基于决策熵的决策树规则提取方法

孙 林,徐久成,马媛媛
(河南师范大学 计算机与信息技术学院,河南 新乡 453007)

摘 要:在决策表中,决策规则的可信度和对象覆盖度是衡量决策能力的重要指标。以知识粗糙熵为基础,提出决策熵的概念,并定义其属性重要性;然后以条件属性子集的决策熵来度量其对决策分类的重要性,自顶向下递归构造决策树;最后遍历决策树,简化所获得的决策规则。该方法的优点在于构造决策树及提取规则前不进行属性约简,计算直观,时间复杂度较低。实例分析的结果表明,该方法能获得更为简化有效的决策规则。

关键词:粗糙集;决策熵;决策树;决策规则
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2007)06-0097-04

Algorithm for Rules Extraction of Decision Tree Based on Decision Information Entropy

SUN Lin, XU Jiu-cheng, MA Yuan-yuan
(College of Computer & Information Technology, Henan Normal University, Xinxiang 453007, China)

Abstract: In decision table, the reliability and objects coverage of decision rules are the most important performance metric for estimating decision ability. Based on rough entropy of knowledge, a new decision information entropy is proposed. The new significance of an attribute is defined, which is based on this entropy. In the process of constructing decision tree, condition attributes are considered to estimate the significance for decision classes. A procedure for reduction of traversing decision rules is also constructed, and helps to get more precise rules. The benefit of the method is that it needn't attribute reduction before extracting decision rules, and its computation is simple and intuitionistic. The experiment and comparison show that the algorithm provides more precise and simple decision rules.

Key words: rough set; decision information entropy; decision tree; decision rules

0 引 言

粗糙集理论是一种新的可以分析模糊和不确定知识的数学工具,已广泛应用于知识发现和机器学习等领域。决策表属性简化、决策规则的简化是粗糙集理论与实际应用的主要研究方向之一。近年来,许多学者通过不同的方法从不同的角度对决策规则获取(值约简)做了深入的研究[1~4]。对于决策表知识约简,确定性粗糙集模型是以一致规则为研究对象的,并没有涉及对不一致规则的处理。文献[2]指出利用文献[3]提出的值约简算法得到的规则,仍存在属性冗余和规则冗余,并给出了反例。文献[4]也指出现有的值约简方法也存在一定的不足。为了解决以上问题,采用目前归纳学习中最有效的决策树分类规则学习方法[5,6],但构造最优决策树已被证明是 NP-Hard 问题[7],所以在属性的选择中,采用更优的启发函数来构造决策树、提取决策规则的方法显然具有重要的实用意义。

在决策表中,决策规则的可信度和对象覆盖度是衡量决策能力的重要指标。由此以知识粗糙熵为基础,给出一种新的粗糙熵定义——决策熵,并定义其属性重要性;然后以条件属性子集的决策熵来度量其对决策分类的重要性,选择决策熵最小且涵盖最多决策分类对象的属性为分枝结点,自顶向下递归构造决策树;最后遍历决策树,简化所获得的决策规则。该方法的优点在于构造决策树及提取决策规则前不进行属性约简,计算直观,时间复杂度较低。理论分析和实验结果表明,该方法是有效的。

1 主要概念

定义 1 五元组 $S=(U,C,D,V,f)$ 是一个决

收稿日期:2006-09-20
基金项目:河南省自然科学基金项目(0511011500);河南省高校新世纪优秀人才支持计划(2006HANCET-19)
作者简介:孙 林(1979-),男,河南南阳人,硕士研究生,研究方向为粗糙集理论、数据挖掘;徐久成,教授,研究方向为粗糙集理论、粒计算、数据挖掘等。

策表,其中 U 为论域;C 为条件属性集;D 为决策属性集且 $C \cap D = \varnothing$;$V = \bigcup \{V_a \mid a \in C \cup D\}$,$V_a$ 为属性 a 的值域;$f: U \times (C \cup D) \rightarrow V$ 是一个信息函数,它对一个对象的每一个属性赋予一个信息值,即 $\forall a \in C \cup D, x \in U$,有 $f(x,a) \in V_a$;每一个属性子集 $P \subseteq C \cup D$ 决定了一个二元不可区分关系 IND(P):IND(P) = $\{(x,y) \in U \times U \mid f(x,a) = f(y,a), \forall a \in P\}$,关系 IND($P$) 可确定 U 的一个划分,用 U/IND(P) 表示,简记为 U/P,U/P 中的任何元素 $[x]_P = \{y \mid f(x,a) = f(y,a), \forall a \in P\}$ 称为等价类(划分块)。

定义 2 设 $X \subseteq U$ 为论域的一个子集,属性集合 $P \subseteq C$,用 $\underline{P}(X) = \bigcup \{[x]_P \mid [x]_P \subseteq X\}$ 表示 X 的 P - 下近似集,决策属性集 D 的 P - 正域 $\mathrm{Pos}_P(D)$ 定义为:

$$\mathrm{Pos}_P(D) = \bigcup \{\underline{P}(X) \mid X \in U/D\} \tag{1}$$

并记 $\gamma_P(D) = |\mathrm{Pos}_P(D)| / |U|$,其中 $|X|$ 表示集合 X 的基数。

定义 3[8] 在决策表 $S = (U,C,D,V,f)$ 中,若 $\mathrm{Pos}_C(D) = U$,则称决策表 S 是一致决策表,否则称决策表 S 为不一致决策表。基于一致决策表获取的知识都是确定的知识,表明其不含有不一致的(冲突的)对象,而不一致决策表则相反。

定义 4 在决策表 $S = (U,C,D,V,f)$ 中,决策规则为隐含式,记为 $(C_1,c1) \wedge (C_2,c2) \wedge \cdots \wedge (C_k,ck) \rightarrow (D,d)$,其中 $ci \in V_{Ci}, C_i \in C, i = 1,2,\cdots,k$,$k \leqslant |C|, d \in V_D$。

2 知识的决策熵与决策树规则提取方法

在决策表 $S = (U,C,D,V,f)$ 中,属性约简的最终目标是在保持决策表 S "决策能力"不变的前提下,去除多余条件属性。在决策应用中,决策规则的可信度和对象覆盖度都是衡量决策能力的重要指标,但是经典粗糙集理论中的知识粗糙熵并没有完全客观地反映决策表决策能力的变化情况。由此本节在知识粗糙熵的基础上,提出了一种新的粗糙熵定义——决策熵,并定义其属性重要性,然后以新的属性重要性为启发信息设计决策树规则的提取方法。

2.1 知识的决策熵

定义 5[9] 设 U 是一个论域,属性集合 R 在 U 上导出的划分 $U/R = \{R_1,R_2,\cdots,R_m\}$,则 R 在 U 上导出划分 U/R 的粗糙熵定义为:

$$E(R) = \sum_{i=1}^{m} \frac{|R_i|}{|U|} \log |R_i| \tag{2}$$

其中 $|R_i| / |U|$ 表示 R_i 在论域 U 上的概率。

为了研究能够体现对象覆盖度的知识粗糙熵,引入下面的引理。

引理 1[10] 设 P,Q 为论域 U 上的两个等价关系集合,则有 $U/(P \cup Q) = U/P \cap U/Q$ 成立。

引理 1 的证明参考文献[10]。

这样,在决策表 $S = (U,C,D,V,f)$ 中,属性集合 $P \cup D(P \subseteq C)$ 的粗糙熵可有如下定义。

定义 6 设 U 是一个论域,条件属性集合 P 在 U 上导出的划分 $U/P = \{X_1,X_2,\cdots,X_n\}$,$D = \{d\}(U/D = \{D_1,D_2,\cdots,D_t\})$ 为决策概念集,则属性集合 $P \cup D$ 的粗糙熵定义为:

$$E(P \cup D) = \sum_{i=1}^{n} \sum_{j=1}^{t} \frac{|X_i \cap D_j|}{|U|} \log |X_i \cap D_j| \tag{3}$$

在 $P \cup D$ 的粗糙熵定义中,$|X_i \cap D_j| / |U|$ 代表了某一决策规则的对象覆盖度,所以该粗糙熵定义就反映了决策表"决策能力"变化的一个重要指标。

定义 7[11] 设 U 是一个论域,$P(U/P = \{X_1, X_2,\cdots,X_n\})$ 为一个条件属性集合,$D = \{d\}(U/D = \{D_1,D_2,\cdots,D_t\})$ 为决策概念集,则决策概念集 D 的粗糙熵定义为:

$$E(D_P) = -\sum_{i=1}^{n} \sum_{j=1}^{t} \frac{|X_i|}{|U|} \log \frac{|X_i \cap D_j|}{|X_i|} \tag{4}$$

由定义 7 知,在条件属性集合的划分 $U/P = \{X_1,X_2,\cdots,X_n\}$ 中,存在两种情况:$x \in X_i \subseteq D_j$ 和 $x \in X_i \not\subset D_j$。其中 $i = 1,2,\cdots,n, j = 1,2,\cdots,t$。在前一种情况下,$D_j$ 中的 x 是完全可确定的,因此,只需考虑 $x \in X_i \not\subset D_j$ 的情况,于是决策概念集 D 关于知识 P 的粗糙熵可简化为:

$$E(D_P) = -\sum_{i=1}^{n} \sum_{j=1}^{k} \frac{|X_i|}{|U|} \log \frac{|X_{ij}|}{|X_i|} \tag{5}$$

其中,$X_{i1},X_{i2},\cdots,X_{ik}(k \leqslant t)$ 是 X_i 与 $D_1,D_2,\cdots,D_t$ 的非空交集。

在决策概念集的粗糙熵定义中,$|X_i \cap D_j| / |X_i|$ 代表了决策表所产生某一决策规则的可信度。这样可以把两种粗糙熵的定义结合起来,使其完全客观地反映决策表决策能力的真实变化情况。在此基础上,提出了一种新的粗糙熵信息论定义形式 —— 决策熵。

定义 8 设 U 是一个论域,P 是 U 上的一个条件属性集合,$D = \{d\}$ 为决策概念集,则 P 关于决策 D 的决策熵记为 $E(D \mid P)$,定义为:

$$E(D \mid P) = E(D_P) + E(P \cup D) \tag{6}$$

有了知识的决策熵定义,可以得到与其相应的属性重要性的度量方式。

定义 9 在决策表 $S = (U,C,D,V,f)$ 中,$B \subseteq$

C,任意属性 $a \in C - B$ 的属性重要性定义为:

$$\mathrm{SGF}(a,B,D) = E(D \mid B) - E(D \mid B \cup \{a\}) \tag{7}$$

特别当 $B = \varnothing$ 时,$\mathrm{SGF}(a,\varnothing,D) = -E(D \mid \{a\})$。

$\mathrm{SGF}(a,B,D)$ 的值越大,说明在已知 B 的条件下,属性 $a \in C - B$ 关于知识 B 就越重要。在计算 $\mathrm{SGF}(a,B,D)$ 的过程中,每次循环时条件属性子集 B 的 $E(D \mid B)$ 均不变,那么求 $\mathrm{SGF}(a,B,D)$ 最大的属性 a 就是求 $E(D \mid B \cup \{a\})$ 最小的属性 a。所以,若把 $\mathrm{SGF}(a,B,D)$ 作为搜索最小或次优知识约简的启发信息时,就只需计算 $E(D \mid B \cup \{a\})$,这样可以避免计算 $E(D \mid B)$,减少了计算量,进而减小了搜索空间。

2.2　基于决策熵的决策树规则提取方法

决策树是指用树形结构来表示决策集合,这些决策集合通过对数据集的分类产生决策规则。若 $S = (U,C,D,V,f)$ 是一致决策表,则决策树的各叶子结点只对应相同决策类的对象,即每个叶子结点对应的是确定性决策规则,其可信度(对象的条件概率分布)等于 1;否则决策树的某些叶子结点对应不同决策类的对象,这样的叶子结点对应的是不确定性决策规则,其可信度小于 1。由此以知识决策熵的属性重要性为启发信息来设计值约简方法。首先从空树 T 开始,逐步加入条件属性,选择最小的知识粗糙熵,以涵盖最多决策分类对象的属性为分枝结点,自顶向下递归构造决策树;然后根据分块处理的思想,尽量以少的属性提取隐含在决策表中有用的决策规则;最后删除所有不影响规则表达的冗余条件属性值,简化决策规则。该方法的具体操作步骤描述如下:

算法 1

输入:决策表 $S = (U,C,D,V,f)$;

输出:最简决策树 T 和决策规则集。

Step1:合并决策表 S 中的相同对象;

Step2:初始化 $B = \varnothing$,T 为空树;

Step3:对任意属性 $a \in C - B$,计算 $E(D \mid B \cup \{a\})$;

Step4:选择使 $E(D \mid B \cup \{a\})$ 最小的属性 a 为决策树 T 的根结点(或分支结点),

1) if 有多个属性同时使 $E(D \mid B \cup \{a\})$ 达到最小值 then 从中选取一个属性 a 使得与 B 的依赖性 $\gamma_{B\cup\{a\}}(D)$ 最大

2) if 仍有多个属性使 $\gamma_{B\cup\{a\}}(D)$ 达到最大值 then 选择顺序靠前的属性

Step5:用选择的属性 a 对 U 进行分类,即计算 $U/\{a\} = \{U_1,U_2,\cdots,U_t\}$,开始建立子决策表(即决策树的分支) $S_i = (U_i,C,D,V,f)$,其中 $i = 1,2,\cdots,t$;

Step6:if 分支 $S_i(i = 1,2,\cdots,t)$ 中的所有对象 U_i 具有相同的决策属性值 then{ 在分支 S_i 下生成一个叶子结点,标识其决策属性值,遍历根到该叶子结点的一条路径,产生相应的决策规则,

if 该决策规则中的任一非叶子结点去掉后,在 S_i 中仍能唯一表示

then 继续去掉第 2,3,… 个非叶子结点,直到不能在 S_i 中唯一表示

}

else $B = B \cup \{a\}$

Step7:if $B = C$ 或 U 被决策树分支完全分类 then 输出决策树 T 和决策规则集,结束

else 转 Step3

用文献[12]的计算划分和正区域方法,分析可得 Setp3 到 Step7 总的最坏时间复杂度为 $O(|C||U|) + O((|C|-1)|U|) + O((|C|-2)|U|) + \cdots + O(|U|) = O(|C|^2|U|)$,因而算法 1 最坏的时间复杂度为 $O(|C|^2|U|)$。

与文献[5]多变量决策树构造方法相比,算法 1 得到的是单变量决策树。其中 Step4 考虑了属性之间的依赖关系,易于消除冗余属性;Step5 采用分块处理的方法,弥补了 ID3 算法容易导致决策树中子树重复和某些属性在同一决策树中被多次选择的不足;Step6 对循环提取的原始决策规则进行化简,删除所有不影响规则表达的冗余条件属性及属性值,这就保证了所提取的决策规则最小,即包含条件属性及其属性值最少,且在约简表中唯一表示。

3　应用实例分析与比较

表 1 给出了一致决策表 $S = (U,C,D,V,f)$,其中 $U = \{1,2,\cdots,14\}$,$C = \{a_1,a_2,a_3,a_4\}$,$D = \{d\}$。

表 1　一致决策表

U	1	2	3	4	5	6	7	8	9	10	11	12	13	14
a_1	1	1	2	0	0	0	2	0	1	1	0	2	2	1
a_2	1	1	1	2	0	0	0	2	0	2	2	2	1	2
a_3	1	1	1	1	0	0	0	0	0	1	1	1	0	0
a_4	0	1	0	0	0	1	1	0	0	0	1	1	0	1
d	0	0	1	1	1	0	1	1	1	0	0	1	1	1

用一致决策表(见表 1)来验证算法 1 的有效性,可得到一棵与最小确定性决策规则集(见表 2)对应的最小决策树(见图 1)。

对于表 1 所示的一致决策表,文献[1]中 RITIO 算法共得到 7 条规则,其中有一条规则是不一致的,它与

表 1 的第 6 个对象矛盾,文献[13]中 LEM2 算法也可得到 7 条规则;但以上两种算法得到规则集的数目均比表 2 所示规则集多。对于不一致决策表而言,由算法 1 得到的决策树,不一致对象对应的决策属性值有两个,且简化后得到的不确定性决策规则的可信度均小于 1。

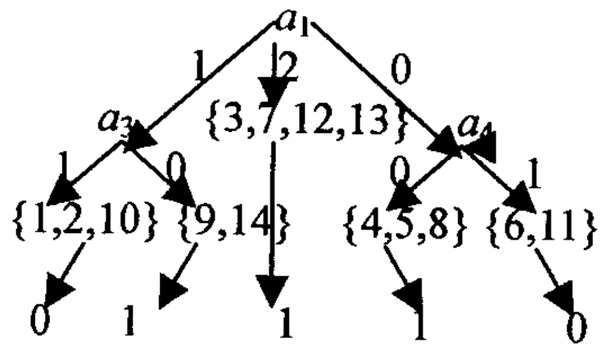

图 1 决策树

表 2 最小确定性决策规则集

序号	决策规则集	可信度	覆盖度
1	$(a_1,1) \wedge (a_3,1) \rightarrow (d,0)$	1	3/14
2	$(a_1,1) \wedge (a_3,0) \rightarrow (d,1)$	1	2/14
3	$(a_1,2) \rightarrow (d,1)$	1	4/14
4	$(a_1,0) \wedge (a_4,0) \rightarrow (d,1)$	1	3/14
5	$(a_1,0) \wedge (a_4,1) \rightarrow (d,0)$	1	2/14

4 结束语

在决策表中,以知识决策熵的属性重要性为启发信息,自顶向下递归构造决策树,然后遍历决策树,并简化所获得的决策规则。实例分析的结果表明,该算法为从决策表中搜索最小决策规则提供了一种有效的方法,并且该研究可进一步扩展粗糙集理论的应用领域。

参考文献:

[1] Wu X D, Urpani D. Induction by attribute elimination[J]. IEEE Transaction on Knowledge and Data Engineering, 1999,11(5):805-812.

[2] 林嘉宜,彭 宏,郑启伦.一种新的基于粗糙集的值约简算法[J].计算机工程,2003,29(4):70-71.

[3] 常犁云,王国胤,吴 渝.一种基于 Rough Set 理论的属性约简及规则提取方法[J].软件学报,1999,10(11):1206-1211.

[4] 黄 兵,周献中.不一致决策表中规则提取的矩阵算法[J].系统工程与电子技术,2005,27(3):441-445.

[5] 苗夺谦,王 珏.基于粗糙集的多变量决策树构造方法[J].软件学报,1997,8(6):425-431.

[6] 刘小虎,李 生.决策树的优化算法[J].软件学报,1998,9(10):797-800.

[7] Hong J R. AE1:An extension matrix approximate method for general covering problem[J]. International Journal of Computer and Information Science,1985,14(6):421-437.

[8] 王国胤.决策表核属性的计算方法[J].计算机学报,2003,26(5):611-615.

[9] Liang J Y, Shi Z Z. The information entropy, rough entropy and knowledge granulation in rough set theory[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,2004,12(1):37-46.

[10] Guan J W, Bell D A. Rough computational methods for information systems[J]. Artificial Intelligences, 1998,105:77-103.

[11] 郑 芳,吴云志,杭小树.粗集理论中知识的粗糙性研究[J].计算机工程与应用,2002,38(4):98-101.

[12] 徐章艳,刘作鹏,杨炳儒,等.一个复杂度为 max($O(|C||U|)$,$O(|C|^2|U|)$)的快速属性约简算法[J].计算机学报,2006,29(3):391-399.

[13] Grzymala-Bausse D M, Grzymala-Busse J W. The usefulness of a machine learning approach to knowledge acquisition [J]. Computational Intelligence,1995,11(2):268-279.

(上接第 96 页)

MFP 算法只需对事务数据库扫描一次,就可把事务数据库转换成 MFP 树,然后对 MFP 树进行挖掘。实验表明:MFP 算法的关联规则挖掘时间效率较高,是一种切实高效的挖掘方法。

参考文献:

[1] 刘乃丽,李玉忱.一种基于 FP_tree 的最大频繁项目集挖掘算法[J].计算机应用,2005,25(5):999-1000.

[2] 毛国君,段立娟.数据挖掘原理与算法[M].北京:清华大学出版社,2005.

[3] Goebel M, Gruenwald L. A Survey of Data Mining and Knowledge Discovery Software Tools[J]. SIGKDD Explorations,1999,1(5):20-23.

[4] 蒋良孝.一种基于 FP-增长的决策规则挖掘算法[J].计算机科学,2003,32(6):23-25.

[5] Han J,Pei J. Freespan:Frequent pattern-projected Sequential pattern Mining[R]. In Technical Report CMPT2000-06,Simon Fraser University,2000:6-12.

[6] 高 俊,何守才.布尔型关联规则挖掘算法[J].计算机工程,2006,32(1):116-118.

[7] Han Jiawei. Data Mining:Concepts and Techniques[D]. Burnaby:Simon Fraser University,2000:155-163.

[8] 张云涛,龚 玲.数据挖掘原理与技术[M].北京:电子工业出版社,2004.