文章编号: 1001-0920(2003)04-0449-04

粗糙集理论中的求核与约简

唐建国[1], 谭明术[2]
(1. 重庆三峡学院 电子工程系, 重庆 万州 404000; 2. 重庆三峡学院 计算机科学系, 重庆 万州 404000)

摘 要: 约简与核是粗糙集理论的两个重要概念, 而直接由定义来计算约简与核是一个典型的NP 难题。发现了分辨矩阵的若干有用性质, 利用这些性质使粗糙集理论中的求核与约简问题得以解决。进而分别讨论了无决策信息系统的约简和有决策信息系统的约简问题。最后举例说明了所得结果的有效性。

关键词: 粗糙集; 分辨矩阵; 求核; 约简
中图分类号: TP202. 1 **文献标识码**: A

On finding core and reduction in rough set theory

TANG Jian-guo[1], *TAN Ming-shu*[2]
(1. Department of Electronic Engineering, Chongqing Three Gorges University, Wanzhou 404000, China;
2. Department of Computer Science, Chongqing Three Gorges University, Wanzhou 404000, China)

Abstract: Reduction and core are two important concepts in rough set theory, while computing reductions and core according to the definitions directly is a typical NP problem. A number of useful natures of the discernable matrix is discovered, and used to solve the NP problem. The problems of reducing systems with and without decision are discussed respectively. The effectiveness of the result obtained is demonstrated by an example.

Key words: Rough set; Discernable matrix; Finding core; Reduction

1 引 言

粗糙集(RS)理论是 20 世纪末发展起来的一种处理不精确、不一致和不完整等各种不完备信息的数学理论[1]。文献[2, 3]对其作了很好的综述, 文献[4]则是国内第一本介绍RS 理论的专著。粗糙集理论中的一个重要概念就是定义了信息系统的“核”, 它是信息系统中不可或缺的部分, 包含在该信息系统的任一简化形式之中。显然, 求核可作为所有约简的基础, 并提供了从信息系统中分析多余属性的途径。因此, “核”和“约简”被认为是RS 理论的精华[2]。

目前, 粗糙集理论并没有提供直接从信息系统中求取核的方法。根据核的定义来求核, 则须先求出该信息系统的所有约简, 再通过求交集得到核, 这样求得的核已失去了指导对信息系统约简的意义。对信息系统进行约简, 求解最小子集是粗糙集理论研究中的一个基本问题, 其计算的复杂性随着数据表的增大呈指数增长, 是一个典型的NP 难题[2]。

文献[5]提出一种分辨矩阵的方法, 可将计算量减半; 文献[6]则采用遗传算法来搜索比较优的约简。但从解决NP 问题的角度看, 都没有取得根本性的突破。本文的主要工作是在文献[5]介绍的分辨矩阵的基础上, 寻求一种直接求取信息系统核的方法, 从而使约简信息系统的工作得以简化。

收稿日期: 2002-04-15; **修回日期**: 2002-10-15。
基金项目: 重庆市教委科研基金资助项目(0109096)。
作者简介: 唐建国(1954—), 男, 重庆开县人, 教授, 硕士, 从事鲁棒控制、粗糙集理论等研究; 谭明术(1962—), 男, 重庆万州人, 副教授, 硕士, 从事计算组合数学等研究。

2 问题表述

对于一个信息系统S, 研究其论域U = {$U_1, U_2, \cdots, U_m$}, U_i 是研究对象。条件属性集 P = {$P_1, P_2, \cdots, P_n$}, P_i 是条件属性。信息系统的一般表达形式如表 1 所示。

表 1 信息系统的一般表达形式

U	P			
	P_1	P_2	⋯	P_n
U_1	p_{11}	p_{12}	⋯	p_{1n}
U_2	p_{21}	p_{22}	⋯	p_{2n}
⋮	⋮	⋮		⋮
U_m	p_{m1}	p_{m2}	⋯	p_{mn}

现根据条件属性集P, 给出RS 理论中的有关定义:

1) 若$U_i \neq U_j$, 则称U_i和U_j是在P 下可分辨的。

2) 若S 中所有U_i都是两两可分辨的, 则称S 是在 P 下可分辨的, 记为 ind(P)。

3) 若去掉P 中某个条件属性P_i后, S 仍是可分辨的, 即有 ind(P - P_i) = ind(P), 则称P_i是P 中可约简的。

4) 若P 中任一条件属性P_i都是不可约简的, 则称 P 为独立的。

注 1 P 为独立的意味着属性集 P 中的任一属性都是必不可少的, 它独立地构成一组表达系统分类的特征。

5) 对于属性子集A ⊆ P, 若满足 ind(A) = ind(P), 且A 是独立的, 则称A 为 P 的一个最小子集, 记为min(P)。

注 2 一个属性集 P 可能有多个最小子集。

6) 核的定义: P 中所有最小子集的交集称为 P 的核, 记为 $P_c = \cap \min_i(P)$。

P_c 是表征系统必不可少的重要属性集。显然, 按定义来求取P_c, 要先求出P 的所有最小子集, 而如何求出P 的最小子集, 还没有较为简便的方法。文献[5] 介绍了一种基于分辨矩阵的方法, 即先构造一个与信息系统有关的分辨矩阵, 再将 P 的每一个可能的子集与该分辨矩阵的每一个元素进行比较, 最终确定该子集是否为 P 的一个最小子集。这是一项很费时的工作。如果一个信息系统研究的对象数目为m, 属性数目为n, 则考察属性集 P 的一个子集是否为最小子集, 要进行$n \times m^2$次比较。n 个属性可构成2^n个子集, 这些子集都有可能是最小子集, 要求出所有最小子集, 理论上需要$2^n \times m^2$次基本操作[1]。这就是所谓的NP 难题。但是, 目前尚未见到比这更好的方法。

本文经过研究发现, 分辨矩阵有许多有用的性质, 可直接用于系统的求核与约简, 而不必对所有的子集进行检验。

在给出主要结果之前, 首先介绍分辨矩阵D 的概念: D 是一个$m \times m$ 矩阵, 其中的每一个元素D_{ij}都是P 的一个子集, 即有$D_{ij} \subseteq P$。D_{ij} 的具体定义如下

$$D_{ij} = \{d_{ij1}, d_{ij2}, \cdots, d_{ijn}\} \quad i, j = 1, 2, \cdots, m \tag{1}$$

其中d_{ijk}确定如下

$$d_{ijk} = \begin{cases} \varnothing, & p_{ik} = p_{jk}, \\ P_k, & p_{ik} \neq p_{jk}, \end{cases} \quad k = 1, 2, \cdots, n \tag{2}$$

注3 D 是一个主对角线为 ∅ 的对称矩阵。

注4 D 与属性集P 有关, 不同的属性集有不同的D。为了区别, 用D (P) 和D (Q) 分别表示对应不同的属性集 P 和Q 的分辨矩阵。

下面举例说明分辨矩阵的形成。对于表 2 所示的一个知识系统, 按上面介绍的方法, 可得到该系统的分辨矩阵如下

表 2 一个知识系统的数据

U	P				
	P_1	P_2	P_3	P_4	P_5
U_1	2	1	1	1	1
U_2	1	1	2	1	2
U_3	1	2	2	1	1
U_4	1	2	2	1	2
U_5	1	2	2	2	2
U_6	1	1	2	1	1

$$D(P) = \begin{bmatrix} 0 & P_1P_3P_5 & P_1P_2P_3 & P_1P_2P_3P_5 & P_1P_2P_3P_4P_5 & P_1P_3 \\ & 0 & P_2P_5 & P_2 & P_2P_4 & P_5 \\ & & 0 & P_5 & P_4P_5 & P_2 \\ & & & 0 & P_4 & P_2P_5 \\ & & & & 0 & P_2P_4P_5 \\ & & & & & 0 \end{bmatrix} \tag{3}$$

3 主要结果

为了便于证明后面的结果, 首先给出两个引理:

引理 1 若存在某个$D_{ij}(P) = \varnothing, i, j \in [1, m], i \neq j$, 则U 中的两个对象$U_i$和$U_j$由属性集P 不

可分辨。

证明　由$D_{ij}(P)$的定义可知,$D_{ij}(P)=\varnothing$意味着U_i和U_j对于P中各个属性的取值都相同,即有$U_i=U_j$,所以由属性集P不能分辨U_i和U_j。□

引理 2　若属性集$Q=P-P_k$,则有

$$D_{ij}(Q)=\begin{cases}D_{ij}(P)-P_k, & P_k\in D_{ij}(P)\\ D_{ij}(P), & P_k\notin D_{ij}(P)\end{cases}$$

证明　因为Q是P去掉属性P_k后剩下的子集,$D_{ij}(Q)$和$D_{ij}(P)$的差别也仅由此产生,而$D_{ij}(Q)$中肯定不含P_k,所以当$P_k\notin D_{ij}(P)$时,有$D_{ij}(Q)=D_{ij}(P)$;而当$P_k\in D_{ij}(P)$时,去掉$D_{ij}(P)$中的P_k就是$D_{ij}(Q)$。□

3.1　核集的求取

这里介绍如何直接由分辨矩阵来求取系统的核集。不失一般性,假定系统S对于属性集P是可分辨的,则有如下结果:

定理 1　P中任一属性$P_k\in P_c$,充要条件为:D(P)中至少存在一个$D_{ij}(P)$,满足$D_{ij}(P)=\{P_k\}$。

证明　必要性:若$P_k\in P_c$,则意味着P_k是不可约简的。令$Q=P-P_k$,则有$\mathrm{ind}(Q)\neq\mathrm{ind}(P)$。由引理1知,系统S对于属性集Q的分辨矩阵中至少存在一个元素$D_{ij}(Q)$,满足$D_{ij}(Q)=\varnothing$,$D_{ij}(P)\neq\varnothing$。由引理2知,系统S对于属性集P的分辨矩阵中对应的元素$D_{ij}(P)$,满足$D_{ij}(P)=D_{ij}(Q)+P_k=\{P_k\}$。必要性得证。

充分性:若D中至少存在一个$D_{ij}(P)$,满足$D_{ij}(P)=\{P_k\}$,则由引理2知,对于Q的分辨矩阵,有$D_{ij}(Q)=D_{ij}(P)-P_k=\varnothing$,这说明$P_k$是P中不可约简的,即$P_k\in P_c$。充分性得证。□

定理1提供了一种直接求核集P_c的方法。例如对于表2的信息系统,由其分辨矩阵(3)和定理1,很容易得到P的核集$P_c=\{P_2,P_4,P_5\}$。

3.2　无决策系统的约简

无决策系统指的是属性集只有条件属性而没有决策属性的系统。上述举例就是一个无决策的系统。这种系统的约简就是求取最小属性集。前已求出P的核集P_c,则子集$B=(P-P_c)$中的属性都是可约简的。但是B中的所有属性通常不能同时约简,除非P只有唯一的一个最小子集,即$P_c=\min(P)$。这只是一种特殊情况。一般说,P有若干个最小子集。

定理 2　设系统S对于属性集P是可分辨的,$B_1\subseteq B$,要使$\mathrm{ind}(P-B_1)=\mathrm{ind}(P)$成立,充要条件为:对所有的$D_{ij}(P)$,$i,j=1,2,\cdots,n$,$i\neq j$,均有$D_{ij}(P)\not\subseteq B_1$。

证明　充分性:因为$B_1\subseteq B$,所以B_1中的任一元素都不属于P_c,即都是可约简的。若定理2条件成立,则意味着将B_1中包含的属性同时约简后,仍能保证$D_{ij}(P-B_1)\neq\varnothing$,对所有$i,j=1,2,\cdots,n$,$i\neq j$成立。充分性得证。

必要性:设B_1中属性是可同时约简的,且至少存在一个$D_{ij}^{*}(P)\subseteq B_1$,则当$B_1$中属性同时约简后,必有$D_{ij}^{*}(P-B_1)=\varnothing$。这说明$B_1$中属性是不可同时约简的,此假设不成立。必要性得证。□

应用定理2来考查表2的知识系统。经计算知,对于$B=\{P_1,P_3\}$,因为$D_{26}=\{P_1,P_3\}=B$,所以P_1和P_3是不能同时约简的。进而可知P存在两个最小属性子集:$\{P_1,P_2,P_4,P_5\}$和$\{P_2,P_3,P_4,P_5\}$。显然,求最小属性子集的工作得以大大简化。

3.3　有决策系统的简化

有决策系统与无决策系统的区别在于:有决策系统的属性集可分为两部分,一部分是条件属性,另一部分是决策属性;而无决策系统则没有决策属性。对于简化的目的,二者也有明显的不同:无决策系统的简化是为了得到最小属性集和核集;有决策系统的简化则是为了得到最简的决策规则。无决策系统简化的理论依据是可分辨性准则;有决策系统的简化依据是协调性原则。

定义 1(协调性定义)　对于决策系统中的两个对象,如果满足如下两个条件之一:1)其条件属性的取值至少有一个属性不同;2)有相同的条件属性取值时,其决策属性的取值是相同的。则称这两个对象是协调的;否则称为不协调的。如果系统中任何一对对象都是协调的,则称该系统是协调的。

有决策系统的简化只是针对条件属性集,而决策属性集是不需简化的。因此,在简化这类系统的过程中,为方便起见,可将不同的决策属性组合用编号表示。这样,决策属性集便可用一个一维编号变量来代替。不失一般性,假设系统是协调且可分辨的,并设P是条件属性集,e是决策编号变量。

(P,e)为一协调算法。若$P_i\in P$,满足((P-P_i),e)仍是协调的,则称P_i是P中可约简的;否则称P_i是不可约简的。

定义 2　如果所有的条件属性$P_i\in P$都是决策算法(P,e)中不可约简的,则决策算法(P,e)是独立的。如果条件属性子集$A\subseteq P$,且决策算法(A,e)是独立且协调的,则称A为决策算法(P,e)中P的简化。决策算法(A,e)称为决策算法(P,e)的简化。

引理3 若至少存在一个$D_{ij}(P,e)\in D(P,e)$，满足$D_{ij}(P,e)=\{e\}$，则决策算法(P,e)是不协调的。

证明 $D_{ij}(P,e)=\{e\}$，表明U_i和U_j的条件属性取值是相同的，但决策属性取值不同。由协调性定义知，U_i和U_j是不协调的。□

定理3 任一条件属性$P_k\in P$是决策算法(P,e)中不可约简的，充要条件为：至少存在一个$D_{ij}(P,e)\in D(P,e)$，满足$D_{ij}(P,e)=\{P_k,e\}$。

证明 必要性：如果条件属性$P_k\in P$是决策算法(P,e)中不可约简的，则意味着至少存在一个$D_{ij}(P,e)\in D(P,e)$，满足$D_{ij}(P,e)-P_k=D_{ij}((P-P_k),e)=\{e\}$，即$D_{ij}(P,e)=\{e\}+P_k=\{P_k,e\}$。必要性得证。

充分性：若存在一个$D_{ij}(P,e)\in D(P,e)$，满足$D_{ij}(P,e)=\{P_k,e\}$，如果要约简P_k，则有$D_{ij}((P-P_k),e)=\{e\}$。由引理3知，系统变成不协调的，所以P_k是不可约简的。充分性得证。□

算法(P,e)可能有若干个不同的简化，所有简化的交集称为算法(P,e)的核。核是所有不可约简的属性之集合，记作$core(P,e)$。

设$B=\{P,e\}-core(P,e)$，即B是算法(P,e)中所有可约简的属性之集合，但B通常不是可同时全部约简的。如果B可以同时全部约简，则说明算法(P,e)只有一个简化，就是$core(P,e)$。

定理4 对于协调算法(P,e)，$B_1\subseteq B$，要使算法$((P-B_1),e)$仍是协调的，充要条件为：对任何含有e的$D_{ij}(P,e)$，均满足$\{D_{ij}(P,e)-e\}\not\subset B_1$。

证明略。

4 举例说明

对于表3所示的一个有决策的信息系统，其分辨矩阵为

表3 一个有决策系统的数据

U	P			
	P_1	P_2	P_3	e
U_1	1	3	2	1
U_2	2	1	1	3
U_3	2	1	2	2
U_4	1	2	2	1
U_5	1	2	1	2

$$D(P,e)=\begin{bmatrix} 0 & P_1P_2P_3e & P_1P_2e & P_2 & P_2P_3e \\ & 0 & P_3e & P_1P_2P_3e & P_1P_2e \\ & & 0 & P_1P_2e & P_1P_2P_3 \\ & & & 0 & P_3e \\ & & & & 0 \end{bmatrix} \tag{5}$$

由以上分辨矩阵可知，满足定理3的元素有$\{P_3,e\}$，即系统的核集为$\{P_3,e\}$，P_3是条件属性集中不可约简的属性，$B=\{P_1,P_2\}$。那么P_1和P_2是否可同时约简呢？考察可知，$D(P,e)$中至少有一个元素，例如$D_{13}=\{P_1,P_2,e\}$，使得$\{(D_{13}-e)\}=\{P_1,P_2\}=B$。由定理4知，$P_1$和$P_2$是不能同时约简的。因此，该系统有两个简化形式，即$(\{P_1,P_3\},e)$和$(\{P_2,P_3\},e)$。

5 结 语

文献[5]在介绍分辨矩阵时，只看到了它的对称性，并将其用于约简信息系统，使工作量减少了将近一半。然而，对于一个NP难题的解决，这是远远不够的。本文给出的若干结果，使得系统的求核与约简大为简化，当系统的分辨矩阵构成后，通过对分辨矩阵简单的观察即可完成。故可认为RS理论中求核与约简这一NP难题基本得到解决。

参考文献(References):

[1] Pawlak Z. Rough sets theory and its applications to data analysis[J]. *Cybernetics and Systems*, 1998, 29: 661-668.

[2] 韩祯祥, 张琦, 文福拴. 粗糙集理论及其应用综述[J]. 控制理论与应用, 1999, 16(2): 153-157.
(Han Zhenxiang, Zhang Qi, Wen Fushuan. A survey on rough set theory and its application[J]. *Control Theory and Application*, 1999, 16(2): 153-157.)

[3] 张文修, 吴伟志. 粗糙集理论介绍和研究综述[J]. 模糊系统与数学, 2000, 14(4): 1-12.
(Zhang Wenxiu, Wu Weizhi. An introduction and a survey for the studies of rough set theory[J]. *Fuzzy Systems and Mathematics*, 2000, 14(4): 1-12.)

[4] 曾黄麟. 粗糙集理论及其应用(修订版)[M]. 重庆: 重庆大学出版社, 1998.

[5] Lin M. Software system for intelligent data processing and discovering based on the fuzzy-rough sets theory[D]. San Diego: San Diego State University, 1995.

[6] Jakub W. Finding minimal reducts using genetic algorithm[R]. Warsaw: Warsaw University of Technology, 1995.