文章编号: 1001- 2486(2008) 05- 0094- 05

基于证据熵对不确定性度量的决策表约简*

宋立军, 胡 政, 杨拥民, 温熙森
(国防科学技术大学 机电工程与自动化学院, 湖南 长沙 410073)

摘 要: 知识约简是粗糙集理论的核心内容之一, 产生的粗糙决策规则往往具有一定的不确定性。在变精度粗糙集的基础上, 本文构造了符合证据理论框架的一组焦元, 利用基本概率分配函数计算了证据的总体信息熵, 度量了决策表的不确定性; 以该度量作为启发信息, 给出了决策表的启发式知识约简算法。计算实例表明了本文方法的有效性。

关键词: 变精度粗糙集; 不确定性度量; 证据熵; 知识约简
中图分类号: TP18 文献标识码: A

Decision Table Reduction Based on Evidence Entropy for Uncertainty Measures

SONG Li-jun, HU Zheng, YANG Yong-min, WEN Xi-sen
(College of Mechatronics Engineering and Automation, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Knowledge reduction is one of the important topics in the research on rough set theory, and rough decision rules are inevitably provided with uncertainty. In this paper, a family of focal sets is constructed within the framework of evidence theory on the basis of variable precision rough set theory. Accordingly, the function of basic probability assignment is defined, and then the total information entropy is calculated for evidence theory, namely the evidence entropy. Uncertainty measure for the decision table is determined by that entropy. Based on the measure, the heuristic algorithm is proposed for decision table reduction. Finally, the experimental results show the validity of the methodology.

Key words: variable precision rough set; uncertainty measure; evidence entropy; knowledge reduction

粗糙集理论由 Pawlak 于 1982 年提出, 并由 Ziarko 于 1993 年扩展为变精度粗糙集[1-2], 是处理不确定和不精确问题的一种新型数学工具, 是知识约简的一个重要手段。粗糙集的约简计算方法, 一般分为代数观点和信息论观点两大类[3-4], 信息论观点已被证明具有更为普遍的适用性[4]。信息论观点下的约简, 需要计算信息熵的等价性, 而这一般是基于 Shannon 熵定义的。

从不确定度问题的研究来看[5], 粗糙集对象的决策表必然具有一定的不确定性, 需要适当地度量。考虑证据理论与粗糙集均关注于对象的"分类", 二者的关系密切[6-7], 可以借助证据理论来对粗糙集进行刻画; 基于 Hartley 熵与 Shannon 熵共同定义的不确定性度量模型[8-9], 有效地表达了系统的总体不确定度, 可以更加完整地评估决策表在约简前后的信息等价性。

本文从粗糙集与证据理论之间的密切联系入手, 构造了基于变精度粗糙集的证据理论框架, 定义了相应的基本概率分配函数, 计算了表示总体不确定度的证据信息熵; 根据约简前后的等价性要求, 用不确定度为启发知识, 给出了有效的决策表约简算法, 并进行了算例验证。

1 变精度粗糙集

在粗糙集理论中, 决策系统 S 可表示为一张数据表 $< U, C \cup D >$, 其中 U 为论域, C、D 分别为条件属性和决策属性[6]。设 R 是属性子集, 则表示了一个不可分辨关系; 以 U/R 表示所有等价类的集

* 收稿日期: 2008- 02- 23
基金项目: 国家部委资助项目(413270303)
作者简介: 宋立军(1980—), 男, 博士生。

合, 则构成了对论域 U 的一个划分。

变精度粗糙集(Variable precision rough set, VPRS), 引入了一个"精度等级"参数 $\beta(0.5<\beta\leqslant 1)$, 是对 Pawlak 粗糙集的扩展[1,7] ($X\subseteq U$):

$$\underline{R}^{\beta}(X)=\bigcup\left\{E\in U/R\mid\frac{|E\cap X|}{|E|}\geqslant\beta\right\} \tag{1}$$

$$\overline{R}^{\beta}(X)=\bigcup\left\{E\in U/R\mid\frac{|E\cap X|}{|E|}>1-\beta\right\} \tag{2}$$

$$BN^{\beta}(X)=\overline{R}^{\beta}(X)-\underline{R}^{\beta}(X) \tag{3}$$

$$NEG^{\beta}(X)=U-\overline{R}^{\beta}(X) \tag{4}$$

式(1)~式(4)分别表达了集合 X 的下近似集、上近似集、边界域和负域。

可见, 在 VPRS 中, 体现的是一种"多数包含"关系。若取 $\beta=1$, 即退化为 Pawlak 粗糙集。

2 VPRS 构造的证据理论框架

2.1 空间划分的简单实例

图 1 是关于集合空间划分的一个简单例子。在图(a)中, 子集 A_1、A_2、A_3 相互重叠, 把全集 U 划分为四类区域, 已标识于图(b)中。区域 iv为单个子集的内部, 区域 ⑦为两个子集的交界, 区域 ④为三个子集的交界, 而区域 ⑤则为所有子集外部的公共区域。

不难得出结论: 所有的这四类区域, 其中的非空集合, 构成了对全空间的一个划分。

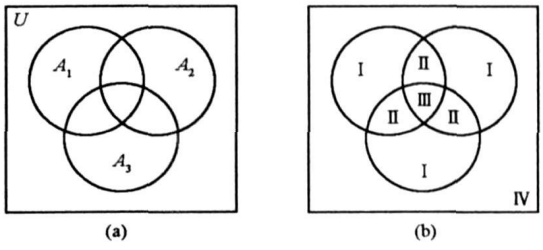

图 1 集合空间划分的简单例子
Fig. 1 Diagram of space division

2.2 基于 VPRS 对论域划分

不妨设决策属性 $D=\{d_1,\cdots,d_m\}$, 对 U 的划分为 $U/D=\{Y_1,\cdots,Y_m\}$, Y_i 是 d_i 唯一对应的等价类。为简便, 在不致引起混淆时, 约定在写法上不再对 Y_i 和 d_i 进行区分。

文献[7]提出了基于 d_i 的下近似集和边界域对 U 进行划分的研究方法。事实上, 经分析可知, 各下近似集之间是不存在重叠的, 相应的区域划分结果稍嫌繁杂。为此, 本文给出了一个新的研究思路, 是基于上近似集和负域来进行划分的。

对条件属性 $R\subseteq C$, d_i 的上近似集、负域分别由式(2)、式(4)确定, 并可定义集合如下:

$$I_i=\overline{R}^{\beta}(d_i)\cap\bigcap_{k\neq i}NEG^{\beta}(d_k) \tag{5}$$

可见, I_i 类似于图 1 中区域 iv, 是与 d_i 有关的集合的内部。相应地, 可定义如下集合:

$$I_0=\bigcap_i NEG^{\beta}(d_i) \tag{6}$$

则 I_0 类似于图 1 中的区域 ⑤。

类似于图 1 中区域⑦、④可给出边界域的相应表达式: ($\forall\theta\subseteq D$, $|\theta|>1$)

$$I_\theta=\left(\bigcap_{d_i\in\theta}\overline{R}^{\beta}(d_i)\right)\cap\left(\bigcap_{d_k\notin\theta}NEG^{\beta}(d_k)\right) \tag{7}$$

可见, I_θ 表达了 $d_i\in\theta$ 的上近似集的公共交界, 而 $\bigcup_{\theta\subseteq D}I_\theta$ 则表示了整个交界区域。

至此, 不难看出, 对于全部的集合 I_i、I_0 和 I_θ, 其中的那些非空集合构成了对论域 U 的一个完全划分。

2.3 证据理论的基本框架

根据上文研究内容, I_0 是与任何 d_i 都无关的区域; 对区域中的对象, 不能由决策属性 D 直接进行判别。为此, 需要相应地添加一个新的 d_0, 则扩充后的识别框架 Θ 定义为

$$\Theta=\{d_1,d_2,\cdots,d_m\}\cup\{d_0\} \tag{8}$$

结合式(5)~式(7),可统一表述如下:

$$I(\theta)=\begin{cases}\phi, & \theta=\phi\\ I_0, & \theta=\{d_0\}\\ I_i, & \theta=\{d_i\}\\ I_\theta, & \theta\subseteq D,|\theta|>1\end{cases} \tag{9}$$

定义映射关系如下: $m: 2^{\Theta}\rightarrow[0,1]$

$$m(\theta)=|I(\theta)|/|U| \tag{10}$$

根据论域 U 的划分关系,不难证明,这是一个基本概率分配(Basic probability assignment, BPA)函数,能够满足表达式: $m(\phi)=0$,且 $\sum_{\theta\subseteq\Theta}m(\theta)=1$。

进一步,由bpa函数可以确定信任函数 Bel 和似然函数 Pl,具体细节不再详述。

$$Bel(\theta)=\sum_{\varphi\subseteq\theta}m(\varphi),\quad Pl(\theta)=\sum_{\varphi|\theta\neq\phi}m(\varphi) \tag{11}$$

3 基于证据熵的知识约简

3.1 证据理论的不确定度

不确定度是一个度量不确定性测度函数所表示对象的不确定程度的模型,不同的不确定性理论用不同的不确定性测度描述。根据Klir的分类,证据理论描述了非特异性和随机性[5]。

基于经典信息论概念的推广,随机性引起的不确定度由Shannon熵表示,而非特异性则需要由Hartley熵表示。为了表示总体不确定度,Pal等采用了两类熵之和的研究思路[5,8]。

不失一般性,总体不确定度计算如下:

$$TU(m)=-\frac{1}{2}\sum_{\theta\subseteq\Theta}m(\theta)\ln\frac{m(\theta)}{|\theta|^2} \tag{12}$$

3.2 决策表的证据信息熵

根据上文研究内容,基于VPRS,可构造出决策表的证据理论框架,因此,考虑借助证据理论的总体不确定度模型,来对粗糙集决策表的不确定性进行适当度量。

对于识别框架 Θ,式(9)将论域 U 详细划分为 2^m 个独立区域,实际运算有所不便。为此,考虑进行简化,将全部交界区域视为一个整体,则 U 最终可划分为 $m+2$ 个区域,即:

$$U=\{I_1\}\cup\ldots\cup\{I_m\}\cup\{I_0\}\cup\bigcup_{\theta\subseteq D,|\theta|>1}I_\theta \tag{13}$$

至于全体边界区域的测度,可计算为:

$$m(\Delta)=1-m(d_0)-\sum_{i=1}^{m}m(d_i) \tag{14}$$

式中,Δ 代指整个边界区域。

根据式(12)和式(14),可以方便地计算出简化情况下的总体不确定度,并将其作为对决策表不确定性的度量。由于需计算证据理论框架的总体信息熵,则称之为证据信息熵,简称证据熵。

3.3 基于证据熵的约简算法

基于VPRS对论域 U 的划分,在一定程度上反映了决策属性 D 与条件属性 $R\subseteq C$ 之间的相互依赖关系。证据熵的计算,实质上是信息论观点下对决策表中分类不确定性的总体评估;证据熵是Hartley熵与Shannon熵的综合形式,比传统上只用单类熵表达的信息更为完整[4,8]。

为书写方便,将证据熵写为如下形式:

$$H(D|R)\triangleq TU(D,R,\beta)=TU(m) \tag{15}$$

对于属性 $r\in R$,可利用证据熵对其重要性进行适当评估,表达式如式(16)。若 $SGF(r)=0$,则该属性不影响分类的不确定性,被认为是不重要的,可以将其约简。

$$SGF(r)=H(D|R)-H(D|R-\{r\}) \tag{16}$$

以属性重要性作为启发信息, 可以给出一个启发式的决策表知识约简算法:

输入: 一个决策表 $S = < U, C \cup D >$

输出: 属性集 C 的一个相对约简 B

Step 1: 初始化 $B = \Phi$;

Step 2: 计算原始的证据熵 $H(D \mid C)$;

Step 3: 对所有 $r \in C$, 计算证据熵 $H(D \mid r)$, 并按照递减的顺序, 将各 r 依次放入 B 中;

Step 4: 对 B 中各属性 b, 依次操作:

Step 4.1: 由式(16), 计算 b 的重要性;

Step 4.2: 若 $SGF(b) = 0$, 则将 b 约简, 即 $B = B - \{b\}$; 否则, B 保持不变。

Step 5: 约简结束, 将 B 输出。

4 柴油机故障诊断实例

为便于对比分析, 这里采用了文献[10]中的实验数据进行验证计算。

利用直列 4135 柴油机进行故障诊断实验(四种工况: 正常, 进气阀开度过小, 进气阀开度过大, 排气阀开度过大), 检测并采样其振动信号, 一共产生 40 组数据样本。

实验中布置了三个测点: 测点 1、2 位于气缸盖上; 测点 3 位于气缸体表面, 对应活塞行程中点处。对每个测点的信号提取 6 个故障特征, 分别为: 频域波形复杂度 I_f、时域波形复杂度 I_r、非周期复杂度 σ、频谱中心频率 C_c、时间序列方差 D_x 和时间序列峭度 α_4。相应的计算公式如下:

$$I_f = \sum_{n=1}^{N/2} X(n) \ln X(n), \quad I_r = \sum_{i=1}^{p} \lambda_i \ln \lambda_i, \quad \sigma = \frac{p}{p-1} \cdot \frac{\sum_{i=2}^{p} \lambda_i^2}{\sum_{i=2}^{p} \lambda_i},$$

$$C_c = \sum_{m=1}^{N/2} \frac{2mX(m)}{N \sum_{n=1}^{N/2} X(n)}, \quad D_x = \frac{1}{l} \sum_{n=1}^{l} [X(n) - \bar{x}]^2, \quad \alpha_4 = \frac{1}{l} \sum_{n=1}^{l} [X(n)]^4 \tag{17}$$

式中各符号含义参见文献[10], 这里不再详述。

首先, 以 40 组数据样本作为论域 U, 以 18 个信号特征作为条件属性 C, 以四种实验工况作为决策属性 D, 构建了决策表; 至于条件属性, 则分别按照"测点 1 特征 1, …, 测点 1 特征 6, 测点 2 特征 1, …, 测点 3 特征 6"的顺序, 依次编号为"1"~"18"。限于篇幅, 不再对决策表进行详细介绍。

然后, 预设一个精度等级, 不妨取 $\beta = 0.9$; 需要说明, β 的取值应视实际情况而定[6, 10], 主要是 VPRS 对数据样本的容错能力需要。依据式(15)、式(12), 可计算原始的证据熵, $H(D \mid C) = 4.3275$。

接着, 需对各属性(特征)进行排序, 这是基于证据熵计算的, 实质上也是对各特征的重要性的一个初步比较。计算结果见表 1。

表 1 各特征(属性)的证据熵

Tab.1 Evidence Entropy for each property

$r1$	$r2$	$r3$	$r4$	$r5$	$r6$	$r7$	$r8$	$r9$
5.3219	5.3219	5.2376	5.3219	5.0501	5.3219	5.017	5.3219	5.3219
$r10$	$r11$	$r12$	$r13$	$r14$	$r15$	$r16$	$r17$	$r18$
5.3219	5.3219	5.0874	4.9610	5.3219	5.3219	5.3219	4.5737	5.0874

按照递减顺序, 各特征的排列为"1, 2, 4, 6, 8, 9, 10, 11, 14, 15, 16, 3, 12, 18, 5, 7, 13, 17"。

按照上面给出的排列顺序, 可依次计算各个特征的属性重要性, 即考虑决策表在约去该属性前后的证据熵变化, 且只保留重要性不为 0 的特征(属性)。作为约简结果, 可以验证, 属性子集{5, 17}的证据熵为 4.3275, 与原始的证据熵相同, 且{5, 17}的任一子集不再满足该条件。

至此, 可以得出结论: 测点 1 的时间序列方差, 以及测点 3 的时间序列方差, 能够有效判断四种实验工况。即是说, 在 0.9 的精度等级上, {5, 17} 可用于替代原特征集(含 18 个特征) 。

若取 β= 1, 约简结果也为“5, 17”。但如果取 β= 0.8, 约简结果则为“1, 12, 17”; 这意味着, 在 0.8 的精度等级上, {1, 12, 17} 与原特征集等效。至于文献[10] 中的结果, 则是基于 Shannon 信息熵计算得到的, 为“5, 17”。可见, 基于 VPRS 的证据熵方法, 具有更好的实用性和普适性。

需要说明: 对于一般的决策表而言, 冲突样本数据是在所难免的, 知识约简的信息熵表示更为适用; 基于 VPRS 进行研究, 可根据实际需要调整“精度等级”, 对于抑制数据中的噪声和不确定性有积极的作用。

5 结 论

(1) 基于变精度粗糙集, 本文分析了决策表中蕴含的证据理论框架, 定义了基本概率分配函数, 计算了用于表征总体不确定度的证据熵, 并以此作为判决依据, 提供了决策表知识约简的一种新的研究思路。

(2) 基于证据熵的决策表知识约简方法, 更全面地考虑了系统的非特异性和随机性, 在信息论观点下更完整地描述了系统的总体不确定度, 计算结果具有更好的可信赖性与适用能力。

参 考 文 献:

[1] Ziarko W. Variable Precision Rough Set Model[J]. Journal of Computer and System Sciences, 1993, 46(1): 39- 59.

[2] Beynon M J, Driffield N. An Illustration of Variable Precision Rough Sets Model: An Analysis of the Findings of the UK Monopolies and Mergers Commission[J]. Computers and Operations Research, 2005, 32(7): 1739- 1759.

[3] Pawlak Z. Rough Sets and Intelligent Data Analysis[J]. Information Sciences, 2002, 147(1- 4): 1- 12.

[4] Wang G Y. Algebra View and Information View of Rough Sets Theory[C]// Proceedings of SPIE, 2001, 4384: 200- 207.

[5] 陈理渊, 黄进. 不确定度问题研究情况综述[J]. 电路与系统学报, 2004, 9(3): 105- 111.

[6] 宋立军, 胡政, 杨拥民, 等. 基于证据理论与粗糙集集成推理策略的内燃机故障诊断[J]. 内燃机学报, 2007, 25(1): 90- 95.

[7] Marszałek-Paszek B, Paszek P. Evidence Theory and VPRS Model[J]. Electronic Notes in Theoretical Computer Science, 2003, 82(4): 153- 163.

[8] Pal N R, Bezdek J C, Hemasinha R. Uncertainty Measures for Evidential Reasoning, II: A new measure of total uncertainty[J]. International Journal of Approximate Reasoning, 1993, 8: 1- 16.

[9] Aczel J, Forte B, Ng C T. Why the Shannon and Hartley Entropies Are ‘Natural’. Advances in Applied Probability, 1974, 6(1): 131- 146.

[10] 王新峰, 邱静, 刘冠军. 基于特征相关性和冗余性分析的机械故障特征选择研究[J]. 中国机械工程, 2006, 17(4): 379- 382.