分结束，作为叶节点，对于属性值 1 和 3 重复上面的步骤，经过以上三步，得到最小规则集如表 2 所示，决策树如图 1 所示。

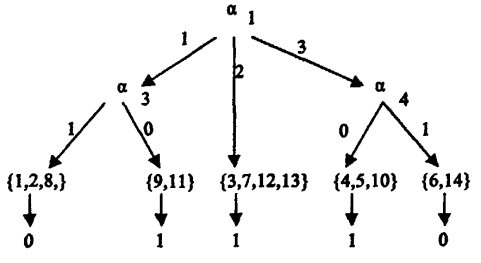

图1 决策树

Fig. 1 Decision tree

表 2 最小规则表

Tab.2 Minimal decision rules

序号	决策规则
1	$a_1=1\wedge a_3=1\rightarrow d=0$
2	$a_1=1\wedge a_3=0\rightarrow d=1$
3	$a_1=2\rightarrow d=1$
4	$a_1=3\wedge a_4=0\rightarrow d=1$
5	$a_1=3\wedge a_4=1\rightarrow d=0$

表 3 为 RITIO 算法[4]对于表 1 的决策表得到的规则集共有 7 条规则，其中第 3 条规则是不确定的，它与表 1 的第 6 个对象矛盾。通过表 2 与表 3 对比可以发现本文算法得到的规则是 5 条，且 1 条规则长度为 1，4 条规则为 2，而 RITIO 算法得到的规则是 7 条，一条规则是不确定的，3 条规则长度为 2，3 条规则为 3，说明本文算法更简洁有效。这是因为 RITIO 算法采用熵测度来度量决策表中条件属性和决策属性的相关性，通过逐步删除最不相关属性并从相容的对象中提取规则，此方法的结果抗噪能力较好，但规则前件比较复杂，有冗余且有些规则不确定。本文提出的算法借鉴统计学独立检验思想和粗糙集的某些概念，以属性重要性与属性值的个数之比作为启发式信息提取规则，一定程度地避免了上述问题。同时，对于表 1 所示的决策表，本文算法得到的决策树与最近研究的成果[5]一样。

表 3 RITIO 算法的最小规则表

Tab.3 Minimal decision rules of the RITIO algorithm

序号	决策规则
1	$a_1=1\wedge a_3=1\rightarrow d=0$
2	$a_1=2\wedge a_3=1\rightarrow d=1$
3	$a_2=0\rightarrow d=1$
4	$a_3=0\wedge a_4=0\rightarrow d=1$
5	$a_1=3\wedge a_3=1\wedge a_4=0\rightarrow d=1$
6	$a_1=3\wedge a_3=0\wedge a_4=1\rightarrow d=0$
7	$a_1=3\wedge a_3=1\wedge a_4=1\rightarrow d=0$

4 结语

本文提出的算法不必先求出属性约简，能解决那种因组合爆炸而引起的规则量太大的问题，而且本文算法得到的决策树与其它算法得到的决策树进行了比较，结果表明本文构造的决策树比较简单，而且该树中的所有属性集一定是一个属性约简。本文算法的复杂性主要集中在各属性的重要性的计算上，可以采用树形的方法来求解 q_{ij}，从而降低复杂性。

参考文献：

[1] 潘巍，王阳生，杨宏戟. 粗糙集理论中求取最小决策规则的研究[J]. 计算机科学，2007，34(4)：185－187.

[2] 张文修，梁怡，吴伟志. 信息系统与知识发现[M]. 北京：科学出版社，2003.1－95.

[3] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2001.

[4] WU XD, UPRANI D. Induction by attribute elimination [J]. IEEE Trans on Knowledge and Data Engineering, 1999, 11 (5)：803－812.

[5] 孙林，徐久成，马媛媛. 基于新的条件熵的决策树规则提取方法[J]. 计算机应用，2007，27(4)：884－887.

(责任编辑 刘存英)

基于粗糙集的决策树规则提取算法

作者： 陈建辉， 陈贞， CHEN Jian-hui， CHEN Zhen
作者单位： 莆田学院,电子信息工程学系,福建,莆田,351100
刊名： 河北工程大学学报（自然科学版） ISTIC
英文刊名： JOURNAL OF HEBEI UNIVERSITY OF ENGINEERING(NATURAL SCIENCE EDITION)
年，卷(期)： 2008,25(1)
被引用次数： 1次

参考文献(5条)

1. 潘巍;王阳生;杨宏戟 粗糙集理论中求取最小决策规则的研究[期刊论文]-计算机科学 2007(04)

2. 张文修;梁怡;吴伟志 信息系统与知识发现 2003

3. 张文修;吴伟志;梁吉业 粗糙集理论与方法 2001

4. WU XD;UPRANI D Induction by attribute elimination[外文期刊] 1999(05)

5. 孙林;徐久成;马媛媛 基于新的条件熵的决策树规则提取方法[期刊论文]-计算机应用 2007(04)

本文读者也读过(10条)

1. 林金山.邵立康.Lin,Jinshan.Shao,Likang 基于串行通信的图像信息采集系统设计[期刊论文]-微计算机信息 2006,22(22)

2. 沈文龙.SHEN Wen-long 实现以RFID卡仿真磁卡的模块设计[期刊论文]-福建农林大学学报（自然科学版） 2007,36(4)

3. 吴晓.陈继信.WU Xiao.CHEN Jixin 新型可倾式重力浇铸机控制系统的研制[期刊论文]-机床与液压2006(7)

4. 黄斌.陈德礼.HUANG Bin.CHEN De-li 基于MPEG-7形状轮廓描述编码的数字水印算法[期刊论文]-陕西科技大学学报（自然科学版）2008,26(6)

5. 王立宏.孙立民.孟佳娜.WANG Li-hong.SUN Li-min.MENG Jia-na 数值离散化中粒度熵与分类精度的相关性[期刊论文]-重庆大学学报（自然科学版）2008,31(1)

6. 林金山.LIN Jin-shan 基于PowerBuilder ActiveX的学籍管理系统设计与开发[期刊论文]-山西农业大学学报(自然科学版)2006,26(4)

7. 洪家军.吴金龙 基于NS-2的Ad Hoc网络路由协议性能仿真[期刊论文]-江汉大学学报（自然科学版）2007,35(1)

8. 陈学军.陶红艳.高云.余成波.CHEN Xue-jun.TAO Hong-yan.GAO Yun.YU Cheng-bo 发电机定子线棒在线监测系统的研制[期刊论文]-压电与声光2007,29(6)

9. 管红波.田大钢 基于属性重要性的决策树规则提取算法[期刊论文]-系统工程与电子技术2004,26(3)

10. 车艳.李少芳.CHE Yan.LI Shao-fang 基于网络考试系统的UML建模[期刊论文]-佳木斯大学学报（自然科学版）2007,25(6)

引证文献(1条)

1. 刘志强.张维 基于多决策属性的刀具选择规则提取算法研究[期刊论文]-锻压装备与制造技术 2012(3)