

粗糙集理论中属性相对约简算法

张腾飞,肖健梅,王锡淮

(上海海事大学电气自动化系,上海 200135)

摘 要: 粗糙集理论是近年来发展起来的一种有效地处理模糊和不确定性知识的数学工具,而求核与约简是粗糙集理论中的两个重要问题,现已证明求决策表所有约简和最小约简是一个典型的 NP 难题.本文在分析粗糙集理论的基础上,发现了正区域的一些有用性质,提出了一种利用正区域直接求核的方法,并利用正区域的启发式信息给出了两种相对约简算法.

关键词: 粗糙集; 求核; 相对约简; 决策表

中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2005) 11-2080-04

Algorithms of Attribute Relative Reduction in Rough Set Theory

ZHANG Teng-fei, XIAO Jian-mei, WANG Xi-huai

(Department of Electrical and Automation, Shanghai Maritime University, Shanghai 200135, China)

Abstract: Rough set is a valid mathematical theory developed in recent years, which has the ability to deal with imprecise, uncertain, and vague information. The core and reduction of attributes are two important topics in the research on rough set theory. It has been proven that computing all the reductions and the optimal (minimal) reduction of decision table is a NP-hard problem. In this paper, Rough set theory is deeply investigated; a number of useful properties of the positive region are discovered. Based on the above findings, we present a calculation algorithm for core directly. And then, two algorithms for relative reduction based on the positive region are designed.

Key words: rough set theory; finding core; relative reduction; decision table

1 引言

粗糙集理论是波兰数学家 Z. Pawlak 教授提出的一种新型的处理模糊和不确定性知识的数学工具^[1,2]. 经过二十多年的研究和发展,粗糙集理论已经在决策与分析、故障诊断、模式识别、数据挖掘、系统建模、动态目标识别及跟踪等领域取得了很大的成功^[3~6].

粗糙集理论是以不可分辨关系为基础,通过引入上近似(upper approximation)集和下近似(lower approximation)集来描述一个集合.信息系统的求核与约简是粗糙集理论和应用研究的焦点问题.信息系统分为无决策信息系统和有决策信息系统,在实际应用中大多为有决策信息系统,用决策表来表示,它是粗糙集研究的主要对象,因此决策表信息系统的求核与约简的研究更为重要.由于求所有属性约简是 NP 难题,因此到目前为止,还没有一个高效的求最佳与所有属性约简的算法.

不过,在实际应用中,往往只要求出某种次优的属性约简就可以了.为此,人们已提出了若干个属性求核和约简算法,其中应用较多的是基于差别矩阵以及在此基础上的一些改进

算法^[7,8];文献 9 给出了基于信息论的方法,用信息熵作为选择重要属性的启发式信息;文献[10]基于变精度粗糙集理论,利用可辨识属性矩阵研究了不协调目标信息系统的知识约简算法;文献[11]将包含度概念和证据理论引入到粗糙集理论中,建立了包含度与粗糙集数据分析中的度量之间的关系,这几种方法的主要缺点是空间复杂度高,计算繁琐,不适用大规模数据.

经过深入分析粗糙集理论,本文发现了正区域的一些性质,提出了一种直接求核的方法,并利用正区域的启发式信息给出了两种属性相对约简算法.实验结果表明,本文的方法简单有效.

2 粗糙集的基本理论

决策表是一类特殊而重要的知识表达系统,多数决策问题都可以用决策表形式来表达,这一工具在粗糙集理论中起着重要的作用.

决策表可以根据知识表达系统定义如下:

设 $S = \langle U, R, V, f \rangle$ 为一知识表达系统, U 是论域, $R = C \cup D$, $C \cap D = \emptyset$, C 称为条件属性集, D 称为决策属性集. V

为属性值的集合, $f: U \times R \rightarrow \mathcal{V}$ 是一个信息函数, 它指定 U 中每一个对象 x 的属性值. 具有条件属性和决策属性的知识表达系统称为决策表.

定义 1 在信息系统 S 中, 对于属性子集 $P \subseteq R$, 不可分辨关系定义为:

$$IND(P) = \{ \langle x, y \rangle \in (U \times U) : \forall a \in P, a(x) = a(y) \}$$

显然 $IND(P)$ 也是等价关系, 对象 x 在属性集 P 上的等价类

$$[x]_{IND(P)} = \{ y : y \in U, xIND(P)y \}.$$

U/R 表示 R 的所有等价类.

定义 2 给定信息系统决策表 $S = \langle U, R, V, f \rangle$, 对于每个子集 $X \subseteq U$ 和不可分辨关系 P , X 的下近似集和上近似集可以分别定义为:

$$RX = \bigcup \{ y \in U/R \mid Y \subseteq X \}$$

$$\overline{RX} = \bigcup \{ y \in U/R \mid Y \cap X \neq \emptyset \}.$$

下近似集 RX 也称为 X 的 R 正区域, 记为: $POS_R(X)$.

众所周知, 知识库中知识(属性)并不是同等重要的, 甚至其中有些知识是冗余的. 这就需要知识约简. 所谓知识约简就是在保持分类能力不变的条件下, 删除其中不相关或不重要的知识.

定义 3 设 R 为一族等价关系, $r \in R$, 如果

$$IND(R) = IND(R - \{ r \}),$$

则称 r 为 R 中不必要的, 否则称 r 为 R 中必要的, 不可以约简的. 如果每一个 $r \in R$ 都为 R 中必要的, 则称 R 为独立的; 否则称 R 为依赖的.

设 $P \subseteq R$, 若 P 是独立的, 且 $IND(P) = IND(R)$.

则称 P 为 R 的一个约简;

R 的约简往往不止一个, 所有约简的交集称为核, 记作 $Core(R)$.

在信息系统决策表的应用中, 需要研究条件属性的分类相对于决策属性的分类之间的关系, 因此相对约简和相对核的概念十分重要.

定义 4 设 P 和 Q 为 U 中的等价关系, Q 的 P 正区域记为 $POS_P(Q)$, 即

$$POS_P(Q) = \bigcup_{x \in U/P} PX$$

Q 的 P 正区域是 U 中所有根据分类 U/P 的信息可以准确划分到关系 Q 的等价类中去的对象集合.

定义 5 设 P 和 Q 为 U 中的等价关系, $r \in P$, 如果 $POS_P(Q) = POS_{(P - \{ r \})}(Q)$, 则称 r 为 P 中 Q 不必要的, 否则 r 为 P 中 Q 必要的.

如果 P 中每个 r 都为 Q 必要的, 则称 P 为 Q 独立的.

设 $R \subseteq P$, 如果 R 为 Q 独立的, 且

$$POS_R(Q) = POS_P(Q).$$

则称 R 为 P 的 Q 约简.

P 的所有 Q 约简的交集称为 P 的 Q 核. 记为 $Core_Q(P)$. P 的 Q 核是 P 中所有 Q 必要的原始关系构成的集合.

3 属性求核算法

在粗糙集理论中, 相对核与约简都是基于正区域而定义

的. 对于信息系统决策表 $S = \langle U, R, V, f \rangle$, 其中 U 为论域, $R = C \cup D$, 设属性 $a_i \in C$, 由不可分辨关系的定义可知关系 C 要比 $(C - \{ a_i \})$ 对 U 的分类细, 再由定义 4, 显然有正区域的如下性质:

性质 1 $POS_{(C - \{ a_i \})}(D) \subseteq POS_C(D)$.

定理 1 a_i 是 C 中的一个属性, 若满足

$$POS_{(C - \{ a_i \})}(D) = POS_C(D)$$

则 a_i 不是 C 相对于决策属性 D 的核属性.

证明: 由定义 5 可知, 若上式成立, 则 a_i 就是 C 中相对于决策属性 D 不必要的, 显然, a_i 不是核属性.

定理 2 C 中的一个属性 a_i 是 C 相对于决策属性 D 的核属性的充分必要条件是

$$POS_{(C - \{ a_i \})}(D) \neq POS_C(D)$$

证明:

充分性. 如果 $POS_{(C - \{ a_i \})}(D) \neq POS_C(D)$, 根据定义 5, a_i 就是 C 中必要的, 即为不可以约简的, 为了说明属性 a_i 为核属性, 采用反证法.

假设 a_i 为非核属性, 根据核的定义, 至少有一个约简 P 不包含属性 a_i . 由于 P 为 C 的约简, 且不包含 a_i , 由性质 1 可得:

$$POS_P(D) \subseteq POS_{(C - \{ a_i \})}(D) \subseteq POS_C(D)$$

根据约简的定义可知必然有

$$POS_P(D) = POS_C(D)$$

所以: $POS_{(C - \{ a_i \})}(D) = POS_C(D)$

这与条件 $POS_{(C - \{ a_i \})}(D) \neq POS_C(D)$ 相矛盾. 充分性得证.

必要性. 如果 $POS_{(C - \{ a_i \})}(D) = POS_C(D)$, 由定理 1 可知 a_i 一定为非核属性. 所以如果 a_i 为核属性, 则 $POS_{(C - \{ a_i \})}(D) \neq POS_C(D)$. 证毕.

根据上述两个定理, 我们可以得到一个基于正区域的直接求核算法.

算法 1 信息系统决策表相对核计算方法.

输入: 信息系统决策表

$S = \langle U, R, V, f \rangle$, $R = C \cup D$ 是属性集合, $C = \{ a_i \mid i = 1, 2, \dots, m \}$ 和 $D = \{ d_i \mid i = 1, 2, \dots, n \}$ 分别称为条件属性集和决策属性集;

输出: 信息系统决策表相对核 $Core$;

Step1 求 $POS_C(D)$;

Step2 令 $Core = \emptyset$;

Step3 对条件属性集 C 中的每个属性 a_i , 如果 $POS_{(C - \{ a_i \})}(D) \neq POS_C(D)$ 则 $Core = Core \cup \{ a_i \}$.

Step4 结束, 集合 $Core$ 为输出.

4 属性相对约简算法

在决策表中不同的属性可能具有不同的重要性, 为找出某些属性的重要性, 通常的方法是从决策表中去掉这个属性,

考察没有该属性后分类的变化情况.若去掉该属性相应变化较大,则说明该属性比较重要;反之,说明该属性不是太重要,即重要性较低.

这里我们应用正区域作为属性重要性的启发式信息,把 $pos = POS_C(D) - POS_{(C-\{r\})}(D)$ 的大小作为属性重要性的判断条件.下面给出两个基于正区域的决策表约简算法.

算法 2 是以核集为基础,逐步选择比较重要的属性加入该集合,直到满足条件 $POS_{Reduct}(D) = POS_C(D)$, $Reduct$ 即为约简.算法 3 是把整个条件属性集 C 作为一个约简,利用正区域的启发式信息逐步将该集合中不必要的属性约去,但仍满足上述条件,保证得到的属性集合 $Reduct$ 为约简.

算法 2 基于核的相对约简算法.

输入:信息系统决策表

$S = \langle U, R, V, f \rangle$, $R = C \cup D$ 是属性集合, $C = \{a_i | i = 1, 2, \dots, m\}$ 和 $D = \{d_i | i = 1, 2, \dots, n\}$ 分别称为条件属性集和决策属性集;

输出:信息系统决策表相对约简 $Reduct$;

Step 1 计算决策属性 D 对于条件属性 C 的正区域 $POS_C(D)$;

Step 2 计算条件属性 C 相对于决策属性 D 的核属性集 $Core_D(C)$;并令 $Reduct = Core_D(C)$; $Rem = C - Core_D(C)$.

Step 3 若 $Reduct = \Phi$,直接转下一步;

若 $Reduct \neq \Phi$,计算 $POS_{Reduct}(D)$.

如果 $POS_{Reduct}(D) = POS_C(D)$,则终止, $Reduct$ 为约简.否则转下一步;

Step 4 从 Rem 中选择属性 a_i ,使下式的值最大:

$$pos = POS_{Rem}(D) - POS_{(Rem - \{a_i\})}(D)$$

$$Reduct = Reduct \cup \{a_i\}$$

$$Rem = Rem - \{a_i\}$$

Step 5 若 $POS_{Reduct}(D) = POS_C(D)$,则终止,输出约简为 $Reduct$;否则转 Step 4.

算法 3 信息系统决策表相对约简计算方法.

输入:信息系统决策表

$S = \langle U, R, V, f \rangle$, $R = C \cup D$ 是属性集合, $C = \{a_i | i = 1, 2, \dots, m\}$ 和 $D = \{d_i | i = 1, 2, \dots, n\}$ 分别称为条件属性集和决策属性集;

输出:信息系统决策表相对约简 $Reduct$;

Step 1 计算决策属性 D 对于条件属性 C 的正区域 $POS_C(D)$;

Step 2 对每个属性 a_i 计算

$$pos = POS_C(D) - POS_{(C-\{a_i\})}(D);$$

Step 3 令 $Reduct = C$;将属性 a_i 按 pos 从小到大的顺序排列,对每个 a_i 执行操作:

若 $POS_{(Reduct - \{a_i\})}(D) = POS_C(D)$,则属性 a_i 应约简, $Reduct = Reduct - \{a_i\}$;

否则 a_i 不能被约简, $Reduct$ 不变;

Step 4 结束.

由粗糙集理论知道,任何决策表的相对核都是唯一的,而

且包含在所有的相对约简之中,算法 2 把相对核作为约简算法的起点,逐步增加对决策分类能力较大的属性,直到满足由相对约简定义的条件,所以基本可以保证得到最小的约简;算法 3 则是以条件属性全体为基础,在保证对决策表分类不变的前提下,逐步消去对决策分类能力较小的属性,算法简单,只需对各个属性扫描一遍即可.

5 实例分析

为了验证上述算法的有效性,本节选择了一个已知核与约简的 Wong-Ziarko 决策表^[12]和 UCI 机器学习数据库^[13]中的三个决策表进行计算,实验结果如表 1 所示.

表 1 求核与约简算法结果

决策表	实例数	条件属性数	算法 1	算法 2	算法 3
Wong-Ziarko decision table	21	9	{D, I}	{A, D, E, I}	{B, D, F, G, I}
BUPA liver disorders	345	6	{}	{A, B, E}	{C, D, E}
Glass Identification	214	10	{}	{A}	{A}
Ionosphere Database	351	34	{}	{4, 18, 24}	{30, 33, 34}

实验结果表明,算法 2 和 3 对于属性相对约简是有效的,并且大多数情况下可以得到最小约简.对于 Wong-Ziarko 决策表,算法 3 虽然没有得到最小约简,但也得到了次优约简.

6 结论

决策表核的确定和属性约简算法是粗糙集理论研究的焦点问题,本文在深入理解粗糙集基本概念的基础上,发现了正区域的一些有用性质,在此基础上给出直接利用正区域的信息求属性核,并给出了两个求属性相对约简的算法,实验结果表明了算法的有效性.由于利用正区域作为属性约简的启发式信息,也仅是从等价关系分类能力的角度对属性重要性进行粗略度量,并不能严格分辨出各个属性的重要程度.因此,如何利用粗糙集理论知识来度量属性重要性以便可以简单的求取最小约简还有待进一步的研究.

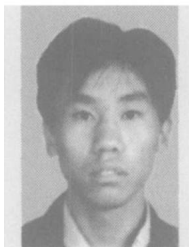
参考文献:

- [1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11(5): 341 - 356.
- [2] Pawlak Z. A rough set view on Bayes' theorem[J]. International Journal of Intelligent Systems, 2003, 18(5): 487 - 498.
- [3] Tay F, et al. Fault diagnosis based on rough set theory[J]. Engineering Applications of Artificial Intelligence, 2003, 16(1): 39 - 43.
- [4] Zhang T F. Dynamic system modeling based on rough sets and RBF neural networks[A]. Proc of the 5th World Congress on Intelligent Control and Automation[C]. Hangzhou, 2004. 185 - 189.
- [5] 徐捷等. 基于粗糙集理论的动态目标识别及跟踪[J]. 电子学报, 2002, 30(4): 605 - 607.

Xu Jie, et al. Dynamic objects identifying and tracing based

- on rough set theory [J]. Acta Electronica Sinica, 2002, 30 (4): 605 - 607. (in Chinese)
- [6] 张文修,等. Rough 集理论与方法[M]. 北京:科学出版社,2001.
- [7] Hu X H, et al. Learning in relational databases: A rough set approach [J]. International Journal of Computational Intelligence, 1995, 11(2): 323 - 338.
- [8] 刘文军,等. 基于可辨识矩阵和逻辑运算的属性约简算法的改进[J]. 模式识别与人工智能, 2004, 17(1): 119 - 123.
- [9] 王国胤,等. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759 - 766.
- [10] Mi J S, et al. Approaches to knowledge reduction based on variable precision rough set model [J]. Information Sciences, 2004, 159(3): 255 - 272.
- [11] Zhang M, et al. A rough set approach to knowledge reduction based on inclusion degree and evidence reasoning theory [J]. Expert Systems, 2003, 20(5): 298 - 304.
- [12] Wong S K M, et al. On optimal decision rules in decision tables [J]. Bulletin of Polish Academy of Science, 1985, 33(11 - 12): 693 - 696.
- [13] <http://www.ics.uci.edu/mlearn/MLRePOStory.html> [DB/OL].

作者简介:



张腾飞 男, 1980 年出生于河南省, 博士研究生, 主要从事粗糙集和神经网络的研究.



肖健梅 女, 1962 年出生于辽宁省大连市, 教授, 主要从事智能控制, 智能信息处理等方面的研究.

王锡淮 男, 1961 年出生于江苏省淮安市, 博士, 博士生导师, 主要从事粗糙集理论, 复杂系统建模与控制等方面的研究.

E-mail: wxh@shmtu.edu.cn.