文章编号：1000 - 8152(2003)01 - 0093 - 04

一种基于粗集理论的动态近似规则挖掘推理方法

张仕念，刘文奇
（昆明理工大学 系统科学与应用数学系，云南 昆明 650093）

摘要：提出一种基于粗集理论的，把属性的重要性和属性值的出现频率综合起来进行规则推理的方法.分析了"激活一个，否则离开"原则的优缺点，指出在近似推理中，大前提中的规则数量应该可变.给出一种根据推理过程中规则的出现频率决定其是否保留，从而实现规则数量的动态变化的方法，证明了动态变化过程中规则的数量不会无限增加.实例表明此法是比较有效的.

关键词：粗集；规则推理；决策表；判别矩阵
中图分类号：TP18 **文献标识码**：A

Dynamics approximate rule mining inference approach based on rough set theory

ZHANG Shi-nian, LIU Wen-qi
(Department of System Science and Applied Mathematics, Kunming University of Science and Technology, Yunnan Kunming 650093, China)

Abstract: Based on rough set theory, a rule inference approach of integrating the importance of attributes with the emergence frequency of attribute value is proposed. After the analysis of the principle of "fall one or leave", it is pointed out that the number of rules in rule inference would be variable. A method is suggested to decide whether a rule is needed or not by its emergence frequency in the inference process, so the number of the rules is dynamic. The dynamic rule is proved not to increase in the process. The results of examples show the efficiency of the approach.

Key words: rough set; rule inference; decision table; judgment matrix

1 引言(Introduction)

1982 年，Z.Pawlak 为了研究软计算问题引入了粗集(rough set)理论[1]，它在处理含噪声、不完整、不精确的信息方面具有强大的功能.文献[2,3]研究了利用粗集进行规则提取.在获取规则库的过程中，规则的数量太多，推理的速度会变慢；规则的数量少时，推理速度快，但经常会出现激不活的现象.本文集中讨论此问题及基于粗集的近似规则推理.

2 理论准备(Preparation of theory)

设 U 为论域，R 为 U 上的等价关系族，设 $U/R = \{X_1, X_2, \cdots, X_n\}$(为方便起见，记 $U/R = U/\mathrm{ind}(R)$，$\mathrm{ind}(R)$ 为 R 的不可分辨关系) 表示 R 对 U 的划分，$\mathrm{card}(U/R)$ 表示分类 U/R 的等价类个数.

设 U 为论域，R 为 U 上的等价关系族，对 U 中的任一概念 X，记

$$R_-(X) = \bigcup \{Y \in U/R \mid Y \subseteq X\},$$

$$R^-(X) = \bigcap \{Y \in U/R \mid Y \cap X \neq \emptyset\},$$

称 $R_-(X)$ 和 $R^-(X)$ 为 X 的 R 下近似和 R 上近似.

知识表达系统可表示为 $S = \langle U, C, D, V, f\rangle$，其中 U 为论域，$A = C \cup D$ 为属性集，C 和 D 分别称为条件属性集和决策属性集，$V = \bigcup_{a \in A} V_a$ 是属性值的集合，V_a 表示属性 $a \in A$ 的属性值范围，$f: U \times A \to V$ 是一个信息函数，它指定 U 中每一个对象 x 的属性值.知识表达系统对应的决策表记为 $T = (U, A, C, D)$.

在 $S = \langle U, C, D, V, f\rangle$ 中，$\forall x \in U$，$\forall a \in C \cup D$，函数 $d_x: A \to V$，$d_x(a) = a(x)$ 称为决策表的决策规划.d_x 对于 C, D 的约束分别记为 $d_x \mid_C$ 和 $d_x \mid_D$，称为 d_x 的条件和决策.若 $S = \langle U, C, D, V, f\rangle$ 是相容的，即对决策规划 d_x，有 $[x]_C \subseteq [x]_D$，$\forall r \in C$，若 $[x]_{C-\{r\}} \not\subseteq [x]_D$，则 r 为 d_x 的核值属性，r 为 d_x 中不可省略的；若 $[x]_{C-\{r\}} \subseteq [x]_D$，则 r 为 d_x 中可省略的，r 不为 d_x 的核值属性.若 r 为 d_x 的核值属性，则称 $f(x, r)$ 为决策规划 d_x 的核值，r 不为 d_x 的核值属性，称 $f(x, r)$ 为决策规划 d_x 的冗余值，表中只有

收稿日期：2000 - 05 - 09；收修改稿日期：2001 - 12 - 29.
基金项目：云南教委基金(20033)资助项目.

核值的决策表称为核值表.

设 P,Q 为 U 中的等价关系,记

$$\mathrm{pos}_P(Q)=\bigcup_{X\in U/Q}P_-(X),$$

称为 Q 的 P 正域.

在 $S=\langle U,C,D,V,f\rangle$ 中,若对应的决策表是相容的,对任意的条件属性 $a\in C$, 其重要性

$$\alpha_a=1-\frac{\mathrm{card}(\mathrm{pos}_{(C-\{a\})}(D))}{\mathrm{card}(U)}.$$

3 问题的分析及近似规则推理方法(The analysis of the problem and approximate rule inference approach)

规则推理的一般模式如下:

已知大前提

f_{11} 且 f_{12} 且 $\cdots$ 且 f_{1m}, 则 d_1,

f_{i1} 且 f_{i2} 且 $\cdots$ 且 f_{im}, 则 d_1,

$f_{(i+1)1}$ 且 $f_{(i+1)2}$ 且 $\cdots$ 且 $f_{(i+1)m}$, 则 d_2,

f_{n1} 且 f_{n2} 且 $\cdots$ 且 f_{nm}, 则 d_h.

且给定小前提

$$c_1 \text{ 且 } c_2 \text{ 且 } \cdots \text{ 且 } c_m,$$

求 d^*.

将 $f_{11},f_{21},\cdots,f_{n1}$ 看成条件属性 a_1 的属性值,$\cdots,f_{1m}f_{2m},\cdots,f_{nm}$ 看成条件属性 a_m 的属性值,$d_1,d_2,\cdots,d_h(d_i\neq d_j,i\neq j,1\leqslant i,j\leqslant h)$ 看成决策属性 d 的属性值,令

$$C=\{a_1,a_2,\cdots,a_m\},D=\{d\},$$

$$A=C\cup D,U=\{1,2,\cdots,n,n+1\},$$

$$B=\begin{pmatrix}\dfrac{\mathrm{card}\{i\in[d_1]\mid f(i,a_1)=c_1\}}{\mathrm{card}[d_1]} & \cdots & \dfrac{\mathrm{card}\{i\in[d_1]\mid f(i,a_m)=c_m\}}{\mathrm{card}[d_1]}\\ \cdots & \ddots & \cdots\\ \dfrac{\mathrm{card}\{i\in[d_h]\mid f(i,a_1)=c_1\}}{\mathrm{card}[d_h]} & \cdots & \dfrac{\mathrm{card}\{i\in[d_h]\mid f(i,a_m)=c_m\}}{\mathrm{card}[d_h]}\end{pmatrix}.$$

定义 2 在不完全信息系统 $S=\langle U,C,D,V,f\rangle$ 中,c_1 且 c_2 且 $\cdots$ 且 c_m 的判别向量 E 定义为 $E=AE_0$,A 为判别矩阵,E_0 为 $(\alpha_{a_1},\alpha_{a_2},\cdots,\alpha_{a_m})^{\mathrm{T}}$ 的归一化向量.$\alpha_{a_1},\alpha_{a_2},\cdots,\alpha_{a_m}$ 为属性 $a_1,\cdots,a_m$ 的重要性.

显然,c_1 且 c_2 且 $\cdots$ 且 c_m 的判别向量 $E=(e_1,e_2,\cdots,e_h)^{\mathrm{T}}$ 中的任一 $e_i(1\leqslant i\leqslant h)$ 表示:c_1 在 $\{f(j,a_1)\mid f(j,a_1)\in V_{a_1},j\in[d_i]\}$ 中出现的频率,c_2 在 $\{f(j,a_2)\mid f(j,a_2)\in V_{a_2},j\in[d_i]\}$ 中出现的频率,$\cdots,c_m$ 在 $\{f(j,a_m)\mid f(j,a_m)\in V_{a_m},j\in[d_i]\}$ 中出现的频率的权重分别为 E_0 的坐标的常权综合.反映了条件属性 $a_1,a_2,\cdots,a_m$ 的属性值分别为 $c_1,c_2,\cdots,c_m$ 时,决策属性 d 的属性值为 d_i 的可能性的大小.由于 d 将 U 分为 h 个类 $[d_1],[d_2],\cdots,[d_h]$,因此,作者认为小前提 c_1 且 c_2 且 $\cdots$ 且 c_m 的输出结果为 d_l,其中 l 满足 $e_l=\max\{e_1,e_2,\cdots,e_h\}$.

判别矩阵 A 定义为 $h\times m$ 阶矩阵 B 的列归一化矩阵(若某列全为 0,则不归一化),列归一化的目的是使同一条件属性 a_i,c_i 在 $\{f(j,a_l)\mid f(j,a_l)\in V_{a_l},j\in[d_l]\}(1\leqslant l\leqslant h)$ 中出现的可能性的大小均匀地发挥作用.当 $e_l=e_k=\max\{e_1,e_2,\cdots,e_h\}$ 且 $k\neq l$ 时,该方法失效,也可认为小前提 c_1 且 c_2 且 $\cdots$ 且 c_m 的输出结果为 d_l 或 d_k.为了提高推理的可信程度,可设常数 β,当 $e_1,e_2,\cdots,e_h$ 中的最大值比次大值大 β,即 $e_l=\max\{e_1,e_2,\cdots,e_h\}>\max(\{e_1,e_2,$

$\cdots, e_h\} \setminus \{e_l\}) + \beta$ 时,小前提 c_1 且 c_2 且 $\cdots$ 且 c_m 的输出结果为 d_l,否则失效.

当决策表中有两个决策属性值 $f(i,d), f(j,d)$ 未知时,可保留其中之一 $f(i,d)$,将 $f(j,d)$ 所在的对象及其对应的所有条件属性值去掉,此时,就转化为只有一个决策属性值未知的情况.有多个决策属性值未知时,可类似地求出.

注 1　判别矩阵随着决策类在决策表中出现的秩序的不同而变化,但并不影响推理结果.例如,在表 1中,将决策值为I和决策值为Ⅳ的决策规则交换,判别矩阵的第 1 行和第 5 行也交换了,决策类在决策表中的变化并不影响推理结果.

注 2　决策表中含有核值的规则不宜推理.例如,在表 1 中,将规则 5 删除,决策值Ⅳ就被去掉,对小前提 4 且 3 且 2 进行推理时就不可能推出对应的决策值是Ⅳ.

表 1　水泥窑生产决策表

Table 1　Decision table of cement production

U	a	b	c	d
1	3	2	1	I
2	2	2	1	Ⅱ
3	3	2	3	Ⅲ
4	4	2	3	Ⅲ
5	4	3	3	Ⅳ

4　大前提中规则的变化(The variation of rules in rule inference)

王国俊在文献[4]中提出了"激活一个,否则离开"的原则进行模糊推理,将这一原则用于规则推理时,其优点是推理速度快,但由于大前提中的规则是事先确定的、不变的,这一原则也有其弱点:一方面,若小前提 c_1 且 c_2 且 $\cdots$ 且 c_m 是经常出现的,而大前提中又没有规则:c_1 且 c_2 且 $\cdots$ 且 c_m 则 $f(i,d)$,此时激不活,不能有效地进行控制,反映了大前提中规则的不足,因此,应将这种规则加入到大前提中去;另一方苗,大前提中的某些规则可能长时间激不活,甚至根本就不起作用,此时,这些规则的存在减慢了推理速度,应将其从大前提中去掉.因此,大前提中的规则应当是可变的.

如何调整大前提中的规则,使其既能充分发挥作用,规则数量又不太多,能满足推理速度的需要是一个必须解决的问题.事实上,设置两个常数 k_0, k_1(k_0, k_1 为自然数,也可设置为时间常数 t_1, t_2),一方面,输入的小前提 c_1 且 c_2 且 $\cdots$ 且 c_m 激不活时,若自 c_1 且 c_2 且 $\cdots$ 且 c_m 输入之时的 k_0 次输入之内 c_1 且 c_2 且 $\cdots$ 且 c_m 再次出现,说明 c_1 且 c_2 且 $\cdots$ 且 c_m 是经常出现的,可利用本文的推理方法推出其对应的决策属性值,形成一条新的规则,并添加到大前提中去;若自 c_1 且 c_2 且 $\cdots$ 且 c_m 输入的 k_0 次输入之内 c_1 且 c_2 且 $\cdots$ 且 c_m 不再出现,说明 c_1 且 c_2 且 $\cdots$ 且 c_m 是不经常出现的,此时离开.另一方面,经过 k_1 次输入小前提,大前提中都未激活的且不含核值规则,说明这些规则在推理过程中所起的作用不大,将其从大前提中去掉.为了保证大前提中适当的规则数量,允许新增加的规则被删除.

大前提中规则的动态变化,具有一定的自适应性,但同时具有遗忘性,为了防止正确的规则因多次未激活而被删除,规定保留大前提中含有核值的规则.

问题是往大前提中增加规则,大前提中的规则是否会无限地增加呢? 以下命题给出否定的答案.

命题　大前提中至多增加 $\left[\frac{k_1-1}{k_0}\right]+1$ 条规则.

证　不妨设大前提中有 l 条规则,其中 m 条含有核值,当 $k_0 \leqslant k_1$ 时,k_0 次输入小前提,大前提中至多增加 1 条规则,不妨记为 d_x,k_0+1 次输入,大前提中至多增加 $\left[\frac{k_0+1}{k_0}\right]$([$x$] 表示 x 向下取整,下同)条规则,此时 d_x 有 1 次未激活,$\cdots$,k_1-1 次输入小前提,大前提中至多增加 $\left[\frac{k_1-1}{k_0}\right]$ 条规则,此时 d_x 有 k_1-k_0-1 次未激活,k_1 次输入时,若 $l > m$,大前提中不含核值的规则将被删除,规则数减少,若 $l = m$,大前提中至多增加 $\left[\frac{k_1}{k_0}\right]$ 条规则,此时 d_x 有 k_1-k_0 次未激活,如此下去,k_1+k_0-1 次输入小前提,大前提中至多增加 $\left[\frac{k_1-1}{k_0}\right]+1$ 条规则,此时 d_x 有 k_1-1 次未激活,若 k_1+k_0 次输入小前提仍未激活,此时 d_x 有 k_1 次未激活,将被删除,规则数减少,因此,至多增加 $\left[\frac{k_1-1}{k_0}\right]+1$ 条规则.同时,$k_0 > k_1$ 时,大前提中至多增加 1 条规则.总之,大前提中至多增加 $\left[\frac{k_1-1}{k_0}\right]+1$ 条规则.

5　仿真算例(Simulation examples)

表 1 是一决策表,表 2 是表 1 的核值表,设有小前提:3 且 2 且 2,2 且 2 且 2,4 且 3 且 2,并设 $k_0=1$, $k_1=3$.对 3 且 2 且 2,由 $k_0=1$,要推出其对应的决策属性值,计算得 $\alpha_a=0.4$, $\alpha_b=0.4$, $\alpha_c=0.4$,故 $E_0=(0.33,0.33,0.33)^{\mathrm{T}}$,由

$$B=\begin{pmatrix}1 & 1 & 0\\ 0 & 1 & 0\\ 0.5 & 1 & 0\\ 0 & 0 & 0\end{pmatrix},$$

故 $$A = \begin{pmatrix} 0.67 & 0.33 & 0 \\ 0 & 0.33 & 0 \\ 0.33 & 0.33 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

从而

$$E = \begin{pmatrix} 0.67 & 0.33 & 0 \\ 0 & 0.33 & 0 \\ 0.33 & 0.33 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.33 \\ 0.33 \\ 0.33 \end{pmatrix} = \begin{pmatrix} 0.33 \\ 0.11 \\ 0.22 \\ 0 \end{pmatrix}.$$

又由 $\max\{0.33, 0.11, 0.22, 0\} = 0.33 = e_1$，因此，其对应的决策属性值是 Ⅰ，将其加到决策表中去，此时有 6 条规则，各条件属性的重要性分别变为 $\alpha_a = 0.67, \alpha_b = 0.67, \alpha_c = 0.5$. 由于决策表中仍没有小前提“2 且 2 且 2”，再进行推理，得到对应的决策值是 Ⅱ，此时，大前提中有 7 条规则，各条件属性的重要性分别变为 $\alpha_a = 0.57, \alpha_b = 0.29, \alpha_c = 0.43$. 输入小前提“4 且 3 且 2” 仍激不活，原决策表中的所有决策规则均有 $k_1 = 3$ 次未激活，同核值表比较知 5 条规则均含有核值，不能从大前提中去掉，从而推出其对应的决策值是 Ⅳ，此时，决策规则变为 8 条. 若再输入小前提“c_1 且 c_2 且 c_3” 仍激不活，第一次加入的规则“3 且 2 且 2 则 Ⅰ”将被删除，因此大前提中至多增加 $\left[\frac{k_1 - 1}{k_0}\right] + 1 = 3$ 条规则.

表 2 表 1 的核值表

Table 2 The core table of table 1

U	a	b	c	d
1	3	—	—	Ⅰ
2	2	—	—	Ⅱ
3	—	—	3	Ⅲ
4	4	2	—	Ⅲ
5	—	3	—	Ⅳ

6 结语(Conclusion)

提出了一个基于粗集理论的近似规则推理方法. 分析了文献[4]所提出的“激活一个，否则离开”原则的优缺点，指出在规则推理中，大前提中的规则数量应该可变，并给出了一种具体的变化方法，实例表明此方法是比较有效的. 从理论上讲，根据不同的推理要求选择适当的 k_0, k_1，但实际应用效果有待检验.

参考文献(References):

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

[2] XIA Y J, LI S Y, XI Y G. A method of inducing decision rules based on rough set theory [J]. *Control and Decision*, 2001, 16(5): 577 - 580 (in Chinese).

[3] MA Z F, XING H C, ZHENG X M. Research on the uncertainty of rule acquisition from decision table [J]. *Control and Decision*, 2000, 15(6): 703 - 707(in Chinese).

[4] WANG G J. Triple I method and interval value for fuzzy reasoning [J]. *Science in China (Series E)*, 2000, 30(4): 331 - 340(in Chinese).

[5] ZENG H L. An intelligent expert system on rough sets [J]. *Engineering Science*, 2001, 3(2): 47 - 51 (in Chinese).

[6] ZENG H L. *Rough Sets Theory and Its Application* [M]. Chongqing: Chongqing University Press, 1998 (in Chinese).

[7] LIU W Q. *Fuzzy Sets' Representation Theory & Its Applications* [M]. Kunming: Yunnan Science & Technology Press, 1999 (in Chinese).

作者简介：

张仕念 (1976 —)，男，昆明理工大学硕士研究生. 研究方向为粗集理论与数据挖掘. Email: mecca@km169.net;

刘文奇 (1965 —)，男，昆明理工大学教授，硕士生导师. 研究方向为系统分析与综合决策等.