文章编号: 1001-0920(2003)03-0313-04

一种基于粗糙集的近似质量求取属性约简的决策算法

徐德友, 胡寿松
(南京航空航天大学 自动化学院, 江苏 南京 210016)

摘 要: 提出一种基于粗糙集的近似质量求取属性约简的算法。该算法以集合近似的质量为迭代准则, 以所有条件属性为初始约简集合, 通过逐步缩减来求取约简, 保证了所求取的约简对问题的分类能力不会减弱。同时给出了该算法的时间复杂度分析,并举例验证了所提出算法的有效性和实用性。

关键词: 粗糙集; 决策表; 集合近似; 约简; 近似质量
中图分类号: TP18 文献标识码: A

Decision algorithm for finding reduct based on approximation quality of rough set

XU De-you, HU Shou-song
(College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: The rough set theory is studied, and an algorithm for finding attribute-oriented reduct based on approximation quality of rough set is presented. With all the condition attributes as the initial reduct, this algorithm takes the approximation quality of rough set as the iterative criterion to assure that the classification ability of the resulted reduct does not decline. The time complexity of the algorithm is analyzed and an example is investigated to verify this algorithm. The results show this algorithm can find the attribute-oriented reduct effectively with less computational effort.

Key words: Rough set; Decision table; Set approximation; Reduct; Approximation quality

1 引 言

粗糙集理论是波兰科学家 Pawlak 于 1982 年提出的一种数学理论[1], 主要用于数据分析, 尤其对不精确和不确定的数据进行分析。该理论提出的核、约简和上下近似等概念, 不仅为信息科学和认知科学提供了新的科学逻辑和研究方法, 而且为智能信息处理提供了有效的处理技术。粗糙集理论的一大特点就是它仅利用数据本身所提供的信息, 不需要任何附加信息或先验知识。例如证据理论中的基本概率赋值、模糊集理论中的隶属度和统计学中的概率分布等。而这些信息有时并不容易得到, 因此, 粗糙集理论相对于许多其他处理不确定知识的方法, 具有更强的客观性。

现已证明 RS 理论中求取全部最小约简的过程是NP-HARD 问题[2]。常用的约简求取方法有基于差别矩阵和差别函数的逻辑化简求取[3], 以及基于信息系统分解的最小约简求取[4]。然而这些方法存在计算复杂、效率低的问题, 在实际应用中遇到大数据时将无法应用。

本文提出一种求取知识属性约简的迭代算法。该算法以粗糙集中集合近似的质量为迭代准则, 每次约简后的近似质量不应小于前次的近似质量。

收稿日期: 2002-01-04; 修回日期: 2002-03-18。
基金项目: 国家自然科学基金重点项目(60234010); 航空科学基金资助项目(02E52025)。
作者简介: 徐德友(1974—), 男, 安徽合肥人, 博士生, 从事智能控制、故障诊断和粗糙集信息分析研究; 胡寿松(1937—), 男, 江苏南京人, 教授, 博士生导师, 从事智能控制、大系统理论和故障诊断与自修复控制研究。

2 粗糙集的基本概念和集合近似质量的定义

2.1 粗糙集中的基本概念

2.1.1 知识、划分与等价的关系

在 RS 理念中, 知识被视为一种对对象进行分类的能力, 可表述如下: 设 $U \neq \varnothing$ 是感兴趣的对象组成的有限集合, 称为论域, 任何子集 $X \subseteq U$ 称为 U 中的一个概念, 则 U 中一族概念称为关于 U 的知识。一个划分定义为 $C = \{X_1, X_2, \cdots, X_n\}$, 使得 $X_i \subseteq U, X_i \neq \varnothing, X_i \cap X_j = \varnothing$, 对 $i \neq j (i, j = 1, 2, \cdots, n)$ 且 $\bigcup_{i=1}^{n} X_i = U$。$X_i$ 称为划分 C 的一个等价类。

2.1.2 决 策 表

在 RS 理论中, 假定现实世界中的信息是用一张表表示的, 称为信息表。本文研究一类特殊的信息表, 称为决策表。决策表是如下形式的 4 元组

$$S = (U, Q, V, f) \tag{1}$$

其中: U 是论域; $Q = \mathrm{CON} \cup \mathrm{DEC}$, CON 是条件属性集合, DEC 是决策属性集合; V 是 Q 中属性值的集合; $f: U \times Q \to V$ 表示一个信息函数, 它指定 U 中每一对象的属性值。对于 $x \in U$ 及 $a \in Q$, 令 $f_x(a) = f(x, a)$。对于 $P \subseteq Q$ 且 $P = \{P_1, P_2, \cdots, P_m\}$, 有如下表示方法:

1) $\overline{p}_x(P)$ 表示 P 到 V_P 的映射, V_P 表示 P 中属性值的集合, 即

$$\overline{p}_x: P \to V_P \tag{2}$$

2) 当 U 是由 P 定义时, 其等价关系表示为

$$\widetilde{P} = \{(x, y) \mid x, y \in U \text{ 且 } \overline{p}_x(P) = \overline{p}_y(P)\} \tag{3}$$

3) 由 $\widetilde{P}$ 所定义且包含 x 的等价类表示为

$$[x]_{\widetilde{P}} = \{y \mid \overline{p}_x(P) = \overline{p}_y(P)\} \tag{4}$$

4) 由 $\widetilde{P}$ 所定义的所有等价类表示为

$$U/\widetilde{P} = \{[x]_{\widetilde{P}} \mid x \in U\} \tag{5}$$

2.2 集合的近似

对于决策系统 $S = (U, \mathrm{CON} \cup \mathrm{DEC}, V, f)$, 设 $X = [x]_{\widetilde{\mathrm{DEC}}}$ 是 U 中的一个概念, 即 $X \subseteq U$。可用包含在 CON 中的信息构造两个集合来逼近 X, 这两个集合分别称为 X 的 CON–下近似和 CON–上近似, 用 $\underline{\mathrm{CON}}X$ 和 $\overline{\mathrm{CON}}X$ 表示, 即

$$\begin{cases} \underline{\mathrm{CON}}X = \{x \in U \mid [x]_{\widetilde{\mathrm{CON}}} \subseteq X\} \\ \overline{\mathrm{CON}}X = \{x \in U \mid [x]_{\widetilde{\mathrm{CON}}} \cap X \neq \varnothing\} \end{cases} \tag{6}$$

集合 $\mathrm{BN}_{\mathrm{CON}}(X) = \mathrm{CON}X - \underline{\mathrm{CON}}X$ 称为 X 的 CON–边界区, 如果集合 $\mathrm{BN}_{\mathrm{CON}}(X)$ 非空, 则称集合 X 为粗糙集。

下面给出本文集合(概念) 近似的近似质量定义:

定义 1 对于 $U/\mathrm{DEC} = \{X_1, X_2, \cdots, X_n\}$, X_i 为第 i 个概念, 即决策表中第 i 个决策类(概念), 其近似质量为

$$\alpha_{\mathrm{CON}}(X_i) = \left| \underline{\mathrm{CON}}X_i \right| \Big\backslash \left| X_i \right| \tag{7}$$

则全部决策类的近似质量定义为

$$\alpha_{\mathrm{CON}}(U/\mathrm{DEC}) = \frac{1}{|U|} \sum_{i=1}^{n} |X_i| \alpha_{\mathrm{CON}}(X_i) = \frac{1}{|U|} \sum_{i=1}^{n} \left| \underline{\mathrm{CON}}X_i \right| \tag{8}$$

$\alpha_{\mathrm{CON}}(U/\mathrm{DEC})$ 表征用条件属性集合 CON 中的信息来近似 U/DEC 的近似质量。

3 基于集合近似质量的约简求取算法

对于决策表 S, 设 $U/\mathrm{DEC} = \{X_1, X_2, \cdots, X_n\}$, 则 U/DEC 的 CON 正域定义为

$$\mathrm{POS}_{\mathrm{CON}}(\mathrm{DEC}) = \bigcup_{X_i \in U/\widetilde{\mathrm{DEC}}} \underline{\mathrm{CON}}X_i \tag{9}$$

正域包含着基于条件属性所得的等价类能够归入基于决策属性所得的等价类的所有对象集合。设 $a \in \mathrm{CON}$, 若有 $\mathrm{POS}_{\mathrm{CON}}(\mathrm{DEC}) = \mathrm{POS}_{\mathrm{CON}-\{a\}}(\mathrm{DEC})$, 则称 a 为 CON 中 DEC 可省略。当 CON 中每个元素都不为 CON 中 DEC 可省略时, 称 CON 为 DEC 独立。当 $\mathrm{CON}' = \mathrm{CON} - \mathrm{CON}^*$ 为 DEC 独立, 且 CON^* 中的所有元素都是 DEC 可省略时, 称 CON' 为 CON 的 DEC 相对约简。从分类角度看, 相对简约就是用一种分类来表达另一种分类必不可少的属性集合。用属性集合 CON' 中的信息来近似 U/DEC 的近似质量与用属性集合 CON 中的信息来近似 U/DEC 的近似质量应是相同的。本文在给出表述上述事实之前, 首先给出如下引理:

引理 1[5] (集合的容斥原理) 对任意 n 个有限集 $A_1, A_2, \cdots, A_n$, 有如下等式

$$\left| \bigcup A_i \right| = |A_1 \cup A_2 \cup \cdots \cup A_n| = \sum |A_i| - \sum |A_i \cap A_j| + \cdots + (-1)^{n-1} |A_1 \cap A_2 \cap \cdots \cap A_n| \tag{10}$$

成立, 对任意 $i \neq j$, 若 $A_i \cap A_j = \varnothing$, 则有

$$\left| \bigcup A_i \right| = \sum |A_i| \tag{11}$$

定理 1 对于决策表 $S = (U, \mathrm{CON}/\mathrm{DEC}, V, f)$, 设 $U/\mathrm{DEC} = \{X_1, X_2, \cdots, X_n\}$, 如果 CON 为

CON 的 DEC 相对约简,则有

$$\alpha_{CON}(U/\ DEC) = \frac{1}{|U|}\sum_{i=1}^{n}|\ \underline{CON}X_i| = \alpha_{CON'}(U/\ DEC) = \frac{1}{|U|}\sum_{i=1}^{n}|\ \underline{CON'}X_i| \quad (12)$$

即约简后的近似质量保持不变。

证明　因为 CON′为 CON 的 DEC 相对约简,所以有 $POS_{CON}(DEC) = POS_{CON'}(DEC)$,即

$$\bigcup_{X_i\in U/\widetilde{DEC}}\underline{CON}X_i = \bigcup_{X_i\in U/\widetilde{DEC}}\underline{CON'}X_i \quad (13)$$

由 $U/\ DEC = \{X_1, X_2, \cdots, X_n\}$ 知 $X_i \cap X_j = \varnothing$,因此根据集合下近似的定义知

$$\begin{cases}\underline{CON}X_i \cap \underline{CON}X_j = \varnothing\\ \underline{CON'}X_i \cap \underline{CON'}X_j = \varnothing\end{cases} \quad (14)$$

由引理 1 及式(13)和(14)知

$$\sum_{i=1}^{n}|\ \underline{CON}X_i| = \sum_{i=1}^{n}|\ \underline{CON'}X_i|$$

故可推得

$$\alpha_{CON}(U/\ DEC) = \alpha_{CON'}(U/\ DEC) \quad \square$$

求取全部相对约简是 NP-HARD 问题[2]。随着问题规模的不断扩大,计算的时间复杂度和空间复杂度将极度扩大,直至超出计算机所能承受的范围。本文基于定理 1 给出一种基于近似质量进行约简求取的算法。该算法以所有的条件属性作为初始约简集合,以集合的近似质量不变为前提,逐步缩减求取约简,可在很短的时间内找出属性约简。

算法 **1**　求取约简

1) 初始化: CON′= CON//CON 是条件属性集;

2) 计算 α= $\alpha_{CON'}(U/\ DEC)$// 初始近似质量;

3) 令 Found = false//Found 表示是否找到约简;

4) While|CON| > 1 and Found = false

5)　for each $c \in CON$

6)　　计算 $\alpha_{CON'-\{c\}}(U/\ DEC)$

7)　　if($\alpha_{CON'-\{c\}}(U/\ DEC) = \alpha$)

8)　　Found = false

9)　　CON′= CON′- {c}

10)　　exit for loop

11)　　else

12)　　Found = true

13)　endif

14)　endfor

15) return CON′// 得到约简

为了分析算法 1 的时间复杂度,下面给出求取 $\alpha_{CON}(U/\ DEC)$ 的算法:

算法 **2**　计算近似质量

1) $U/\ DEC = \{X_1, X_2, \cdots, X_n\}$

　$U/\ CON = \{R_1, R_2, \cdots, R_k\}$

2) for($i = 0; i < n; i++$)

3) for($j = 0; j < k; j++$)

4) if($X_i \supseteq R_j$)

5) $\underline{CON}X_i = \underline{CON}X_i \cup R_j$

6) 计算 $\alpha_{CON}(U/\ DEC) = \frac{1}{|U|}\sum_{i=1}^{n}|\ \underline{CON}X_i|$

对于算法 2,第 1) 步时间复杂度分别为 $O(mn)$ 和 $O(smk)$[6],m 为 U 中对象个数,s 为条件属性集合 CON 中元素个数。因为 $n, k \le m$,因此第 2) 步～第 5) 步的时间复杂度为 $O(m^2)$。整个算法 2 的时间复杂度为 $O(mn) + O(smk) + O(m^2) \in O(mn + smk + m^2)$。由于 $k \le m$,有 $O(smk) \in O(sm^2)$,故有如下命题成立:

命题 **1**　算法 2 的时间复杂度为 $O(sm^2)$,其中: s 为条件属性的个数,m 为 U 中对象个数。

对于算法 1,由命题 1 知第 2) 步和第 6) 步的时间复杂度均为 $J = O(sm^2)$。由于算法 1 中的外循环和内循环均最多循环 s 步,故第 4) 步～第 14) 步的时间复杂度为 $O(s^2J)$。整个算法 1 的时间复杂度为 $O(sm^2) + O(s^2J)$,因此有如下命题成立:

命题 **2**　算法 1 的时间复杂度为 $O(s^3m^2)$,其中: s 为条件属性的个数,m 为 U 中对象个数。

4　应用示例

表 1 所示的决策系统,论域 U 由 8 个对象组成,其条件属性集 CON 包括 4 个属性,即{学历,经验,

表 **1**　一个两类决策系统

	学历	经验	是否说法语	证明	是否接收
x_1	MBA	中	是	优秀	是
x_2	MBA	低	是	一般	否
x_3	MCE	低	是	良好	否
x_4	MSc	高	是	一般	是
x_5	MSc	中	是	一般	否
x_6	MSc	高	是	优秀	是
x_7	MBA	高	否	良好	是
x_8	MCE	低	否	优秀	否

是否说法语, 证明}。$V_{学历}$ = {MBA, MCE, MSc}, MBA 为工商管理硕士, MCE 为土木工程学硕士, MSc 为理学硕士; $V_{经验}$ = {低, 中, 高}; $V_{法语}$ = {是, 否}; $V_{证明}$ = {优秀, 良好, 一般}。决策属性为是否接收 V_d = {接收, 拒绝}。

对于表 1 所示的决策系统, 有

$$U/\ \mathrm{CON} = \{\{x_1\},\{x_2\},\{x_3\},\{x_4\},\{x_5\},\{x_6\},\{x_7\},\{x_8\}\}$$

$$U/\ \mathrm{DEC} = \{\{x_1,x_4,x_6,x_7\},\{x_2,x_3,x_5,x_8\}\} = \{X_1,X_2\}$$

$$\underline{\mathrm{CON}}(X_1) = \{x_1,x_4,x_6,x_7\}$$

$$\underline{\mathrm{CON}}(X_2) = \{x_2,x_3,x_5,x_8\}$$

初始近似质量为 1。

利用算法 1 的迭代算法对表 1 进行化简, 可得到约简集{经验, 证明}, 其对 DEC 的近似质量仍为 1。可以验证, {经验, 证明} 是决定一个对象属于哪一类的最小集。对于表 1 所示的决策系统, {经验, 证明} 和{学历, 经验, 是否说法语, 证明} 对 U 的分类能力是相同的。因此可以删除属性{学历} 和{是否说法语}, 从而得到简化的决策系统。需要指出的是, 最小约简集不是唯一的。

5 结 语

本文以粗糙集中集合近似的近似质量为基础, 证明了约简后的近似质量保持不变, 并基于此给出了基于近似质量求取约简的算法, 分析了该算法的时间复杂度。该算法避免了求取差别矩阵和差别函数的复杂过程, 简单明了, 可用较少的计算时间求出约简。

参考文献(References):

[1] Pawlak Z. Rough sets[J]. *Int J of Information and Computer Science*, 1982, 11(5): 341-356.

[2] Ziarko W. The discovery, analysis and representation of data dependencies in databases[A]. *Knowledge Discovery in Databases*[C]. Cambrige: AAAI/MIT Press, 1990. 213-228.

[3] Skowron A. The discernibility matrices and functions in information systems[A]. *Intelligent Decision Support-Handbook of Advances and Applications of the Rough Set Theory*[C]. Dordrecht, Boston, London: Kluwer Academic Publishers, 1992. 311-369.

[4] Kryszkiewicz M, Rybinski H. Finding reducts in composed information systems[A]. *Proc From Int Workshop on Rough Sets and Knowledge Discovery RSKD'93*[C]. Banff, 1993. 259-268.

[5] 姜泽渠. 离散数学[M]. 重庆: 重庆大学出版社, 1997.

[6] Sever H. Knowledge structuring for database mining and text retrieval using past optimal queries[D]. The University of Southwestern Louisiana, 1995.

(上接第 312 页)

参考文献(References):

[1] Wang H, Daley S. Actuator fault diagnosis: An adaptive observer-based technique[J]. *IEEE Trans on Automatic Control*, 1996, 41(7): 1073-1078.

[2] Christoher Edwards, Sarah K Spurgeon, Ron J Patton. Sliding mode observers for fault detection and isolation [J]. *Automatica*, 2000, 36: 541-553.

[3] 周川, 吴晓蓓, 陈庆伟, 等. 一类基于非线性状态观测器的鲁棒故障检测[J]. 信息与控制, 1998, 29(4): 297-303.
(Zhou Chuan, Wu Xiaopei, Chen Qingwei, et al. Robust fault detection based on a class of nonlinear state observer [J]. *Information and Control*, 1998, 29(4): 297-303.)

[4] Anshul A Jain, Michael A Demetrion. A neural network based actuator fault detection and diagnostic scheme for a sacra manipulator[A]. *Proc of the 15th IEEE Int Symposium on Intelligent Control* [C]. Greece, 2000. 297-302.

[5] 赵众, 顾幸生, 蒋慰孙. 基于 WaveARX 神经网络的间歇过程工况监测[J]. 控制与决策, 1998, 13(2): 151-155.
(Zhao Zhong, Gu Xingsheng, Jiang Weisun. Batch process monitoring based on WaveARX neural network [J]. *Control and Decision*, 1998, 13(2): 151-155.)

[6] Michael Demetriou. Robust adaptive techniques for sensor fault detection and diagnosis[A]. *Proc of the 37th IEEE Conf on Design and Control*[C]. Florida, 1998. 1143-1148.

[7] 李庆国, 冯玉珠, 佟绍成, 等. 基于神经网络的非线性系统故障检测及容错控制方法[J]. 信息与控制, 1998, 27(6): 440-445.
(Li Qingguo, Feng Yuzhu, Tong Shaocheng, et al. Fault detection and fault tolerant control of nonlinear systems using neural networks[J]. *Information and Control*, 1998, 27(6): 440-445.)

[8] Demetriou Michael A, Polycarpou M M. Incipient fault diagnosis of dynamical systems using online approximators [J]. *IEEE Trans on Automatic Control*, 1998, 43(11): 1612-1617.