

决策表中规则获取的不确定性研究*

马志锋 邢汉承

郑晓妹

(东南大学计算机科学与工程系 南京 210096) (南京航空航天大学计算机科学与工程系)

摘 要 知识获取的不确定性主要来源于有限的分辨能力以及对于数据描述的不确定性。首先将 Rough 集理论与不确定问题中的证据理论以及模糊集合理论进行比较,然后介绍不确定性数据的模糊描述。通过引入模糊区别矩阵和扩展近似集方法延伸了 Rough 集理论,并从模糊决策表中导出合理的决策规则。

关键词 不确定性, Rough 集理论, 决策制定, 模糊决策表

分类号 TP 18

Research on the Uncertainty of Rule Acquisition from Decision Table

Ma Zhifeng, Xing Hancheng

Zheng Xiaomei

(Southeast University)

(Nanjing University of Aeronautics and Astronautics)

Abstract The uncertainty of knowledge acquisition mainly results from two aspects. One is caused by the limited discernibility of decisions in terms of condition attributes. The other is originated from the uncertain description of data. Rough set theory is compared with evidence theory and fuzzy set theory. Then the different descriptions of uncertain data are introduced. Rough set theory is extended to induce the reasonable decision rules from fuzzy decision table by introducing fuzzy discernibility matrix and modified approximations.

Key words uncertainty, rough set theory, decision making, fuzzy decision table

1 引 言

不确定环境下的知识获取是智能信息处理中的关键问题之一。如何在信息不完全、不精确或模糊的情况下,根据决策系统中已有的决策数据获取知识,一直为众多学者所关注^[1,2]。

近年来,基于 Rough 集理论的知识获取方法已成为一种重要的方法^[2-6]。这是一种新型的处理不确定性知识的数学工具,其基本思想是在保持分类能力不变的前提下,通过知识的约简导出概念的分类规则。该方法的最大优点在于无需人为的额外假设条件,而是完全由已知数据来如实地回答问题,从

而开辟了一条与传统的方法所截然不同的新途径。然而当采用 Rough 集传统方法进行数据分析时,它所基于的数据往往要求是精确的或确定的^[1],这为实际应用带来了困难,有时甚至会使系统难以实现。

本文要探讨的是利用模糊区别矩阵和扩展近似集对决策表中规则的不确定性进行分析。该方法拓宽了融入不确定数据的模糊决策表的应用范围,为决策的不确定性研究提供了新思路。

2 决策表及其不确定性

Rough 集理论一般采用决策表来描述决策系统^[1,3,7]。决策表 DT 是由 4 元组 $U, A = \{d\}, V, \rho$ 构成的,其中, U 为对象的集合, A 为条件属性的集合, d 为决策属性, $V = \{v_a | a \in A\}$ 为属性值域, $\rho = U \times$

* 国家自然科学基金项目(69673010)

$A = \{d\}$ V 为决策函数。设有 $B \subseteq A, x \in U$, 则决策规则可用来对未知对象进行决策分类, 并具有如下形式

$$\left(\begin{matrix} \rho(x, b) = m_b \\ \rho(x, d) = m_d^0 \\ \dots \\ \rho(x, d) = m_d^k \end{matrix} \right) \Rightarrow \left(\begin{matrix} \rho(x, d) = m_d^0 \\ \rho(x, d) = m_d^1 \\ \dots \\ \rho(x, d) = m_d^k \end{matrix} \right)$$

如果决策规则中 $k = 0$, 即规则的右端仅含一个析取项, 则称此规则为确定规则, 否则称为不确定规则。

2.1 Rough 集及其它不确定性理论

对决策分类的不同描述

Rough 集理论中的不确定性是通过边界的含糊性表达的。为客观地刻画这种不确定性, 每个决策类 $X \subseteq U$ 由一个二元组 $\{\mathcal{R}X, \mathcal{R}X\}$ 表示。其中, $\mathcal{R}X = \{[x] \in \mathcal{A} \mid [x] \subseteq X\}$ 称为关于 X 的下近似集, $[x] \in \mathcal{R}$ 为由不分明关系 \mathcal{R} 所确定的包含 x 的不分明类; $\mathcal{R}X = \{[x] \in \mathcal{A} \mid [x] \cap X \neq \emptyset\}$ 称为关于 X 的上近似集。决策类 X 的下近似集包含了所有能确切分类到 X 的对象集合, 上近似集包含了所有可能做出 X 决策的对象集合, 而由 $\mathcal{R}X - \mathcal{R}X$ 所形成的边界域中的对象则不能做出肯定或否定属于决策类 X 的判断。显然若边界域非空, 则决策类 X 便带有某种程度的不确定性。

对于不确定性知识的处理, 除了采用 Rough 集方法之外, 还有其它一些方法, 如传统的统计学方法, Dempster-Shafer 证据理论方法, 模糊集合理论方法等。Rough 集理论与这些理论相比, 尽管在对不确定性的处理上具有一定的相似之处, 但从概念的本质上说却存在很大的差别。

传统的统计学方法所基于的是决策表中决策数据的概率分布, 它采用 Bayes 概率定理进行规则提取。例如在医疗诊断决策表中, 设 D_i 是一组可能的疾病集合, C 是从病人观察到的一种病症。如果从大量统计中得知先验概率 $P(D_i)$ 以及所观察到的发生疾病 D_i 时病症 C 的条件概率 $P(C \mid D_i)$, 那么便可通过 Bayes 定理算出 $P(D_i \mid C)$ 。取 $P(D_k \mid C) = \max_i \{P(D_i \mid C)\}$, 则 D_k 即为病症 C 最可能得的疾病。然而由于该方法在推导过程中所用的先验概率和条件概率均需建立在大量的统计资料之上, 所以它不适合于只有少量决策数据的决策表。Rough 集方法恰好相反, 它不是减少数据的不确定性, 而是描述这种不确定性, 即便在决策数据相当少的情况下也可最大限度地导出决策规则。

的信任度量, 用似然度表示不否定某个概念的信任程度。信任函数与似然函数分别定义为

$$\text{Bel}(X) = \sum_{Y \subseteq X} m(Y)$$

$$\text{Pl}(X) = 1 - \text{Bel}(\bar{X}) = \sum_{Y \cap X \neq \emptyset} m(Y)$$

其中 $m: 2^U \rightarrow [0, 1]$ 为基本概率分配函数, 表示证据对 U 的子集 X 成立的一种信任度量。显然有 $0 \leq \text{Bel}(X) \leq \text{Pl}(X) \leq 1, \text{Pl}(X) - \text{Bel}(X)$ 表示既不信任 X 也不信任 \bar{X} 的一种度量, 可表示对 X 未知的度量。用区间 $(\text{Bel}(X), \text{Pl}(X))$ 来描述 X 的不确定性: $(\text{Bel}(X), \text{Pl}(X)) = (1, 1)$ 表示 X 为真; $(\text{Bel}(X), \text{Pl}(X)) = (0, 0)$ 表示 X 为假; $(\text{Bel}(X), \text{Pl}(X)) = (0, 1)$ 表示对 X 一无所知。通常信任度和似然度是由某个或某些专家给出的, 因而带有很大的主观性。而 Rough 集的上、下近似集则完全是根据已有数据采用确定方法计算出来的, 并不需要关于数据的任何附加信息。

模糊集合理论是对普通集合理论的拓广。它将不确定性理解为可能性, 模糊决策类则被描述为一个模糊子集, 然而模糊隶属函数必须事先人为确定。而在 Rough 集中, 对象 x 与决策类 X 之间的隶属关系完全是根据决策分类知识客观计算出来的, 无需主观给定, 即

$$\mu_{\mathcal{R}}^X(x) = \frac{|\{[x] \in \mathcal{A} \mid [x] \subseteq X\}|}{|\{[x] \in \mathcal{A} \mid [x] \cap X \neq \emptyset\}|}$$

2.2 不确定性数据的模糊决策表描述

实际问题中所遇到的决策表, 往往由于某些原因, 决策数据并非可以精确描述。这给采用传统 Rough 集方法进行决策带来了一定的麻烦, 因为决策分类的不确定性尽管可以通过 Rough 近似集加以体现, 但决策表中数据本身的不确定性在方法中并未得到有效处理。

事实上, 决策表中数据的不确定性均可通过模糊决策表的形式予以表达。

- 1) 定量属性的不确定性: 未经处理的定量数据一般很难用 Rough 集分析, 这里可将定量数据的不确定性转化成对于不同定性术语的模糊隶属度来描述;
- 2) 含糊数据的不确定性: 可将这种不确定性变成对于各个不明确概念的可能性分布;
- 3) 噪声数据的不确定性: 可由专家给出若干个取值子区间, 而不必拘泥于精确取值, 通过模糊隶属于不同的子区间来表达其不确定性;
- 4) 不完全数据的不确定性: 数据的丢失或无法获取, 可根据已有决策数据推测出可能落入某个取

值子区间的可能性。

以上各种不确定性均可采用模糊数据的形式加以描述。模糊数的表达有多种方式^[8,9], 如: 以模糊集合表示的模糊数, 以模糊区间数表示的模糊数, 以模糊中心数表示的模糊数等。

定义 1 4元组 $DT = (U, A, \{d\}, V, \pi)$ 称为模糊决策表。其中, A 为条件属性, d 为决策属性, $V = \bigcup_{a \in A} V_a$ 为属性值域, V 为定义在 V 上的模糊数的集合(其模糊数可采用上述各种表达形式), 映射 $\pi: U \times A \rightarrow \{d\} \rightarrow V$ 为模糊决策函数。由此导出的模糊决策规则可用于对未知对象的决策分类。

3 模糊决策表中的规则获取

为利用 Rough 集从决策表中提取决策规则, 通常可采用两种方法: 区别矩阵方法或 Rough 近似集方法。然而, 当决策表中的决策数据取为模糊数据时, 这两种传统的方法均不适用, 因而有必要扩充现有的方法。

3.1 模糊区别矩阵

Skowron 在文献[7]中提出采用区别矩阵和区别函数来计算决策表的约简方法。区别矩阵

$$M(DT) = (m_{ij})_{n \times n}$$

$$m_{ij} = \{a \in A \mid (\rho(x_i, a) \neq \rho(x_j, a)) \vee (\rho(x_i, d) \neq \rho(x_j, d))\}$$

$$i, j = 1, \dots, n$$

其中 $x_i, x_j \in U, n = |U/\text{IND}(A)|$, $\text{IND}(A)$ 为由属性集 A 所确定的不分明关系。决策表 DT 的区别函数 Ψ_{DT} 定义为含有 $k = |A|$ 个布尔变量 a_1^*, \dots, a_k^* 的布尔函数, 即

$$\Psi_{DT}(a_1^*, \dots, a_k^*) = \bigwedge_{j < i} \{m_{ij}^* \mid 1 \leq j < i \leq n, m_{ij} \neq \emptyset\}$$

$$m_{ij}^* = \{a^* \mid a \in m_{ij}\}$$

区别函数的所有最小蕴含项对应于决策表 DT 的约简。

当决策表中的决策数据允许为模糊数据时, 如何采用区别函数从模糊决策表中获取决策规则呢? 这需要对区别矩阵的定义做适当修改。问题的关键在于模糊决策表中两模糊数据间的语义距离做何解释。设 $DT = (U, A, \{d\}, V, \pi)$ 为模糊决策表, D_1, \dots, D_t 为 DT 上的决策类, 其中 $t = |\text{IND}(d)|$ 。

定义 2 $s_a: V_a \times V_a \rightarrow \mathbf{R}_+$ 为属性 $a \in A$ 上的两个模糊数 $v_{a_i}, v_{a_j} \in V_a$ 间的距离, 则有:

1) 当模糊数以模糊集合的形式表达时, 若隶属

函数取值分别为 $\mu(v, v_{a_i})$ 和 $\mu(v, v_{a_j})$, 则可定义距离函数 $s_a(v_{a_i}, v_{a_j})$ 为 $|\mu(v, v_{a_i}) - \mu(v, v_{a_j})|$ 的某种范数 $\mu(v, v_{a_i}) - \mu(v, v_{a_j})$ 。

① 切比雪夫距离

$$s_a(v_{a_i}, v_{a_j}) = \max_v |\mu(v, v_{a_i}) - \mu(v, v_{a_j})|$$

② 欧几里得距离

$$s_a(v_{a_i}, v_{a_j}) = \left\{ \sum_v |\mu(v, v_{a_i}) - \mu(v, v_{a_j})|^2 \right\}^{1/2}$$

③ 加权明科夫斯基距离

$$s_a(v_{a_i}, v_{a_j}) = \left\{ \sum_v w(x) |\mu(v, v_{a_i}) - \mu(v, v_{a_j})|^p \right\}^{1/p}$$

其中 $w(x)$ 为权函数, $p \geq 1$ 。

2) 当模糊数以模糊区间数表示时, 设 $v_{a_i} = [e_1, f_1]/p_1, v_{a_j} = [e_2, f_2]/p_2, \mathcal{Q}(\cdot, \cdot)$ 为事先定义的一种距离, 则有:

① $s_a(v_{a_i}, v_{a_j}) =$

$$w_1 [\mathcal{Q}(e_1, e_2) + \mathcal{Q}(f_1, f_2)] + w_2 |p_1 - p_2|$$

其中 $w_1 + w_2 = 1, w_1, w_2 \geq 0$ 为权系数;

② $s_a(v_{a_i}, v_{a_j}) =$

$$[w_1 |e_1 - e_2|^p + w_2 |f_1 - f_2|^p + w_3 |p_1 - p_2|^p]^{1/p}, \quad p \geq 1$$

其中 $w_1 + w_2 + w_3 = 1, w_1, w_2, w_3 \geq 0$ 为权系数。

3) 当模糊数以模糊中心数表示时, 令 $v_{a_i} = (c_1, r_1, p_1), v_{a_j} = (c_2, r_2, p_2)$, 于是有:

① 先将 v_{a_i} 和 v_{a_j} 转化成区间数 $[c_1 - r_1, c_1 + r_1]/p_1$ 和 $[c_2 - r_2, c_2 + r_2]/p_2$, 然后通过 2) 求出 $s_a(v_{a_i}, v_{a_j})$;

② $s_a(v_{a_i}, v_{a_j}) =$

$$[w_1 \mathcal{Q}(c_1, c_2)^p + w_2 |r_1 - r_2|^p + w_3 |p_1 - p_2|^p]^{1/p}, \quad p \geq 1$$

其中 $w_1 + w_2 + w_3 = 1, w_1, w_2, w_3 \geq 0$ 为权系数。

4) 当一个为精确数 v_{a_i} , 另一个为区间数或中心数 v_{a_j} 时, 可将精确数 v_{a_i} 表示为区间数 $[v_{a_i}, v_{a_i}]/1$ 和中心数 $[v_{a_i}, 0, 1]$, 然后采用上述方法分别计算 $s_a(v_{a_i}, v_{a_j})$ 。

需要指出的是, 当模糊数以模糊集合组、模糊区间数组和模糊中心数组的形式出现时, 假设有变换函数 $g: \mathbf{R} \rightarrow \mathbf{R}_+$, 其中 m_a 为关于属性 a 的模糊数

组的阶,表示将 m_a 维距离空间映射到一维空间上。

定义3 $M(DT) = (m_{ij})_{n \times n}$ 为模糊决策表 DT 的模糊区别矩阵,其中

$$m_{ij} = \{ a \in A \mid S_d(\pi(x_i, a), \pi(x_j, a)) \leq \delta_a, S_d(\pi(x_i, d), \pi(x_j, d)) \leq \delta_d, i, j = 1, 2, \dots, n \}$$

$$x_i, x_j \in U, n = |U|$$

定义4 模糊决策表 DT 的区别函数定义为

$$\Psi_{DT}(a_1^*, \dots, a_k^*) = \{ m_{ij}^* \mid 1 \leq j < i \leq n, m_{ij} \in \emptyset \}$$

其中, $k = |A|$, $m_{ij}^* = \{ a \in A \mid m_{ij} \in \emptyset \}$ 表示所有能够近似区分模糊决策对象 x_i 和 x_j 的属性子集所对应的变量。

化简区别函数 $\Psi_{DT}(a_1^*, \dots, a_k^*)$ 为关于变量 a_1^*, \dots, a_k^* 的主析取范式,所得极小项即为模糊决策表的约简属性。定义3中增加了条件 $S_d(\pi(x_i, d), \pi(x_j, d)) \leq \delta_d$,其目的是为了表明无需区别具有相似决策值的条件属性。另外,考虑到模糊区别矩阵中,一般 n 的取值会较大,这样不利于约简的计算。因此,解决好这一问题已成为能否有效地应用模糊区别矩阵的关键。文献[5]介绍了一种采用相似关系来浓缩原始决策表以减小区别矩阵规模的方法。限于篇幅,此处不再赘述。

3.2 扩展近似集计算

传统 Rough 集理论中,一个概念(决策类)的模糊性是指其具有不可明确划分的边界。相应地,决策规则可分为确定规则与可能规则,其中,前者对应于下近似集,后者对应于上近似集。然而在模糊决策表中,直接采用上、下近似集的定义进行规则获取有时是难以实现的。因为这里牵涉到模糊聚类的问题,聚类效果将直接影响到规则的优劣。因此,若能找出一种无需模糊聚类的方法,则无疑可大大提高模糊决策的效率。

事实上,上、下近似集定义中的关键在于两个集合间的包含与相交程度。为此,本节仿照 Rough 近似集的定义,将模糊决策表中属性的各模糊取值视为关于对象论域 U 上的模糊集合,结合模糊集合的运算规则,给出以下关于两模糊数间的包含与相交测度算子的概念。

定义5 设有任意两个模糊集合 v 和 v' ,则它们之间的包含与相交测度算子 $\alpha(v \subseteq v')$ 和 $\beta(v \cap v')$ 分别为

$$\alpha(v \subseteq v') = \alpha(v \cap v') / \alpha(v)$$

$$AvgMax(1 - \mu(u, v), \mu(u, v'))$$

$$\beta(v \cap v') = \frac{1}{|U|} \sum_{u \in U} \min(\mu(u, v), \mu(u, v'))$$

其中 Avg, Max, Min 分别表示论域中各对象隶属于模糊集合的隶属度的平均值、最大值和最小值。

定理1 若 v 和 v' 为两个普通经典集合,则有:

- 1) 当且仅当 $v \subseteq v'$ 时, $\alpha(v \subseteq v') = 1$, 否则 $0 < \alpha(v \subseteq v') < 1$;
- 2) 当且仅当 $v \cap v' \neq \emptyset$ 时, $\beta(v \cap v') = 1$, 否则 $\beta(v \cap v') = 0$ 。

证明略。

由此可见,经典集合在包含与相交测度算子 $\alpha(v \subseteq v')$ 和 $\beta(v \cap v')$ 的意义下乃是模糊集合的特例。

下面介绍如何利用包含和相交测度算子从模糊决策表中获取决策规则。设有模糊决策表 $DT = (U, A, \{d\}, V, \pi, v_a \in V_a)$ 是属性 $a \in A$ 上的一个模糊数取值, V_a 由一组定义在 V_a 上的模糊概念的集合 $\Omega_a = \{V_{a_1}, \dots, V_{a_n}\}$ 来描述,即 $v_a = \{ \mu(v_a, V_{a_1}) / V_{a_1}, \dots, \mu(v_a, V_{a_n}) / V_{a_n} \}$ 。同样, $v_d \in V_d$ 是决策属性 d 上的一个模糊决策值,亦可表示为 $v_d = \{ \mu(v_d, V_{d_1}) / V_{d_1}, \dots, \mu(v_d, V_{d_n}) / V_{d_n} \}$ 。这样,关于 Ω_a 中的任一模糊概念 V_{a_i} ,便可形成一个模糊向量 $v = \{ \mu(\pi(x_1, a_i), V_{a_i}), \dots, \mu(\pi(x_h, a_i), V_{a_i}) \}$,它代表了模糊决策中各对象的 V_{a_i} 情况,其中 $h = |U|$ 为模糊决策表 DT 中决策对象的个数。 $v = \{ \mu(\pi(x_1, d_i), V_{d_i}), \dots, \mu(\pi(x_h, d_i), V_{d_i}) \}$,它代表了可能作出 V_{d_i} 的决策情况。

考虑所有的 v 以及不同 v 的相交组合对于 v 的包含与相交测度 $\alpha(\Pi \subseteq v)$ 和 $\beta(\Pi \cap v)$ 其中 $\Pi = (v)$ 。若 $\alpha(\Pi \subseteq v) \geq \delta_\alpha$,则可得出模糊确定规则为

$$\text{if des}(\Pi) \text{ then des}(v) \text{ with certainty} = \alpha(\Pi \subseteq v)$$

若 $\beta(\Pi \cap v) \geq \delta_\beta$,则可得出模糊可能规则为

$$\text{if des}(\Pi) \text{ then des}(v) \text{ with possibility} = \beta(\Pi \cap v)$$

其中, $\delta_\alpha, \delta_\beta$ 分别为包含和相交测度的最小阈值, $\text{des}(\cdot)$ 表示对于所有出现在 (\cdot) 中的模糊概念的描述。

通过对部分实验数据的模拟计算发现,采用模糊区别矩阵和扩展近似集计算所得的规则能很好地体现模糊决策表中的关键决策信息,而采用传统 Rough 集方法则很难从模糊决策表中获取信息。

4 结 论

本文研究了决策表中规则获取的不确定性。通过分析发现这种不确定性主要源于两方面,即由于条件属性对于决策属性的有限分辨能力所引起的决策规则的不一致性,以及由于决策数据描述本身所带来的不确定性。与不确定问题中的统计学理论、Dempster-Shafer 证据理论以及模糊集合理论对比, Rough 集仅仅利用决策表中所提供的决策数据进行分析,不需要提供任何关于决策数据的附加信息。在探讨了决策表中不确定数据的各种模糊描述的基础上,分析了模糊数据间的距离,从而为在 Rough 集中应用不确定数据进行决策开辟了新的途径。实践证明文中所给出的方法对于模糊决策表的处理是可行且高效的。

参 考 文 献

- 1 Pawlak Z. Rough sets-theoretical aspects of reasoning about data. Dordrecht: Kluwer Academic Publishers, 1991
- 2 Slowinski R, Stefanowski J. Handling various types of uncertainty in the rough set approach. In: Rough Sets, Fuzzy Sets and Knowledge Discovery. London: Springer-Verlag, 1994. 366~376
- 3 马志锋, 邢汉承, 郑晓妹. 基于多元组 Rough 集的不相容决策. 东南大学学报, 1999, 29(3): 28~33
- 4 Ma Zhifeng, Xing Hancheng, Zheng Xiaomei. A multi-tuple rough set approach for information retrieval. J of Southeast University (English Edition), 1999, 15(1):

61~66

- 5 马志锋, 邢汉承, 郑晓妹. 基于不分明与相似关系的 Rough 集的超图描述. 计算机科学, 1999, 26(9): 35~39
- 6 Ma Zhifeng, Xing Hancheng, Zheng Xiaomei. An attribute-oriented multi-tuple rough set approach for knowledge discovery in relational databases. In: The Fifth Int Conf for Young Computer Scientists. Beijing: International Academic Publisher, 1999. 460~465
- 7 Skowron A. The discernibility matrices and functions in information systems. In: Intelligent Decision Support: Handbook of Applications and Advances of Rough Sets Theory. Dordrecht: Kluwer Academic Publishers, 1992. 331~362
- 8 何新贵. 模糊知识处理的理论与技术. 北京: 国防工业出版社, 1994
- 9 Worm J A. Measuring uncertainty by extracting fuzzy rules using rough sets. Technical Report of Research Institute for Computing and Information Systems, University of Houston-Clear Lake, 1991

作 者 简 介

马志锋 男, 1970 年生。1997 年于东南大学自动控制系获硕士学位, 现为东南大学计算机系博士研究生。研究方向为人工智能, 知识获取, Rough 集理论及应用。

邢汉承 男, 1938 年生。1960 年毕业于哈尔滨工业大学, 现为东南大学计算机系教授, 博士生导师。主要研究方向为人工智能, 模式识别与图象处理。

郑晓妹 女, 1973 年生。1995 年毕业于上海铁道学院计算机系, 现为南京航空航天大学计算机系硕士研究生。研究方向为人工智能, 数据库理论及应用。