

基于粗糙集理论的多准则决策分析^{*}

Multicriteria Decision Analysis Based on Rough Sets

邓方安^{1,2} 刘三阳¹

(西安电子科技大学应用数学系 西安710071)¹ (陕西理工学院数学与计算机科学系 汉中723000)²

Abstract In this paper, the application of rough sets theory to multicriteria decision analysis is discussed. Firstly, some basic and crucial concepts of rough sets theory are introduced. Then, the fuzzy measure represents the payoff obtained by the coalition in a cooperative game; in a multicriteria decision problem, the fuzzy measure can be interpreted as the conjoint importance of the criteria from the coalition. Finally, the application of this method is illustrated with an example.

Keywords Rough sets, Fuzzy measure, Multicriteria decision analysis

1 前言

粗糙集理论(Rough Set, RS)是由波兰学者 Z. Pawlak 在 1982 年提出的一种刻画不完整性和不确定性的数学工具。粗糙集理论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。近二十年来,RS 理论已发展成为人工智能领域中一个较新的学术热点,引起了学术界科研人员的广泛关注。本文将对决策论中的 Shapley 值和 Banzhaf 值作为多准则决策分析中特殊准则的平均重要程度的度量指标,最后,用实例阐述了粗糙集理论在解决综合评判问题方面的应用。

2 粗糙集理论的基本概念

2.1 信息系统与不可分辨关系

设 $K=(U, R)$ 为一个信息系统,其中 $U=\{u_1, u_2, \dots, u_n\}$ 为有限非空的集合(论域), R 是 U 上的一族等价关系,是一个知识库(或有限属性集)。设 $R=\{r_1, r_2, \dots, r_k\}$, $\forall r \in R$ 是 U 上的一个属性,则在 U 上的元素关于 r 属性的等价类记为 $[x]_r, x \in U$ 。

设 $P \subseteq R, P \neq \emptyset, P=\{r_1, \dots, r_m\}$, P 的全体等价关系的交记为 $\cap P = \cap_{i=1}^m r_i$, 则 $\cap P$ 也是一个等价关系,记为 $\text{IND}(P)$, 称此为不可分辨关系。定义如下:

$$\text{IND}(P) = \{(x, y) \in U^2 : \forall a \in P, a(x) = a(y)\}$$

由这种关系把 U 的元素划分为不可分辨的元素类,其每一个元素必包含在某一子族 P 中。用 $U/\text{IND}(P)$ 表示等价关系 $\text{IND}(P)$ 的等价类族,它代表与 P 有关的知识,被称为知识库 K 中关于 U 的 P -基本知识。利用 U 上的任意特殊等价关系 $Q \in R$, 得到的 Q 等价类被称为知识 R 的 Q -基本范畴或概念。对于 $X \subseteq U$ 和一个等价关系 R , 我们用 U/R 来表示关系 R 的等价类。

2.2 粗糙集的下近似、上近似

对于任意一个等价关系 R 和任意一个子集 X , 定义 X 关于 R 的下近似和上近似分别为:

$$R_-(X) = \{x \in U | [x]_R \subseteq X\} = \bigcup \{Y \in U/R : Y \subseteq X\}$$

$$R_+(X) = \{x \in U | [x]_R \cap X \neq \emptyset\} = \bigcup \{Y \in U/R : Y \cap X \neq \emptyset\}$$

根据粗糙集的定义容易验证下列重要关系成立:

$$R_-(X) \subseteq X \subseteq R_+(X); \quad R_-(X) = U - R_+(U - X)$$

2.3 决策表、决策规则与决策表的化简

(1) 决策表 信息系统又可理解为信息表,一个信息表是一个四元组 $S=(U, Q, V, f)$ 。设 U 是对象的有限集, $Q=\{q_1, q_2, \dots, q_m\}$ 是属性的有限集, V_q 是属性 q 的值域, $V=\bigcup_{q \in Q} V_q$ 。一个函数 f 被称为信息函数,如果 $f: U \times Q \rightarrow V$ 是一个全函数且满足 $f(x, q) \in V_q, \forall q \in Q, x \in U$ 。因此, U 的每一个对象 x 都可以用一个向量 $\text{Des}_Q(x)=[f(x, q_1), f(x, q_2), \dots, f(x, q_m)]$ 来表示。

在一个信息表中,属性集合 Q 被划分为条件属性($C \neq \emptyset$)和决策属性($D \neq \emptyset$), $C \cup D = Q$, 且 $C \cap D = \emptyset$, 这样的表称为决策表。在条件属性独立的情况下,通过不可分辨关系 I_D 对决策表的约简可得到决策属性诱导出的一个 U 的划分,称这样的 D -基础集为决策类。

(2) 决策规则 每一个属性 $a \in R$ 是一个映射 $a: U \rightarrow V_a$, V_a 为 a 的值集。

对 $\forall x \in U$, 函数 $d_x: R \rightarrow V$, 使得 $d_x(a) = a(x)$, 称 d_x 为一条决策规则。其规则的语法可表示为: 如果 $f(x, q_1) = r_{q_1}^1$, 且 $f(x, q_2) = r_{q_2}^2$, 且 $\dots, f(x, q_p) = r_{q_p}^p$, 则 $x \in Y_1$ 或 Y_2 或 Y_j , 其中 $(r_{q_1}^1, r_{q_2}^2, \dots, r_{q_p}^p) \in V_{q_1} \times V_{q_2} \times \dots \times V_{q_p}$ 且 $\{q_1, q_2, \dots, q_p\} \subseteq C$, $Y_1, Y_2, \dots, Y_j$ 是某些分类的决策类。如果这个结果是单义的($k=1$), 则称这个规则是精确的, 否则, 称该规则是近似的或意义不明确的。

(3) 决策表的约简 设 $r \in R$, 如果 $\text{IND}(R) = \text{IND}(R - \{r\})$, 则称 r 是 R 中的一个可约简的属性, 否则称 r 是 R 的一个不可约简的属性。如果 $\forall r \in R$ 都不可约简, 则称集合 R 是独立的, 否则称集合 R 是相关的。如果 $Q \subseteq P$ 是独立的, 且 $\text{IND}(Q) = \text{IND}(P)$, 则称 Q 是 P 一个约简。所有约简的交集是属性(决策表)的核(core)。“约简”和“核”是粗糙集中很重要的两个概念。由于决策表中每一个对象都蕴涵着一条分类规则, 因此决策表又可以看作是一组逻辑规则的集合, 而约简决策表的过程就是抽取分类规则的过程。

3 模糊测度与粗糙集

3.1 粗糙近似的精度、质量及粗糙隶属函数

^{*}国家自然科学基金资助(6997206); 陕西省自然科学基金资助(2001SL08)。

(1) 近似精度 对于一个粗糙集 $X \subseteq U, X \neq \emptyset$, 我们可以利用属性 R 定义 X 的近似精度为: $\alpha_R(X) = |R_+(X)| / |R^+(X)|$, 这里 $|S|$ 表示一个有限集 S 的基数。显然, $0 \leq \alpha_R(X) \leq 1$; 如果 $\alpha_R(X) = 1$, 则 X 关于 R 是一个普通集; 如果 $\alpha_R(X) < 1$, X 关于 R 是一个粗糙(模糊)的集。

(2) 近似质量 对于一个粗糙集 $X \subseteq U, X \neq \emptyset$, 利用属性 R 可以定义 X 的另一个比值, 称为 X 关于 R 的近似质量:

$$\gamma_R(X) = |R_+(X)| / |X|$$

上式的数量 $\gamma_R(X)$ 表示利用属性 R 将对象正确分类的相对频率。对于 $\gamma_R(X)$ 容易得到:

$$0 \leq \alpha_R(X) \leq \gamma_R(X) \leq 1$$

$$\gamma_R(X) = 0 \Leftrightarrow \alpha_R(X) = 0$$

$$\gamma_R(X) = 1 \Leftrightarrow \alpha_R(X) = 1$$

注: 对于子集 $X \subseteq U$ 的近似的定义也可以推广到 U 的一个划分: $S = \{S_1, S_2, \dots, S_n\}$, 其中子集 $S_i, i = 1, 2, \dots, n$, 是 S 的不相交类。引入记号:

$$R_+ S = \{R_+ S_1, R_+ S_2, \dots, R_+ S_n\},$$

$$R^+ S = \{R^+ S_1, R^+ S_2, \dots, R^+ S_n\}$$

则系数 $\gamma_R(S) = \sum_{i=1}^n |R_+ S_i| / |U|$ 被称为利用属性 R 对 S 进行分类的近似质量。

从形式上看, 粗糙集分类的质量满足模糊测度的集函数的性质, 而模糊测度是建立联盟重要性模型的十分有用的工具。基于粗糙集的方法, 预测粗糙集分类质量的模糊测度已经广泛用于估计由每个属性提供的信息相关值和分析属性之间的相互影响。

下面我们先介绍对策论的 N 人合作对策的 Shapley 值及 Banzhaf 值的定义, 并把它们作为度量多属性决策的个体属性之间的相互影响的指标。

设 $N = \{1, 2, \dots, n\}$ 是一个有限集, 它的元素可以看成对策论中的局中人, 或在一个多准则决策中的准则, 或在数据表中的属性等。设 $P(N)$ 表示 N 的幂集, 在 N 上的一个模糊测度是满足如下公理的集函数 $\mu: P(N) \rightarrow [0, 1]$:

$$(1) \mu(\emptyset) = 0, \mu(N) = 1;$$

$$(2) A \subseteq B \Rightarrow \mu(A) \leq \mu(B), \forall A, B \in P(N).$$

在下面研究过程中, 我们将公理(1)中 $\mu(N) = 1$ 放宽为 $\mu(N) \leq 1$ 。

设 $A \subseteq N$, 按照 Shapley 和 Banzhaf 在对策论中处理合作对策的方法, 函数 $\mu(A)$ 被称为特征函数, 它表示由联盟 A 得到的支付。在对策论中已经引入了一些作为合作对策特殊解的指标, 即就是用于描述某局中人 i 单独对整个联盟的特殊的加权平均贡献, 其中最重要的是 Shapley 和 Banzhaf 分别提出的 Shapley 值和 Banzhaf 值, 其定义分别为:

$$\varphi_S(i) = \sum_{K \subseteq N - \{i\}} \frac{(n - |K| - 1)! |K|!}{n!} [\mu(K \cup \{i\}) - \mu(K)];$$

$$\varphi_B(i) = \frac{1}{2^{n-1}} \sum_{K \subseteq N - \{i\}} [\mu(K \cup \{i\}) - \mu(K)].$$

1992年, Murofushi 在进行多属性决策分析时, 把模糊测度用于表征准则的联合重要程度, 提出将 Shapley 值和 Banzhaf 值用于表示特殊准则的平均重要性。除了特殊准则之外, 其他欲测量的指标必然受到“准则对”之间的相互影响。1993年, Murofushi 和 Soneda 提出用下面定义的指标来测量相互影响: 对于 $i, j \in N$,

$$I_{MS}(i, j) = \sum_{K \subseteq N - \{i, j\}} \frac{(n - |K| - 2)! |K|!}{(n - 1)!} \times [\mu(K \cup \{i, j\}) - \mu(K \cup \{i\}) - \mu(K \cup \{j\}) + \mu(K)]$$

1996年, Roubens 定义了另一测量相互影响的指标: 对于 $i, j \in N$,

$$I_R(i, j) = \frac{1}{2^{n-2}} \sum_{K \subseteq N - \{i, j\}} [\mu(K \cup \{i, j\}) - \mu(K \cup \{i\}) - \mu(K \cup \{j\}) + \mu(K)]$$

上述相互影响的指标 $I_{MS}(i, j)$ 和 $I_R(i, j)$ 可以理解为把元素 i, j 同时放入每一个可能的联盟中产生的一种特殊的平均增加值。对于 $I_{MS}(i, j)$ 和 $I_R(i, j)$ 应注意下列情况:

① $I_{MS}(i, j) > 0$ ($I_R(i, j) > 0$): i, j 是互补的;

② $I_{MS}(i, j) < 0$ ($I_R(i, j) < 0$): i, j 是可以替代的;

③ $I_{MS}(i, j) = 0$ ($I_R(i, j) = 0$): i, j 是独立的。

对于相互影响的指标的定义也可以从“无序对 $i, j \in N$ ”扩充到任意子集 $A \subseteq N, A \neq \emptyset$ 。

1996年, Grabischba 把 A 的元素的 Shapley 相互影响指标定义为:

$$I_S(A) = \sum_{A \subseteq N - A} \frac{(n - |K| - |A|)! |K|!}{(n - |A| + 1)!} \sum_{L \subseteq A} (-1)^{|A| - |L|} \mu(L \cup K)$$

1996年, Roubens 把 A 的元素的 Banzhaf 相互影响指标定义为:

$$I_B(A) = \frac{1}{2^{n - |A|}} \sum_{K \subseteq N - A} \sum_{L \subseteq A} (-1)^{|A| - |L|} \mu(K \cup L)$$

除此之外, 还有一个用于解释模糊测度的概念, 它是 μ 的 Möbius 表示, 也就是集函数 $m: P(N) \rightarrow R$, 定义如下:

$$m(A) = \sum_{B \subseteq A} (-1)^{|A - B|} \mu(B), \forall A \subseteq N$$

注: 上述的三个指标对研究所考虑的属性间的信息依赖及选择最好的约简是十分有用的。

4 例子^[1]

粗糙集方法的一个重要优点是它能够处理对象被条件属性不可区分但可由决策属性区分决策问题。除此之外, 粗糙集还能提供有关特殊属性和在决策类中它们的近似子集的有用信息, 而且为生成包含关联属性的决策规则作好了准备。下面的例子将用于解释上面所提到的有关概念。

学生情况统计表

学生编号	数学	物理	语文	综合评定成绩
1	好	好	坏	好
2	中等	坏	坏	坏
3	中等	坏	坏	好
4	坏	坏	坏	坏
5	中等	好	好	坏
6	好	坏	好	好

上述表中各分量为: $U = \{1, 2, 3, 4, 5, 6\}, Q = \{A_1, A_2, A_3, A_4\}, V_1 = \{\text{坏, 中等, 好}\}, V_2 = \{\text{好, 坏}\}, V_3 = \{\text{好, 坏}\}, V_4 = \{\text{好, 坏}\}$, 其信息函数 $f(x, q)$ 的取值: $f(1, A_3) = \text{坏}, f(5, A_3) = \text{好}$, 等。

在上述表格中, 每个学生关于属性 A_1, A_2, A_3, A_4 各不相同, 从而可以利用这四个属性来区分六个学生。基于这四个属性的不可分辨关系是:

$$I_Q = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$$

也就是说, 不存在两个不同的学生 $(x, y) \in I_Q$ 。但就学生 2, 3来讲, 由于关于属性 A_1, A_2, A_3 有相同的值, 因此他们关于 $P = \{A_1, A_2, A_3\}$ 是不可分辨的, 且基于 P 的不可分辨关系是 $I_P = \{(1, 1), (2, 2), (2, 3), (3, 2), (3, 3), (4, 4), (5, 5), (6,$

6))。

对于每一个 $P \subseteq Q$ 确定一个对论域 U 的划分 U/I_P , 在相应的等价类中的对象关于属性 P 有相同的描述, 如, 对于 $P' = \{A_1, A_2\}$, $U/I_{P'} = \{\{1\}, \{2, 3, 4\}, \{5\}, \{6\}\}$ 。这样, $\{1\}, \{2, 3, 4\}, \{5\}, \{6\}$ 是 P' -基础集。

若选属性集为 $P = \{A_1, A_2, A_3\}$, 则总体评价为“好”的学生的近似集 $X = \{1, 3, 6\}$, 因为 $U/I_P = \{\{1\}, \{2, 3\}, \{4\}, \{5\}, \{6\}\}$, 于是

$$P(X) = \{1, 6\}, \bar{P}(X) = \{1, 2, 3, 6\}, B_{P'} = \{2, 3\}$$

以上结果表明 P -边界集 $B_{P'}(X)$ 非空, 学生2和学生3属于同一 P -边界集, 但学生2的评价为“完全坏”, 而学生3的评价为“好”。基于属性 P 的信息, 我们可以看出:

① 从 X 的 P -下近似得知, 学生1和学生6一定属于“好学生”的集合;

② 从 X 的 P -上近似得知, 学生1、2、3和学生6可能属于“好学生”的集合;

③ 从 X 的 P -边界得知, 学生2和学生3是“好学生”的集合中不确定成员。

在属性集 Q 中, 如果把条件属性 $C = \{A_1, A_2, A_3\}$ 和决策属性 $D = \{A_4\}$ 加以区别, 则数据表就可以视为决策表。下面给出经过约简后的决策规则:

- (1) 如果 $f(x, A_1) = \text{好}$, 则 $f(x, A_4) = \text{好}$; (1, 6)
- (2) 如果 $f(x, A_1) = \text{坏}$, 则 $f(x, A_4) = \text{坏}$; (4)
- (3) 如果 $f(x, A_1) = \text{中等}$, 且 $f(x, A_2) = \text{好}$, 则 $f(x, A_4) = \text{坏}$; (5)
- (4) 如果 $f(x, A_1) = \text{中等}$, 且 $f(x, A_2) = \text{坏}$, 则 $f(x, A_4) = \text{好或坏}$; (2, 3)

由上述规则不难看出, 规则(1), (2), (3) 只有一个单义结果, 所以它们是确切的规则, 但规则(4)的结论有两个结果, 这是因为(4)是一个近似规则。

最后, 我们把计算出的近似质量、在属性 C 中的属性所有子集的相互影响指标 I_S 和 I_B 及 Möbius 表示列在表2中。在表2中的第三列相应于 $\{A_1, A_2, A_3\}$ 的负值将作为在三个属

性联合贡献中信息冗余的度量。在第四列中的前三个值是 Shapley 值, 且可以解释为在粗糙近似中相应属性重要程度的度量, 其中 A_2 和 A_3 是互补的, 而 A_1 和 A_2 与 A_1 和 A_3 是可以互相替代的, 这列中相应影响指标的负值说明 A_1, A_2, A_3 间有一个冗余。对于第五列也可类似地做出解释。

表2 近似质量、Möbius 表示、相互影响指标

属性	质量	Möbius 表示	Shapley	Banzhaf
$\{A_1\}$	0.5	0.5	0.44	0.5
$\{A_2\}$	0	0	0.11	0.17
$\{A_3\}$	0	0	0.11	0.17
$\{A_1, A_2\}$	0.67	0.17	-0.17	-0.17
$\{A_1, A_3\}$	0.67	0.17	-0.17	-0.17
$\{A_2, A_3\}$	0.5	0.5	0.17	0.17
$\{A_1, A_2, A_3\}$	0.67	-0.67	-0.67	-0.67

结束语 本文在叙述有关粗糙集的基本概念的基础上, 重点介绍用模糊测度来度量粗糙质量的基本方法: 将对策论的 N 人合作对策的 Shapley 值及 Banzhaf 值作为度量多属性决策的个体属性之间的相互影响的指标, 为多准则决策分析提供了新的有效方法。

参考文献

(上接第83页)

- 4 Wexelblat A, Maes P. Using history to assist information browsing. RIAO'97: Computer-assisted information retrieval on the Internet. Montreal, 1997
- 5 Fink J, Kobza A, Nill A. User-oriented adaptivity and adaptability in the AVANTI project. Designing for the Web: Empirical studies, Microsoft Usability group, Redmond, WA
- 6 Schechter S, Krishnan M, Smith M D. Using path profiles to predict HTTP requests. In: Proc. of 7th Intl. World Wide Web Conf. Brisbane, Australia, 1998
- 7 Cooley R, Mobasher B, Srivastava J. Data preparation for mining World Wide Web browsing patterns. Journal of Knowledge and Information Systems, 1999(1)
- 8 Buchner A, Mulvenna M D. Discovering internet marketing intelligence through online analytical Web usage mining. SIGMOD Record, 1999, 27(4)

- 9 Nasraoui O, et al. Mining Web access logs using relational competitive fuzzy clustering. To appear in the Proceedings of the Eight International Fuzzy Systems Association World Congress, Aug. 1999
- 10 汪晓岩, 胡庆生, 李斌, 庄镇泉. 面向 Internet 的个性化智能信息检索. 计算机研究与发展, 1999. 9
- 11 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc. of the 20th VLDB Conf. Santiago, Chile, 1994. 487~499
- 12 Savasere A, Omiecinski E, Navathe S. An efficient algorithm for mining association rules in large databases. In: Proc. of the 21th VLDB Conf. Zurich, Switzerland, 1995. 432~443
- 13 Yan T, Jacobsen M, Garcia-Molina H, Dayal U. From user access patterns to dynamic hypertext linking. In: Proc. of the 5th Intl. World Wide Web Conf. Paris, France, 1996