

混合语义下的不完备多尺度决策系统及其最优尺度选择

刘微, 胡军

1. 重庆邮电大学 计算智能重庆市重点实验室, 重庆 400065; 2. 重庆邮电大学 计算机科学与技术学院, 重庆 400065

摘要: 现有不完备多尺度信息系统的定义存在局限性, 即只考虑了遗漏型未知属性值并且认为对象在某个属性下缺失则在所有尺度下全部缺失. 本研究从未知属性值的两种语义解释出发, 定义混合语义下的不完备多尺度信息系统. 为保持混合语义下的不完备多尺度信息系统尺度间的偏序关系, 根据粒信息转换函数填补数据并扩展特征关系模型. 研究在不同尺度下信息粒度的表示及其相互关系, 并进一步定义基于扩展后特征关系的集合的上、下近似, 讨论了它们的性质. 在混合语义下的不完备多尺度决策系统中构建序贯三支决策模型, 并基于该模型给出最小化不确定性的最优尺度选择方法. 实验结果验证了数据填补和特征关系模型扩展的有效性, 并且表明通过序贯三支决策模型能够选择不确定性最小的尺度作为最优尺度.

关键词: 多尺度; 不完备; 特征关系; 三支决策; 最优尺度选择

中图分类号: TP18 **文献标志码:** A

文章编号: 1673-9868(2023)03-0002-17

开放科学(资源服务)标识码(OSID):



Incomplete Multi-Scale Decision System with Mixed Semantics and Its Optimal Scale Selection

LIU Wei, HU Jun

1. Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China;

2. College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract: The definition of the existing incomplete multi-scale information system has limitations. Namely, only the missing unknown value is considered, and objects unknown at an attribute are missing at all scales of the attribute. First of all, the incomplete multi-scale information system with mixed semantics is

收稿日期: 2022-07-23

基金项目: 国家自然科学基金项目(61936001, 62276038); 重庆市自然科学基金项目(cstc2021ycjh-bgzxm0013); 重庆市教委重点合作项目(HZ2021008).

作者简介: 刘微, 硕士研究生, 主要从事粗糙集、粒计算研究.

通信作者: 胡军, 教授, 博士研究生导师.

defined in this paper based on the two kinds of semantic interpretations of unknown value. Secondly, in order to maintain the partial order relation between scales in incomplete multi-scale information systems with mixed semantics, the data is supplemented according to the granular information transformation function and the characteristic relation model is extended. Then, the representation of information granularity at different scales and their interrelations are given. The concepts of upper and lower approximation of sets based on extended characteristic relation are further defined, and their properties are discussed. Finally, a sequential three-way decisions model is constructed in an incomplete multi-scale decision system with mixed semantics, and an optimal scale selection method for minimizing uncertainty is given based on this model. The experimental results verify the validity of data filling and the extension of the characteristic relation model, and show that the scale with the least uncertainty can be selected as the optimal scale by the sequential three-way decision model.

Key words: multi-scale; incomplete; characteristic relation; three-way decisions; optimal scale selection

Zadeh^[1]在1979年首次提出信息粒化的概念,认为人类认知能力可概括为粒化、组织和因果3个主要特征,并由此引出粒计算的概念。目前粒计算已成为智能信息处理领域的一个热门研究方向,在复杂问题求解、海量数据挖掘和模糊信息处理中具有独特优势。它模拟人类思考方式,以粒为基本计算单位,强调对问题多角度、多尺度的理解和描述。由波兰学者 Pawlak^[2]提出的粗糙集理论是一种处理不精确、不确定信息的数学工具,其利用等价关系诱导的划分粒化数据样本集,使用由等价类描述的粒代替样本对数据进行表示和处理,并通过约简对数据集进行特征提取^[3]。在众多粒计算研究方法中,粗糙集理论对推动和发展粒计算研究发挥着重要的作用。

现实世界中由于数据测量的误差、对数据的理解或获取的限制等因素,使得数据呈现不完备性。如何科学地处理数据的缺失是当前知识发现领域的难点之一。使用粗糙集理论处理不完备信息系统(incomplete information system, IIS)的方法大致分为两类:① 间接处理,即采用数据补齐或数据删除将 IIS 转化为完备信息系统^[4-5];② 直接处理,即将粗糙集模型在 IIS 中进行扩展。就目前关于 IIS 的研究来看,一般都是将未知属性值分为“缺席型”和“遗漏型”两种语义进行讨论。其中,缺席型未知属性值指存在且有用的缺失值,只是由于某种原因而丢失或者没有获取到;遗漏型未知属性值指对决策或分类结果没有影响的缺失值,被称为“不关心”条件。针对 IIS 中所有未知属性值都是缺席型的情况,Stefanowski 等^[6]构建了相似关系(自反性和传递性)。当 IIS 中所有未知属性值都是遗漏型时,Kryszkiewicz^[7]构建了满足自反性和对称性的容差关系。然而无论是相似关系还是容差关系都只考虑了未知属性值的一种语义解释,为了处理同时包含遗漏和缺席型未知属性值的 IIS,Grzymala-Busse^[8]利用改进的属性-值对构建其中的特征关系,此广义不可分辨关系是结合容差关系和相似关系的一种更为一般的表现形式。此后,学者在 IIS 中进一步研究了特征关系^[9-11]。

在传统的粗糙集数据分析模型中,一个对象在每个属性下只能取唯一的属性值,这样的数据描述结构称为单粒度信息系统。然而,单一粒度或单一尺度结构下的知识表示和数据挖掘已不能满足实际应用的需求。因此,Qian 等^[12]提出多粒化粗糙集模型,认为多粒度是由属性选择引起的,该模型根据信息系统中多个属性子集构成知识多粒度空间;Zhu 等^[13]以邻域粗糙集为基础提出多粒度邻域粗糙集模型,根据对象邻域半径的大小对论域进行粒化,然后选择合适的粒度进行分类或聚类。多粒化粗糙集模型和多粒度邻域粗糙集模型处理的数据形式实质上还是单尺度的。然而在实际应用中,对象在同一属性下根据不同的需求以不同的尺度进行测量会取不同的属性值,而在不同尺度下进行决策可能会导致不同的结果。基于这一情况,Wu 等^[14]认为对象的属性取值的多尺度是引起论域的多粒度粒化的一个原因,由此提出多尺度数据的粒计算分析模型,并进一步讨论了最优尺度选择和规则获取问题^[15-16]。

She 等^[17]研究了多尺度信息系统的局部最优尺度和规则提取. Chen 等^[18]和郑嘉文等^[19]从信息熵的角度分别讨论多尺度覆盖决策系统和多尺度决策系统的最优尺度选择问题. Zhang 等^[20]在考虑不确定性的同时, 结合代价敏感选择最优尺度. She 等^[21]通过引入粒度树和切割的概念在多尺度决策系统中定义了最优切割, 又进一步讨论了泛化约简^[22]. Wang 等^[23]在多尺度信息系统中构建多粒度粗糙集模型以选择最优尺度. 上述研究假设所有属性拥有相同的尺度级数^[14-23], Li 等^[24-25]提出广义多尺度信息系统的概念, 即不同属性可以拥有不同数目的尺度级数, 并研究了最优尺度组合选择问题. 此后, 许多学者也对最优尺度组合进行了研究^[26-28]. Hao 等^[29]基于三支决策的思想在多尺度决策系统中构建序贯三支决策模型, 并基于该模型在动态变化的多尺度决策系统中选择最优尺度. 上述关于多尺度信息系统的研究只针对完备数据, 于是 Wu 等^[30]将容差关系引入到多尺度信息系统中以应对数据的缺失, 并进一步研究最优尺度选择和规则获取问题^[31-32]. 顾沈明等^[33]在不完备多粒度决策系统中研究了局部最优粒度选择问题.

现有关于不完备多尺度信息系统的研究只考虑了遗漏型未知属性值, 并且假设对象在某个属性下缺失时则在该属性的所有尺度下全部缺失^[30-33]. 此外, 系统的不确定性随着尺度细化存在保持不变后减小的情况, 此时 Hao 等^[29]选择的最优尺度的不确定性不是最小. 考虑到上述情况, 本研究定义混合语义下的不完备多尺度信息系统, 即未知属性值包含两种语义解释且对象在属性下的缺失既可以全部缺失也可以部分缺失, 并且融合数据填补和模型扩展的方法以处理此类更为一般的不完备多尺度信息系统. 最后建立序贯三支决策模型, 利用三支决策表示决策系统的不确定性, 基于该模型给出最小化不确定性的最优尺度选择方法.

1 基础知识

本节主要介绍基于特征关系的不完备信息系统以及基于容差关系的不完备多尺度信息系统的基本概念.

定义 1^[8] 信息系统 S 可以表示为一个二元组 (U, A) , 其中 $U = \{x_1, x_2, \dots, x_n\}$ 是一个非空有限对象集, 称为论域; $A = \{a_1, a_2, \dots, a_m\}$ 是一个非空有限属性集, 对于任意的 $a \in A$, 满足 $a: U \rightarrow V_a$ 即 $a(x) \in V_a, x \in U$, 其中 $V_a = \{a(x) \mid x \in U\}$ 称为 a 的值域. $[(a, v)]$ 为属性-值对 (a, v) 的块, 记:

$$[(a, v)] = \{x \in U \mid a(x) = v\} \quad (1)$$

定义 2^[8] $S = (U, A)$ 是一个不完备信息系统, $V_a = \{a(x) \mid x \in U\}$ 为属性 $a \in A$ 的值域. 符号“*”和“ $\emptyset$ ”分别表示遗漏和缺席型未知属性值. 对于属性-值对的块 $[(a, v)]$, 有: ① 若 $a(x) = *$, 则 $x \in [(a, v)]$; ② 若 $a(x) = \emptyset$, 则 $x \notin [(a, v)]$. 其中 $v \in V_a$. 对于 $\forall x \in U, a \in B \subseteq A$, 记:

$$K(x, a) = \begin{cases} U & a(x) = * \vee a(x) = \emptyset; \\ [(a, a(x))] & \text{否则} \end{cases} \quad (2)$$

$K(x, a)$ 为 x 关于属性 a 的不可分辨对象集, 则特征集 $K_B(x)$ 定义为:

$$K_B(x) = \bigcap_{a \in B} K(x, a) \quad (3)$$

对于 $\forall x, y \in U$, 特征关系 R_B 为 U 上关于属性子集 $B \subseteq A$ 的广义不可分辨关系, 记:

$$(x, y) \in R_B \Leftrightarrow y \in K_B(x)$$

显然, 特征关系 R_B 仅满足自反性, 而不一定满足对称性和传递性. 若 IIS 中的所有未知属性值都是遗漏型, 则特征关系 R_B 退化为容差关系^[7]; 若 IIS 中的所有未知属性值都是缺席型, 则特征关系 R_B 退化为相似关系^[6]. 因此, 可以将特征关系看作是容差关系和相似关系的一种泛化表现形式.

设 (U, A) 是一个不完备信息系统, $\forall X \subseteq U, B \subseteq A$, 则 X 关于特征关系 R_B 的上近似集和下近似集分别定义为

$$\begin{aligned} \overline{R_B}(X) &= \{x \in U \mid K_B(x) \cap X \neq \emptyset\} \\ \underline{R_B}(X) &= \{x \in U \mid K_B(x) \subseteq X\} \end{aligned}$$

根据 X 关于 R_B 的上、下近似集,可将论域 U 划分为 3 个互不相交的区域,分别为正域、边界域及负域,定义为

$$\begin{aligned} POS(X) &= \underline{R_B}(X) \\ BND(X) &= \overline{R_B}(X) - \underline{R_B}(X) \\ NEG(X) &= U - \overline{R_B}(X) = U - \overline{R_B}(X) \end{aligned}$$

$POS(X)$ 表示一定属于 X 的对象集, $NEG(X)$ 表示一定不属于 X 的对象集, $BND(X)$ 表示不确定是否属于 X 的对象集. X 关于 R_B 的近似精度定义为:

$$\alpha_{R_B}(X) = \frac{|\underline{R_B}(X)|}{|\overline{R_B}(X)|}$$

其中 $|X|$ 表示集合 X 的基数. X 关于 R_B 的粗糙度定义为:

$$\rho_{R_B}(X) = 1 - \alpha_{R_B}(X)$$

定义 3^[14] $S=(U, A)$ 是一个多尺度信息系统, 其中 $U=\{x_1, x_2, \dots, x_n\}$ 为论域, $A=\{a_1, a_2, \dots, a_m\}$ 是一个非空有限属性集, 且每个属性都是多尺度属性. 假设所有属性都有 I 个相同的尺度级数, 则多尺度信息系统可以表示为 $(U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\})$, 其中, $a_j^k: U \rightarrow V_j^k$, V_j^k 是属性 a_j 在第 k 个尺度的值域, 对于 $j=1, 2, \dots, m, k=1, 2, \dots, I-1$, 存在满射函数 $g_j^{k,k+1}: V_j^k \rightarrow V_j^{k+1}$ 使得 $a_j^{k+1} = g_j^{k,k+1} \circ a_j^k$, 即:

$$a_j^{k+1}(x) = g_j^{k,k+1}(a_j^k(x)) \quad x \in U$$

称 $g_j^{k,k+1}$ 为粒信息转换函数. 对于 $k \in \{1, 2, \dots, I\}$, 记 $A^k = \{a_1^k, a_2^k, \dots, a_m^k\}$, 则一个多尺度信息系统 $S=(U, A)$ 可以分解为 I 个信息系统 $S^k=(U, A^k)$, $k=1, 2, \dots, I$.

若 $S^1=(U, A^1)$ 是一个不完备信息系统, 则称 $(U, A)=(U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\})$ 为不完备多尺度信息系统^[30]. 用符号“*”表示未知属性值, 即如果 $a_j^k(x)=*$, 就认为 x 在属性 a_j^k 上的值是未知的. 此时不同尺度层次之间的属性值变换为:

$$a_j^{k+1}(x) = \begin{cases} * & a_j^k(x) = * \\ g_j^{k,k+1}(a_j^k(x)) & \text{其他} \end{cases}$$

其中, $j=1, 2, \dots, m; k=1, 2, \dots, I-1; x \in U$.

2 不完备多尺度信息系统的扩展

本节主要给出混合语义下的不完备多尺度信息系统的定义, 并讨论数据填补和特征关系模型的扩展.

2.1 混合语义下的不完备多尺度信息系统

现有不完备多尺度信息系统的研究通过引入容差关系处理多尺度信息系统中的数据缺失, 但只考虑了遗漏型未知属性值, 并且认为多尺度信息系统 S 的不完备是由于 S^1 的不完备, 因此对象 x 在多尺度属性 a_j 下缺失, 则 $a_j^1(x)=a_j^2(x)=\dots=a_j^I(x)=*$, 即 x 在 a_j 的所有尺度下全部缺失. 而在实际应用中, 可能存在遗漏型和缺席型未知属性值共存于一个多尺度信息系统, 并且对象 x 只是在多尺度属性的某些尺度下缺失. 因此, 针对已有研究的不足, 本小节将定义混合语义下的不完备多尺度信息系统.

定义 4 多尺度信息系统中的某些属性值未知, 并且未知属性值是两种语义的任意解释, 则称该信息系统为混合语义下的不完备多尺度信息系统, 仍表示为 $S=(U, A)=(U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\})$. 符号“*”和“ $\emptyset$ ”分别表示遗漏和缺席型未知属性值, 即若 $a_j^k(x)=*$ 或 $a_j^k(x)=\emptyset$, 则认为 x 在属性 a_j^k 上的值未知, 否则 x 在属性 a_j^k 上的值已知.

例 1 表 1 是一个具有 3 个尺度和 2 个属性的混合语义下的不完备多尺度信息系统 $S=(U, A)=(U, \{a_j^k \mid k=1, 2, 3, j=1, 2\})$, 其中, $U=\{x_1, x_2, \dots, x_8\}$, $A=\{a_1, a_2\}$. S 同时具有遗漏和缺席型未知

属性,对象在多尺度属性下存在全部缺失也存在部分缺失,如 x_2 在 a_1 下部分缺失, x_5 在 a_2 下全部缺失.

表 1 混合语义下的不完备多尺度信息系统

U	a_1^1	a_1^2	a_1^3	a_2^1	a_2^2	a_2^3
x_1	1	S	Y	1	S	N
x_2	1	$\emptyset$	$\emptyset$	2	M	Y
x_3	2	S	Y	*	*	N
x_4	$\emptyset$	S	Y	3	M	Y
x_5	3	M	N	*	$\emptyset$	$\emptyset$
x_6	*	M	*	4	L	Y
x_7	4	L	N	*	L	$\emptyset$
x_8	5	L	N	5	L	Y

2.2 数据填补

定理 1 $(U, \{a_j^k \mid k=1, 2, \cdots, I, j=1, 2, \cdots, m\})$ 是一个多尺度信息系统,存在粒信息转换函数 $g_j^{k,k+1}: V_j^k \rightarrow V_j^{k+1}, k=1, 2, \cdots, I-1$, 对于 $\forall x, y \in U$:

- 1) $a_j^{k+1}(x) = g_j^{k,k+1}(a_j^k(x))$;
- 2) $a_j^{k+1}(x) \neq a_j^{k+1}(y) \rightarrow a_j^k(x) \neq a_j^k(y)$;
- 3) $a_j^k(x) = a_j^k(y) \rightarrow a_j^{k+1}(x) = a_j^{k+1}(y)$.

在不完备多尺度信息系统中,对于在多尺度属性 a_j 上部分缺失的对象,如果该对象在 a_j 的粗尺度下未知,而在细尺度下已知,此时根据定理 1 的 1) 可以填补粗尺度下的属性值. 因此,下面基于粒信息转换函数给出数据填补方法.

定义 5 $S=(U, A)=(U, \{a_j^k \mid k=1, 2, \cdots, I, j=1, 2, \cdots, m\})$ 是混合语义下的不完备多尺度信息系统, $g_j^{k,k+1}: V_j^k \rightarrow V_j^{k+1}$ 是相对 S 完备的多尺度信息系统的粒信息转换函数. 对于 $k \in \{1, 2, \cdots, I\}$, 记 $A^k = \{a_1^k, a_2^k, \cdots, a_m^k\}, S^k = (U, A^k)$, 数据填补后的 S^k 记为 S^{k*} , 数据填补后的 S 记为 S^* , 由 $S^{1*}, S^{2*}, \cdots, S^{I*}$ 组成, 其中 $S^{1*} = S^1$.

对于 $k \in \{1, 2, \cdots, I-1\}$, 根据 $g_j^{k,k+1}$ 及 S^{k*} 对 S^{k+1} 进行数据填补得到 $S^{k+1*}: \forall x \in U, a_j^k \in A^k$, 若 $a_j^{k+1}(x) = * \vee a_j^{k+1}(x) = \emptyset$ 并且 $a_j^k(x) \neq * \wedge a_j^k(x) \neq \emptyset$, 则对 $a_j^{k+1}(x)$ 进行填补有 $a_j^{k+1}(x) = g_j^{k+1}(a_j^k(x))$.

定理 2 设 $S=(U, A)=(U, \{a_j^k \mid k=1, 2, \cdots, I, j=1, 2, \cdots, m\})$ 是一个混合语义下的不完备多尺度信息系统, $a_j \in A, x \in U$, 对于数据填补后的 S^* 有:

- 1) 若 $a_j^k(x) = *$ 或 $a_j^k(x) = \emptyset$, 则有 $a_j^m(x) = *$ 或 $a_j^m(x) = \emptyset, m \in \{1, 2, \cdots, k\}$;
- 2) 若 $a_j^k(x) \neq *$ 且 $a_j^k(x) \neq \emptyset$, 则有 $a_j^m(x) \neq *$ 且 $a_j^m(x) \neq \emptyset, m \in \{k, k+1, \cdots, I\}$;
- 3) 若 x 在 a_j 下部分缺失, 即 $\exists k \in \{1, 2, \cdots, I-1\}$ 使得 $a_j^k(x) = * \vee a_j^k(x) = \emptyset$ 而 $a_j^{k+1}(x) \neq * \wedge a_j^{k+1}(x) \neq \emptyset$, 则有 $a_j^m(x) = * \vee a_j^m(x) = \emptyset, m \in \{1, 2, \cdots, k\}$, 并且 $a_j^n(x) \neq * \wedge a_j^n(x) \neq \emptyset, n \in \{k+1, k+2, \cdots, I\}$.

例 2 已知例 1 中 S 对应的完备多尺度信息系统的粒信息转换函数为: $g_1^{1,2}(1) = g_1^{1,2}(2) = S, g_1^{1,2}(3) = M, g_1^{1,2}(4) = g_1^{1,2}(5) = L; g_1^{2,3}(S) = Y, g_1^{2,3}(M) = g_1^{2,3}(L) = N; g_2^{1,2}(1) = S, g_2^{1,2}(2) = g_2^{1,2}(3) = M, g_2^{1,2}(4) = g_2^{1,2}(5) = L; g_2^{2,3}(S) = N, g_2^{2,3}(M) = g_2^{2,3}(L) = Y$.

对 S 进行数据填补:

根据 $g_j^{1,2}$ 及 $S^{1*} = S^1$ 对 S^2 进行数据填补得到 $S^{2*}: a_1^2(x_2) = \emptyset$ 并且 $a_1^1(x_2) = 1$, 则 $a_1^2(x_2) = g_1^{1,2}(1) = S$;

根据 $g_j^{2,3}$ 及 S^{2*} 对 S^3 进行数据填补得到 S^{3*} : $a_1^3(x_2) = \emptyset$ 并且 $a_1^2(x_2) = S$, 则 $a_1^3(x_2) = g_1^{2,3}(S) = Y$; $a_1^3(x_6) = *$ 并且 $a_1^2(x_6) = M$, 则 $a_1^3(x_6) = g_1^{2,3}(M) = N$; $a_2^3(x_7) = \emptyset$ 并且 $a_2^2(x_7) = L$, 则 $a_2^3(x_7) = g_2^{2,3}(L) = Y$.

由 S^{1*} , S^{2*} , S^{3*} 得到 S^* , 如表 2 所示.

表 2 数据填补后的混合语义下的不完备多尺度信息系统

U	a_1^1	a_1^2	a_1^3	a_2^1	a_2^2	a_2^3
x_1	1	S	Y	1	S	N
x_2	1	S	Y	2	M	Y
x_3	2	S	Y	*	*	N
x_4	$\emptyset$	S	Y	3	M	Y
x_5	3	M	N	*	$\emptyset$	$\emptyset$
x_6	*	M	N	4	L	Y
x_7	4	L	N	*	L	Y
x_8	5	L	N	5	L	Y

2.3 特征关系的模型扩展

由于数据填补后可能仍然存在未知值, 可以将特征关系引入数据填补后的不完备多尺度信息系统. 针对例 2 中数据填补后的 S^* , 根据公式 2 可得 x_7 关于属性子集 $\{a_2\}$ 在不同尺度下的特征集: $K_{\{a_2^1\}}(x_7) = U$, $K_{\{a_2^2\}}(x_7) = \{x_3, x_6, x_7, x_8\}$, $K_{\{a_2^3\}}(x_7) = \{x_2, x_4, x_6, x_7, x_8\}$. 显然不满足多尺度中尺度间的偏序关系 $K_{\{a_2^1\}} \subseteq K_{\{a_2^2\}} \subseteq K_{\{a_2^3\}}$.

观察表 2 可以发现, 在第 2 尺度下 $a_2^2(x_7)$ 与 $a_2^2(x_1)$, $a_2^2(x_2)$, $a_2^2(x_4)$ 不相等, 在第 3 尺度下 $a_2^3(x_7)$ 与 $a_2^3(x_3)$ 不相等, 由定理 1 知, 在第 1 尺度下 $a_2^1(x_7)$ 与 $a_2^1(x_1)$, $a_2^1(x_2)$, $a_2^1(x_3)$, $a_2^1(x_4)$ 不相等; 因 $a_2^2(x_5) = \emptyset$ (不可比较), 则有 $x_5 \notin [(a_2^2, L)] = K(x_7, a_2^2)$, 而 $x_5 \in K(x_7, a_2^1)$, 为满足尺度间的偏序关系, 使得 $x_5 \notin K(x_7, a_2^1)$. 由此得到 $K_{\{a_2^1\}}(x_7) = \{x_6, x_7, x_8\}$.

根据上述的分析, 需要扩展特征关系模型以满足数据填补后混合语义下的不完备多尺度信息系统尺度间的偏序关系.

定义 6 $S = (U, A) = (U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\})$ 是一个多尺度信息系统, $[(a_j^k, v)]$ 为属性-值对 (a_j^k, v) 的块, 记:

$$[(a_j^k, v)] = \{x \in U \mid a_j^k(x) = v\} \quad (4)$$

定义 7 $S = (U, A) = (U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\})$ 是一个混合语义下的不完备多尺度信息系统. 对于数据填补后的 S^* , $V_j^k = \{a_j^k(x) \mid x \in U\}$ 是属性 a_j 在第 k 个尺度的值域, 对于属性-值对的块 $[(a_j^k, v)]$, 有:

1) 若 $a_j^k(x) = *$, 则 $x \in [(a_j^k, a_j^k(z))]$, 其中 $z \in U - Y$, $Y = \{y \in U \mid (m \neq 0 \wedge a_j^m(x) \neq a_j^m(y), m = m_{a_j^k}(x)) \vee (\exists a_j^n(x) = \emptyset, k < n \leq I)\}$;

2) 若 $a_j^k(x) = \emptyset$, 则 $x \notin [(a_j^k, v)]$, 其中 $v \in V_j^k$.

对于 $\forall x \in U, k \in \{1, 2, \dots, I\}, a_j \in B \subseteq A$, $K(x, a_j^k)$ 是 x 关于属性 a_j^k 的广义不可分辨对象集, 定义如下:

1) 若 $a_j^k(x) \neq \emptyset$ 且 $a_j^k(x) \neq *$, 则

$$K(x, a_j^k) = [(a_j^k, a_j^k(x))] \quad (5)$$

2) 若 $a_j^k(x) = \emptyset$ 或 $a_j^k(x) = *$:

当 $a_j^I(x) = \emptyset$ 或 $a_j^I(x) = *$ 时,

$$K(x, a_j^k) = U \quad (6)$$

当 $a_j^I(x) \neq \emptyset$ 且 $a_j^I(x) \neq *$ 时,

$$K(x, a_j^k) = U - \{y \in U \mid (m_{a_j^k}(y) \neq 0 \wedge a_j^m(x) \neq a_j^m(y)) \vee (\exists a_j^n(y) = \emptyset, m_{a_j^k}(x) \leq n \leq I)\} \quad (7)$$

其中

$$m = \max(m_{a_j^k}(x), m_{a_j^k}(y)), m_{a_j^k}(x) = \begin{cases} 0 & a_j^I(x) = \emptyset \text{ 或 } a_j^I(x) = * \\ \min\{k_1 \mid a_j^{k_1}(x) \neq * \wedge a_j^{k_1}(x) \neq \emptyset, k < k_1 \leq I\} & \text{否则} \end{cases}$$

$m_{a_j^k}(x)$ 表示对象 x 在属性 a_j 的 k 尺度后属性值已知的最细尺度, 若 k 尺度后属性值全部未知则 $m_{a_j^k}(x) = 0$. 记:

$$K_{B^k}(x) = \bigcap_{a_j^k \in B^k} K(x, a_j^k) \quad (8)$$

$K_{B^k}(x)$ 是对象 x 关于属性集 B 在第 k 尺度下的特征集.

对于 $\forall x, y \in U, k \in \{1, 2, \dots, I\}, B \subseteq A$, 特征关系 R_{B^k} 是 S^* 在第 k 尺度下由属性子集 B 导出的一个广义不可分辨关系:

$$(x, y) \in R_{B^k} \Leftrightarrow y \in K_{B^k}(x)$$

对于 $x \in U, k \in \{1, 2, \dots, I-1\}$, 混合语义下的不完备多尺度信息系统 S^* 的粒信息转换函数不是满射函数, 仅当 $a_j^k(x) \neq \emptyset$ 且 $a_j^k(x) \neq *$ 时, 满足 $a_j^{k+1}(x) = g_j^{k+1}(a_j^k(x))$. 一个混合语义下的不完备多尺度信息系统 S^* 可以分解为 I 个信息系统 $S^{*k} = (U, A^k), k = 1, 2, \dots, I$.

例 3 例 2 中数据填补后的 S^* , 由定义 7 可得到 x_7 关于 $B = A$ 在不同尺度下的特征集, 可得 $K_{B^1}(x_7) \subseteq K_{B^2}(x_7) \subseteq K_{B^3}(x_7)$.

$$K_{B^1}(x_7) = K(x_7, a_1^1) \cap K(x_7, a_2^1) = \{x_7\} \cap \{x_6, x_7, x_8\} = \{x_7\}$$

$$K_{B^2}(x_7) = K(x_7, a_1^2) \cap K(x_7, a_2^2) = \{x_7, x_8\} \cap \{x_6, x_7, x_8\} = \{x_7, x_8\}$$

$$K_{B^3}(x_7) = K(x_7, a_1^3) \cap K(x_7, a_2^3) = \{x_5, x_6, x_7, x_8\} \cap \{x_2, x_4, x_6, x_7, x_8\} = \{x_6, x_7, x_8\}$$

定义 8 S_1 和 S_2 是非空集合 U 上的 2 个混合语义下的不完备多尺度信息系统, 对于数据填补后的 S_1^* 和 S_2^* 分别有扩展后的特征关系 R_1 和 R_2 , 若对于 $\forall x \in U$ 总有 $K_{R_1}(x) \subseteq K_{R_2}(x)$, 那么就称 U/R_1 较 U/R_2 更细, 或 U/R_2 较 U/R_1 更粗, 记为 $U/R_1 \subseteq U/R_2$.

定理 3 $(U, \{A^k \mid k = 1, 2, \dots, I\})$ 是一个混合语义下的不完备多尺度信息系统, 对于数据填补后的 $S^*, k = 1, 2, \dots, I, B \subseteq A$, 记:

$$U/R_{B^k} = \{K_{B^k}(x) \mid x \in U\}$$

如下性质成立: $\forall x \in U$.

- 1) R_{B^k} 是自反的, 不一定是对称和传递的;
- 2) $R_{B^k} = \bigcap_{a \in B} R_{a^k}$;
- 3) 当 $C \subseteq B \subseteq A$ 时, $R_{A^k} \subseteq R_{B^k} \subseteq R_{C^k}$;
- 4) 当 $C \subseteq B \subseteq A$ 时, $K_{A^k}(x) \subseteq K_{B^k}(x) \subseteq K_{C^k}(x)$;
- 5) $R_{B^k} \subseteq R_{B^{k+1}}$, 其中 $k = 1, 2, \dots, I-1$;
- 6) $K_{B^k}(x) \subseteq K_{B^{k+1}}(x)$, 其中 $k = 1, 2, \dots, I-1$;
- 7) $U/R_{B^k} \subseteq U/R_{B^{k+1}}$, 其中 $k = 1, 2, \dots, I-1$.

定义 9 $X \subseteq U$, X 关于 R_{B^k} 的下近似集和上近似集定义为:

$$\begin{aligned} \underline{R_{B^k}}(X) &= \{x \in U \mid K_{B^k}(x) \subseteq X\} \\ \overline{R_{B^k}}(X) &= \{x \in U \mid K_{B^k}(x) \cap X \neq \emptyset\} \end{aligned}$$

集合 $BN_{B^k}(X) = \overline{R_{B^k}}(X) - \underline{R_{B^k}}(X)$ 称为 X 的 R_{B^k} 边界域.

由定理 3 和上述上、下近似的定义易得到如下定理:

定理 4 $(U, \{A^k \mid k=1, 2, \dots, I\})$ 是一个混合语义下的不完备多尺度信息系统, 对于数据填补后的 S^* , $k=1, 2, \dots, I$, $B \subseteq A$, 有如下性质成立: $\forall X \subseteq U, Y \subseteq U$.

- 1) $\overline{R_{B^k}}(X) = \sim \underline{R_{B^k}}(\sim X)$;
- 2) $\underline{R_{B^k}}(X) = \sim \overline{R_{B^k}}(\sim X)$;
- 3) $\underline{R_{B^k}}(\emptyset) = \overline{R_{B^k}}(\emptyset) = \emptyset$;
- 4) $\underline{R_{B^k}}(U) = \overline{R_{B^k}}(U) = U$;
- 5) $\underline{R_{B^k}}(X \cap Y) = \underline{R_{B^k}}(X) \cap \underline{R_{B^k}}(Y)$;
- 6) $\overline{R_{B^k}}(X \cup Y) = \overline{R_{B^k}}(X) \cup \overline{R_{B^k}}(Y)$;
- 7) $X \subseteq Y \Rightarrow \underline{R_{B^k}}(X) \subseteq \underline{R_{B^k}}(Y)$;
- 8) $X \subseteq Y \Rightarrow \overline{R_{B^k}}(X) \subseteq \overline{R_{B^k}}(Y)$;
- 9) $\underline{R_{B^k}}(X \cup Y) \supseteq \underline{R_{B^k}}(X) \cup \underline{R_{B^k}}(Y)$;
- 10) $\overline{R_{B^k}}(X \cap Y) \subseteq \overline{R_{B^k}}(X) \cap \overline{R_{B^k}}(Y)$;
- 11) $\underline{R_{B^k}}(X) \subseteq X \subseteq \overline{R_{B^k}}(X)$;
- 12) $\underline{R_{B^{k+1}}}(X) \subseteq \underline{R_{B^k}}(X)$, $k=1, 2, \dots, I-1$;
- 13) $\overline{R_{B^k}}(X) \subseteq \overline{R_{B^{k+1}}}(X)$, $k=1, 2, \dots, I-1$.

X 关于 R_{B^k} 的近似精度定义为:

$$\alpha_{B^k}(X) = \frac{|\underline{R_{B^k}}(X)|}{|\overline{R_{B^k}}(X)|}$$

X 关于 R_{B^k} 的粗糙度定义为:

$$\rho_{B^k}(X) = 1 - \alpha_{B^k}(X)$$

可得到不同的尺度下 X 的近似精度与粗糙度的关系.

定理 5 $(U, \{A^k \mid k=1, 2, \dots, I\})$ 是一个混合语义下的不完备多尺度信息系统, 对于数据填补后的 S^* , $k=1, 2, \dots, I-1$, $B \subseteq A$, 则 $\forall X \subseteq U$.

- 1) $\alpha_{B^{k+1}}(X) \leq \alpha_{R_{B^k}}(X)$;
- 2) $\rho_{B^k}(X) \leq \rho_{R_{B^{k+1}}}(X)$.

定理 5 表明, 尺度越小(细), 集合 X 的近似精度越高而粗糙度越小.

3 基于序贯三支决策模型的最优尺度选择

在第 2 节的工作基础之上, 本节将给出混合语义下不完备多尺度信息系统的序贯三支决策模型, 并基于此模型在混合语义下的不完备多尺度决策系统中选择不确定性最小的尺度作为最优尺度.

3.1 混合语义下的不完备多尺度信息系统的序贯三支决策模型

定义 10 $S = (U, A) = (U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\})$ 是一个混合语义下的不完备多尺度信息系统, 记 $A^k = \{a_1^k, a_2^k, \dots, a_m^k\}$. 对于数据填补后的 S^* , $X \subseteq U$ 关于 A^k 的三支决策(接受、拒绝和不确定)分别定义为 $ACP(A^k, X)$, $REJ(A^k, X)$ 和 $UNC(A^k, X)$. 其中 $ACP(A^k, X)$ 表示 U 中在特征关系 R_{A^k} 下一定属于 X 的对象集, $REJ(A^k, X)$ 表示 U 中在特征关系 R_{A^k} 下一定不属于 X 的对象集, $UNC(A^k, X)$ 表示 U 中在特征关系 R_{A^k} 下不确定是否属于 X 的对象集.

定理 6 设 $S = (U, A) = (U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\})$ 是一个混合语义下的不完备多尺度信息系统. 对于数据填补后的 S^* , $k=1, 2, \dots, I-1$, 根据 X 关于 A^k 的三支决策, 可得到 X 关于 A^{k+1} 的三支决策.

- 1) 接受: $ACP(A^{k+1}, X) = \{x \in ACP(A^k, X) \mid K_{A^{k+1}}(x) \subseteq X\}$;
- 2) 拒绝: $REJ(A^{k+1}, X) = \{x \in REJ(A^k, X) \mid K_{A^{k+1}}(x) \cap X = \emptyset\}$;
- 3) 不确定:

$$UNC(A^{k+1}, X) = UNC(A^k, X) \cup \{x \in ACP(A^k, X) \mid K_{A^{k+1}}(x) \subsetneq X\} \cup \\ \{x \in REJ(A^k, X) \mid K_{A^{k+1}}(x) \cap X \neq \emptyset\}$$

证 由定理 3 知 $K_{B^k}(x) \subseteq K_{B^{k+1}}(x)$:

1) 对于 $\forall x \in ACP(A^k, X)$, 始终满足 $K_{A^k}(x) \subseteq X$, 于是有 $K_{A^{k+1}}(x) \cap X \neq \emptyset$. 因此 $x \notin REJ(A^{k+1}, X)$; 若 $K_{A^{k+1}}(x) \subseteq X$ 则有 $x \in ACP(A^{k+1}, X)$.

2) 对于 $\forall x \in REJ(A^k, X)$, 始终满足 $K_{A^k}(x) \cap X = \emptyset$, 于是有 $K_{A^{k+1}}(x) \subsetneq X$. 因此 $x \notin ACP(A^{k+1}, X)$; 若 $K_{A^{k+1}}(x) \cap X = \emptyset$ 则有 $x \in REJ(A^{k+1}, X)$.

3) 对于 $\forall x \in UNC(A^k, X)$, 始终满足 $K_{A^k}(x) \subsetneq X$ 并且 $K_{A^k}(x) \cap X \neq \emptyset$, 于是有 $K_{A^{k+1}}(x) \subsetneq X$ 且 $K_{A^{k+1}}(x) \cap X \neq \emptyset$. 因此 $x \notin ACP(A^{k+1}, X)$ 且 $x \notin REJ(A^{k+1}, X)$.

由以上得到 1)2) 的证明;

对于 $\forall x \in ACP(A^k, X)$ 始终有 $K_{A^k}(x) \subseteq X$, 因此 $K_{A^{k+1}}(x) \cap X \neq \emptyset$, 若 $K_{A^{k+1}}(x) \subsetneq X$ 则有 $x \in UNC(A^{k+1}, X)$; 对于 $\forall x \in REJ(A^k, X)$ 始终有 $K_{A^k}(x) \cap X = \emptyset$, 因此 $K_{A^{k+1}}(x) \subsetneq X$, 若 $K_{A^{k+1}}(x) \cap X \neq \emptyset$ 则有 $x \in UNC(A^{k+1}, X)$; 对于 $\forall x \in UNC(A^k, X)$ 始终有 $K_{A^k}(x) \subsetneq X$ 并且 $K_{A^k}(x) \cap X \neq \emptyset$, 因此 $K_{A^{k+1}}(x) \subsetneq X$ 并且 $K_{A^{k+1}}(x) \cap X \neq \emptyset$, 即 $x \in UNC(A^{k+1}, X)$. 由此得到 3) 的证明.

由定理 6 知三支决策 $ACP(A^k, X)$, $REJ(A^k, X)$ 和 $UNC(A^k, X)$ 从 k 尺度到 $k+1$ 尺度是单调变化的, 变化过程如定理 7 所示.

定理 7 设 $S = (U, A) = (U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\})$ 是一个混合语义下的不完备多尺度信息系统. 对于数据填补后的 S^* , $k=1, 2, \dots, I-1$, 有:

- 1) $ACP(A^k, X) \supseteq ACP(A^{k+1}, X)$;
- 2) $REJ(A^k, X) \supseteq REJ(A^{k+1}, X)$;
- 3) $UNC(A^k, X) \subseteq UNC(A^{k+1}, X)$.

混合语义下的不完备多尺度信息系统的序贯三支决策模型是在尺度增加(粗化)的过程中对 $ACP(A^k, X)$ 和 $REJ(A^k, X)$ 的进一步决策, 使得属于 $ACP(A^k, X)$ 或 $REJ(A^k, X)$ 的对象可能会增加到 $UNC(A^{k+1}, X)$ 中. 三支决策从最细尺度到最粗尺度的整个动态过程如图 1 所示.

3.2 基于序贯三支决策模型的最优尺度选择算法

在不完备多尺度决策系统中, 尺度越细的决策系统的不确定性越小而在尺度越粗的决策系统中获取的规则越简单^[30]. 本节将利用序贯三支决策的不确定(UNC)量化决策系统的不确定性, 选择不确定性最小且尽可能粗的尺度作为最优尺度.

定义 11 一个混合语义下的不完备多尺度决策系统可以表示为:

$$S = (U, A \cup \{d\}) = (U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\} \cup \{d\})$$

其中 $(U, A) = (U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\})$ 是一个混合语义下的不完备多尺度信息系统. A 为条件属性集, $d \notin A$ 为一个单尺度属性, 称为决策属性, $d: U \rightarrow V_d$, V_d 为决策属性 d 的值域.

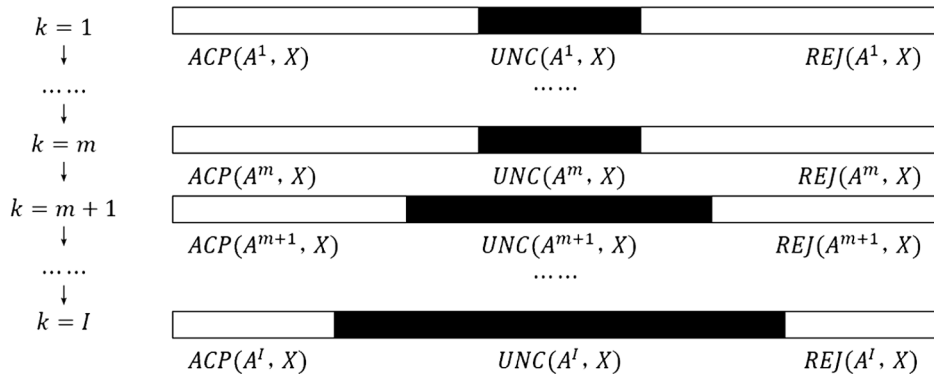


图1 不完备多尺度信息系统的序贯三支决策模型

定义 12 $d(x)$ 为 x 在决策属性 d 下的属性值, 关于决策属性 d 的等价关系定义为:

$$R_d = \{(x, y) \in U \times U \mid d(x) = d(y)\}$$

包含 x 的决策类 $[x]_{R_d} = \{y \in U \mid (x, y) \in R_d\}$. 由决策属性 d 导出的划分 $U/R_d = \{[x]_{R_d} \mid x \in U\}$.

定理 8 设 $S = (U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\} \cup \{d\})$ 是一个混合语义下的不完备多尺度决策系统, 记 $A^k = \{a_1^k, a_2^k, \dots, a_m^k\}$, 对于数据填补后的 S^* , $U/R_d = \{D_1, D_2, \dots, D_p\}$, $k=1, 2, \dots, I-1$, 根据 $D_i \in U/R_d$ 关于 A^k 的三支决策, 可得到 D_i 关于 A^{k+1} 的三支决策.

- 1) 接受: $ACP(A^{k+1}, D_i) = \{x \in ACP(A^k, D_i) \mid K_{A^{k+1}}(x) \subseteq D_i\}$;
- 2) 拒绝: $REJ(A^{k+1}, D_i) = \{x \in REJ(A^k, D_i) \mid K_{A^{k+1}}(x) \cap D_i = \emptyset\}$;
- 3) 不确定:

$$UNC(A^{k+1}, D_i) = UNC(A^k, D_i) \cup \{x \in ACP(A^k, D_i) \mid K_{A^{k+1}}(x) \not\subseteq D_i\} \cup \{x \in REJ(A^k, D_i) \mid K_{A^{k+1}}(x) \cap D_i \neq \emptyset\}$$

显然, 随着尺度增加, $UNC(A^k, D_i)$ 单调不减. 定理 6 与定理 8 的不同之处在于三支决策的目标对象不同. 具体来说, 前者利用不同尺度下特征关系导出的覆盖对目标概念进行序贯三支决策, 而后者则是决策属性导出的划分 U/R_d 的序贯三支决策.

定义 13 $S = (U, A \cup \{d\}) = (U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\} \cup \{d\})$ 是一个混合语义下的不完备多尺度决策系统, 记 $A^k = \{a_1^k, a_2^k, \dots, a_m^k\}$, 对于数据填补后的 S^* , $x \in U$, 记:

$$\partial_{A^k}(x) = \{d(y) \mid y \in K_{A^k}(x)\}$$

$\partial_{A^k}(x)$ 是 x 关于属性集 A 在第 k 尺度下的广义决策值.

定理 9 设 $S = (U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\} \cup \{d\})$ 是一个混合语义下的不完备多尺度决策系统. 对于数据填补后的 S^* , $U/R_d = \{D_1, D_2, \dots, D_p\}$. $\forall D_i \in U/R_d$, 若 $x \in UNC(A^k, D_i)$, 则有 $|\partial_{A^k}(x)| > 1$.

证 $\forall x \in UNC(A^k, D_i)$ 始终满足 $K_{A^k}(x) \not\subseteq D_i$ 并且 $K_{A^k}(x) \cap D_i \neq \emptyset$, 于是存在 $x_1, x_2 \in K_{A^k}(x)$ 和 $x_3 \in D_i$ 满足 $d(x_3) = d(x_1)$ 而 $d(x_3) \neq d(x_2)$, 因此 $d(x_2) \neq d(x_1)$, 也就是 $|\partial_{A^k}(x)| = |\{d(y) \mid y \in K_{A^k}(x)\}| > 1$.

定义 14 设 $S = (U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\} \cup \{d\})$ 是一个混合语义下的不完备多尺度决策系统. 对于数据填补后的 S^* , 记 $A^k = \{a_1^k, a_2^k, \dots, a_m^k\}$, $U/R_d = \{D_1, D_2, \dots, D_p\}$. 若 $\exists D_i \in U/R_d$ 使得 $UNC(A^k, D_i) \subset UNC(A^{k+1}, D_i)$ 并且 $\forall D_i \in U/R_d, \forall l < k$ 满足 $UNC(A^k, D_i) = UNC(A^l, D_i)$, 那么就称 k 为基于不确定性的最优尺度.

由定理 7 可以得到随着尺度的增加(粗化) $UNC(A^k, D_i)$ 单调不减. 通过定义 14 知, 当 k 为最优尺度时, 存在 $D_i \in U/R_d$ 使得 $UNC(A^k, D_i)$ 随着尺度增加而增大. 因此, 最优尺度是使得 $UNC(A^k, D_i)$ 最

小的所有尺度中最粗的尺度. 从不确定性最小即为最优的角度来说, 在最优尺度进行决策是有效的.

在尺度细化的过程中, 文献[29]选择不确定性不再改变的尺度作为最优尺度. 然而, 当增加的尺度是冗余的时, 不确定性很容易保持不变^[20], 此时选择的最优尺度的不确定性不是最小. 定义 14 给出的最优尺度选择方法弥补了此缺陷, 得到的最优尺度的不确定性总是最小. 对于一个混合语义下的不完备多尺度决策系统, 最优尺度选择步骤如算法 1 所示.

算法 1 混合语义下的不完备多尺度决策系统最优尺度选择

输入: 数据填补后混合语义下的不完备多尺度决策系统 $S^* = (U, A \cup \{d\}) = (U, \{a_j^k \mid k=1, 2, \dots, I, j=1, 2, \dots, m\} \cup \{d\})$;

输出: 最优尺度 k .

- 1) 计算决策属性的划分: $U/R_d = \{D_1, D_2, \dots, D_p\}$;
- 2) **for each** $i \in \{1, 2, \dots, p\}$
- 3) 计算 $ACP(A^1, D_i), REJ(A^1, D_i), UNC(A^1, D_i)$;
- 4) **end for**
- 5) $k = 1$;
- 6) **while** $k \leq I - 1$
- 7) $end_condition = 0$;
- 8) **for each** $i \in \{1, 2, \dots, p\}$
- 9) initialize $M = \emptyset$;
- 10) **for each** $x \in ACP(A^k, D_i) \cup REJ(A^k, D_i)$
- 11) **if** $K_{A^{k+1}}(x) \not\subseteq D_i$ **and** $K_{A^{k+1}}(x) \cap D_i \neq \emptyset$ **then**
- 12) $M \leftarrow M \cup \{x\}$;
- 13) **end for**
- 14) **if** $M \neq \emptyset$ **then**
- 15) $UNC(A^{k+1}, D_i) = UNC(A^k, D_i) \cup M$;
- 16) $end_condition = 1$;
- 17) **break**;
- 18) **else**
- 19) $UNC(A^{k+1}, D_i) = UNC(A^k, D_i)$;
- 20) **end for**
- 21) **if** $end_condition = 1$ **then**
- 22) **break**;
- 23) $k = k + 1$;
- 24) **end while**
- 25) **return** k .

算法 1 首先计算了由决策属性 d 导出的划分 U/R_d , 时间复杂度为 $O(n^2)$. 2—4 在第 1 尺度下计算所有决策类的三支决策, 时间复杂度是 $O(n \mid p \mid)$, 其中每次计算涉及特征集的计算, 而每个特征集的计算需要时间复杂度为 $O(nmI)$, 所以时间复杂度为 $O(n^2mI \mid p \mid)$. 接着根据序贯三支决策(UNC)计算基于不确定性的最优尺度, 有 3 层循环嵌套, 同样内循环涉及特征集的计算, 所以在最坏的情况下时间复杂度为 $O(n^2I^2m \mid p \mid)$. 所以总的时间复杂度为 $O(n^2I^2m \mid p \mid)$.

例 3 表 3 是一个完成数据补充后的不完备多尺度决策系统 $S^* = (U, A = \{a_j^k \mid k=1, 2, \dots, 5, j=1, 2, 3\} \cup \{d\})$, 其中 $U = \{x_1, x_2, \dots, x_{10}\}$, 关于决策属性的论域划分 $U/R_d = \{D_1, D_2\}$, $D_1 = \{x_1,$

$x_2, x_3, x_6, x_7\}, D_2 = \{x_4, x_5, x_8, x_9, x_{10}\}.$

表 3 补充数据后的不完备多尺度决策系统

U	a_1^1	a_1^2	a_1^3	a_1^4	a_1^5	a_2^1	a_2^2	a_2^3	a_2^4	a_2^5	a_3^1	a_3^2	a_3^3	a_3^4	a_3^5	d
x_1	1	I	A	H	Y	2	I	A	H	Y	2	I	A	H	Y	1
x_2	2	II	A	H	Y	3	II	B	H	Y	3	I	A	H	Y	1
x_3	3	II	A	H	Y	2	I	A	H	Y	$\emptyset$	$\emptyset$	$\emptyset$	M	N	1
x_4	4	III	B	S	Y	*	*	*	*	*	4	II	B	M	N	0
x_5	6	V	D	M	N	4	III	C	M	Y	5	III	B	M	N	0
x_6	4	III	B	S	Y	5	IV	C	M	Y	4	II	B	M	N	1
x_7	5	IV	C	M	N	4	III	C	M	Y	5	III	B	M	N	1
x_8	6	V	D	M	N	$\emptyset$	V	D	L	N	7	IV	C	M	N	0
x_9	7	VI	E	L	N	6	V	D	L	N	*	*	*	L	N	0
x_{10}	8	VI	E	L	N	7	V	D	L	N	8	V	D	L	N	0

根据算法 1, 从 S^* 中选择最优尺度的过程如下:

1) 在第 1 尺度下, 关于 $A^1 = \{a_1^1, a_2^1, a_3^1\}$ 有 $K_{A^1}(x_1) = \{x_1\}, K_{A^1}(x_2) = \{x_2\}, K_{A^1}(x_3) = \{x_3\}, K_{A^1}(x_4) = \{x_4, x_6\}, K_{A^1}(x_5) = \{x_5\}, K_{A^1}(x_6) = \{x_4, x_6\}, K_{A^1}(x_7) = \{x_7\}, K_{A^1}(x_8) = \{x_8\}, K_{A^1}(x_9) = \{x_9\}, K_{A^1}(x_{10}) = \{x_{10}\}$. 显然, $UNC(A^1, D_1) = \overline{R_{A^1}}(D_1) - \underline{R_{A^1}}(D_1) = \{x_4, x_6\}, UNC(A^1, D_2) = \{x_4, x_6\}$;

2) 在第 2 尺度下, 关于 $A^2 = \{a_1^2, a_2^2, a_3^2\}$ 有 $K_{A^2}(x_1) = \{x_1\}, K_{A^2}(x_2) = \{x_2\}, K_{A^2}(x_3) = \{x_3\}, K_{A^2}(x_5) = \{x_5\}, K_{A^2}(x_7) = \{x_7\}, K_{A^2}(x_8) = \{x_8\}, K_{A^2}(x_9) = \{x_9, x_{10}\}, K_{A^2}(x_{10}) = \{x_9, x_{10}\}$. 由定理 8 可得到 $UNC(A^2, D_1) = UNC(A^1, D_1) \cup \emptyset = \{x_4, x_6\}, UNC(A^2, D_2) = \{x_4, x_6\}$;

3) 在第 3 尺度下, 关于 $A^3 = \{a_1^3, a_2^3, a_3^3\}$, 同理可得 $UNC(A^3, D_1) = UNC(A^3, D_2) = \{x_4, x_6\}$;

4) 在第 4 尺度下, 关于 $A^4 = \{a_1^4, a_2^4, a_3^4\}$, 有 $UNC(A^4, D_1) = UNC(A^4, D_2) = \{x_4, x_5, x_6, x_7\}$.

因此有 $UNC(A^3, D_i) = UNC(A^2, D_i) = UNC(A^1, D_i) (i = 1, 2)$ 而 $UNC(A^3, D_i) \subset UNC(A^4, D_i) (i = 1, 2)$. 最后选择第 3 尺度作为 S^* 的最优尺度.

4 实验结果与分析

为了验证在混合语义下的不完备多尺度决策系统中的数据填补和模型扩展以及最优尺度选择算法的可行性与有效性, 本研究从 UCI 数据库中选出 5 个数据集进行数值模拟, 分别为 Seeds, Glass Identification, Wholesale Customers, Balance Scale, Car Evaluation, 数据集的详细信息如表 4 所示.

表 4 原始数据集信息

数据集	对象	条件属性	决策属性
Seeds	210	7	1
Glass Identification	214	9	1
Wholesale Customers (Region)	440	6	1
Balance Scale	625	4	1
Car Evaluation	1 728	6	1

为满足实验要求, 本研究借鉴文献[29]的方法将表 4 中的数据集重新构造为多尺度决策系统. 表 5 列出了经过预处理后的数据集信息. 在多尺度决策系统中随机丢失一定百分比的条件属性值, 从而得到不同缺失率的不完备多尺度决策系统. 若属性值的缺失使得所在属性的值域减小时, 将该未知属性值解释为缺

席型，否则解释为遗漏型。最后得到混合语义下的不完备多尺度决策系统。

表 5 预处理后的数据集信息

数据集	对象	条件属性	决策属性	尺度
Seeds	210	7	1	6
Glass Identification	214	9	1	6
Wholesale Customers (Region)	440	6	1	6
Balance Scale	625	4	1	5
Car Evaluation	1 728	6	1	5

为验证 2.2 和 2.3 小节中的数据填补和模型扩展的有效性，将文献[34]定义的广义不可分辨关系的分类精度 μ_R 作为评价标准。具体公式定义为：

$$\mu_R = \frac{\sum_{x \in U} \frac{|[x]_E \cap R(x)|}{|[x]_E \cup R(x)|}}{|U|}$$

其中，在完备信息系统 $S_1 = (U, A)$ 中， E 为 U 上的等价关系， $[x]_E$ 表示 x 的等价类；在 S_1 中引入一定百分比的缺失值后有不完备信息系统 $S_2 = (U', A)$ ， R 为 U' 上广义不可分辨关系(特征关系)， $R(x)$ 表示 x 的广义不可分辨对象集(特征集)。当不完备信息系统 S_2 导出的覆盖 U'/R 越接近完备信息系统 S_1 导出的划分 U/E ， S_2 的分类精度越高；当 S_2 导出的覆盖完全还原 S_1 的划分，则取最大值 1。

表 6~表 10 分别是 5 个数据集在 35%缺失率下 4 种阶段的单尺度分类精度，图 2~图 6 为 5 个数据集在 0~50%的缺失率下 4 种阶段的多尺度分类精度，多尺度分类精度即为所有单尺度分类精度的平均值。其中，CR 表示将特征关系直接引入不完备多尺度决策系统；CR-add 表示在引入特征关系的基础上对系统进行数据填补；CR-extend1 指在 CR-add 的基础上针对特征关系不满足偏序关系的第一次模型扩展，即通过粗尺度下已知属性值的不相等预判细尺度下存在未知属性值的两个对象是不相等的；CR-extend2 则是在 CR-extend1 的基础上针对缺席型未知属性值“ $\emptyset$ ”的不可比较导致不满足偏序关系而进行的第二次模型扩展。

表 6 Seeds 的单尺度分类精度

尺度 k	CR	CR-add	CR-extend1	CR-extend2
1	0.177	0.177	0.966	0.966
2	0.179	0.828	0.972	0.972
3	0.170	0.932	0.978	0.978
4	0.149	0.956	0.963	0.963
5	0.177	0.973	0.973	0.973
6	0.394	0.989	0.989	0.989

表 7 Glass Identification 的单尺度分类精度

尺度 k	CR	CR-add	CR-extend1	CR-extend2
1	0.276	0.276	0.963	0.968
2	0.253	0.85	0.963	0.963
3	0.173	0.895	0.957	0.957
4	0.169	0.947	0.958	0.958
5	0.193	0.97	0.976	0.976
6	0.404	0.976	0.976	0.976

表 8 Region 的单尺度分类精度

尺度 k	CR	CR-add	CR-extend1	CR-extend2
1	0.069	0.069	0.907	0.907
2	0.077	0.455	0.907	0.907
3	0.081	0.795	0.907	0.907
4	0.178	0.935	0.959	0.959
5	0.355	0.971	0.98	0.98
6	0.66	0.988	0.988	0.988

表 9 Balance Scale 的单尺度分类精度

尺度 k	CR	CR-add	CR-extend1	CR-extend2
1	0.013	0.013	0.608	0.608
2	0.013	0.177	0.608	0.608
3	0.073	0.691	0.934	0.934
4	0.074	0.863	0.934	0.934
5	0.199	0.968	0.968	0.968

表 10 Car Evaluation 的单尺度分类精度

尺度 k	CR	CR-add	CR-extend1	CR-extend2
1	0.008	0.008	0.598	0.598
2	0.008	0.145	0.598	0.598
3	0.049	0.632	0.921	0.921
4	0.048	0.854	0.921	0.921
5	0.131	0.959	0.959	0.959

具体地, 针对 Seeds 数据集, 实验结果如表 6 和图 2 所示. 表 6 中第 1 尺度下数据得不到填补, 因此 CR-add 的分类精度等于 CR 的分类精度, 其余尺度下的分类精度 CR-add 大于 CR; 在第 6 尺度下, CR-extend1 与 CR-add 的分类精度相同, 其余尺度下 CR-extend1 大于等于 CR-add 的分类精度; CR-extend2 与 CR-extend1 的分类精度保持不变, 对于 Glass Identification 数据集, CR-extend2 相对于 CR-extend1 的分类精度在尺度 1 下增高 0.005, 总体而言变化不大. 针对图 2 中的多尺度分类精度, 在 0~50% 的缺失率下, CR-add, CR-extend1 相对于前一阶段都得到了或多或少的提升, CR-extend2 和 CR-extend1 持平, 即因缺席型未知属性值“ $\emptyset$ ”的不可比较, 未保持偏序关系的模型扩展对分类精度影响不大. 表 7~表 10 以及图 3~图 6 分别是其他 4 个数据集的分类精度结果, 同样符合 Seeds 数据集的分析结论. 因此进行数据填补和模型扩展后(CR-extend2)的分类精度相对于直接引入特征关系(CR)的分类精度更高, 有利于后续知识获取达到更高的确定性, 验证了其有效性.

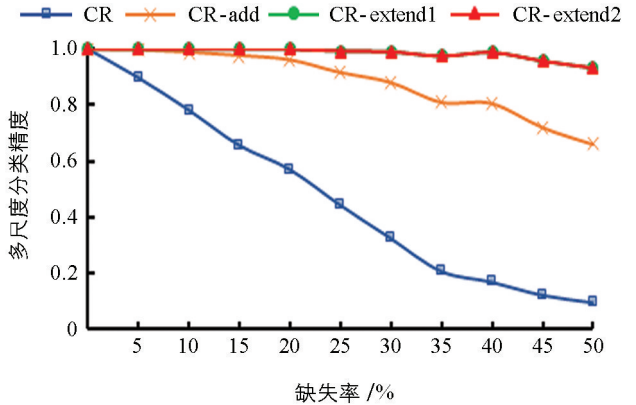


图 2 Seeds 的多尺度分类精度

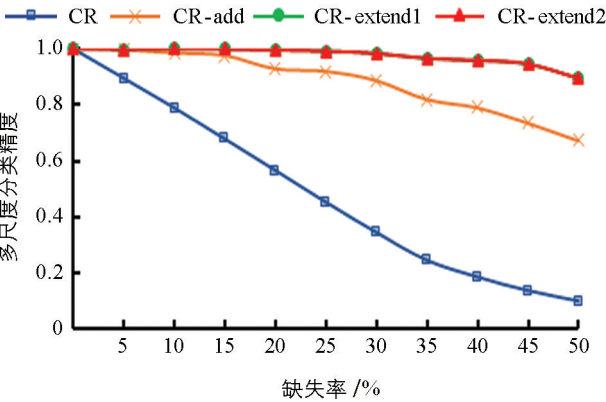


图 3 Glass Identification 的多尺度分类精度

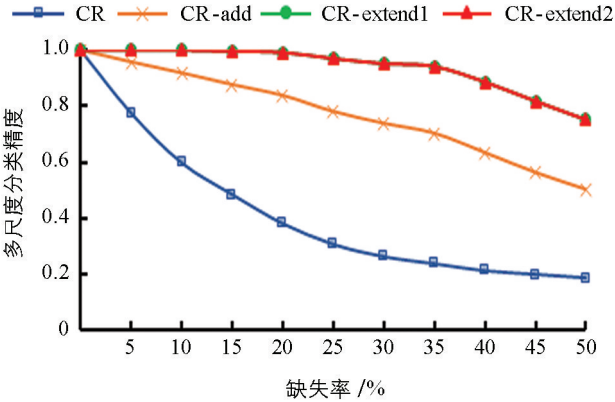


图 4 Region 的多尺度分类精度

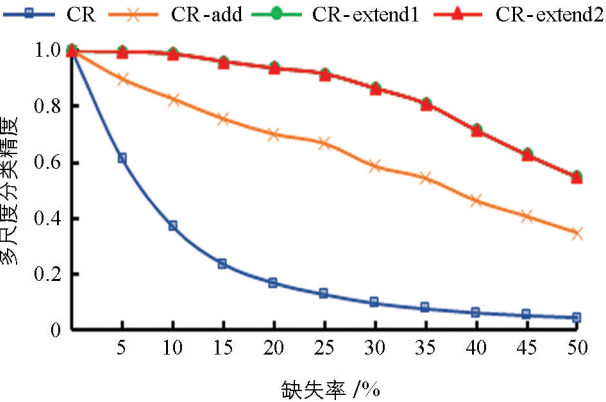


图 5 Balance Scale 的多尺度分类精度

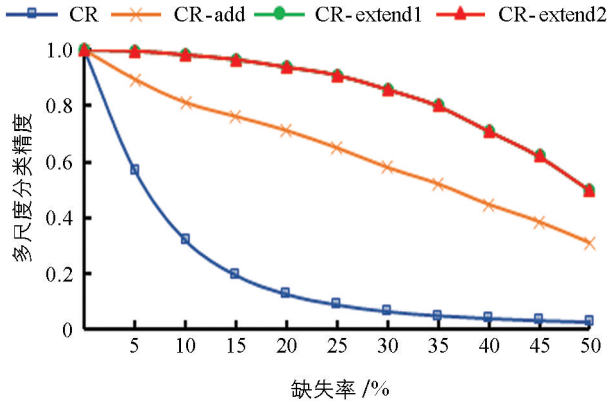


图 6 Car Evaluation 的多尺度分类精度

表 11 是 5 个数据集在 5% 的缺失率下, 经数据填补和模型扩展后算法 1 的最优尺度实验结果. 其中 OLS 是算法 1 得到的基于不确定性的最优尺度. HC'OLS 表示文献[29]中定义的最优尺度, 其标准是随着尺度变细, 不确定性不变的尺度即为最优尺度. 从实验结果可以看出, 随着尺度级数的变小(细化), 不确定性存在保持不变后减小的情况. 对于 Seeds, Balance Scale 和 Car Evaluation 数据集, HC'OLS 的不确定性不是最小的, 而本研究提出的方法可得到不确定性最小的最优尺度.

表 11 基于序贯三支决策的最优尺度

数据集	大小	UNC 占 U 的百分比/%						OLS	HC'OLS
		尺度 1	尺度 2	尺度 3	尺度 4	尺度 5	尺度 6		
Seeds	$210 \times (7 \times 6 + 1)$	0	0	0	5.71	5.71	70	3	5
Glass Identification	$214 \times (9 \times 6 + 1)$	0.93	0.93	9.34	20.56	51.86	77.10	2	2
Region	$440 \times (6 \times 6 + 1)$	40.90	40.90	40.90	64.09	78.63	93.40	3	3
Balance Scale	$625 \times (4 \times 5 + 1)$	0.64	0.64	48.64	48.64	88.48	\	2	4
Car Evaluation	$1\,728 \times (6 \times 5 + 1)$	0.34	0.34	29.62	29.62	41.66	\	2	4

5 结论

本研究分析发现目前研究中对不完备多尺度信息系统的定义存在局限, 因此定义混合语义下的不完备多尺度信息系统, 即未知属性值可以是任意的语义解释(遗漏和缺席型未知属性值)并且对象在多尺度属性下缺失时不局限于全部缺失. 引入特征关系以处理此类更为一般的不完备多尺度信息系统, 为保持尺度间偏序关系对混合语义下的不完备多尺度信息系统进行数据填补和特征关系的模型扩展, 同时给出在不同尺

度下信息粒度的表示及其相互关系,定义了集合的上、下近似集概念,并讨论了它们的性质.基于上述工作,在混合语义下的不完备多尺度决策系统中建立序贯三支决策模型,基于该模型选择不确定性最小的尺度作为最优尺度.本研究所定义的最优尺度与已有的方法略有不同,选择不确定性最小的尺度作为最优尺度是直观且容易理解的,并且无需从一致和不一致的不完备多尺度决策系统两个角度分别讨论最优尺度.通过实验证明,经数据填补和模型扩展后,不完备多尺度决策系统的分类精度得到提升,并且选择的最优尺度是所有尺度中不确定性最小的.

参考文献:

- [1] ZADEH L A. Fuzzy Sets and Information Granularity [J]. *Advances in Fuzzy Set Theory and Applications*, 1979, 11: 3-18.
- [2] PAWLAK Z. *Rough Sets: Theoretical Aspects of Reasoning about Data* [M]. Berlin: Springer Science and Business Media, 1991.
- [3] 周涛,陆惠玲,任海玲,等.基于粗糙集的属性约简算法综述[J].*电子学报*, 2021, 49(7): 1439.
- [4] ZHU W H, ZHANG W, FU Y Q. An Incomplete Data Analysis Approach Using Rough Set Theory [C] //2004 International Conference on Intelligent Mechatronics and Automation, Proceedings. IEEE, 2004: 332-338.
- [5] BAI X L, ZHANG M C, WU Q T, et al. A Novel Data Filling Algorithm for Incomplete Information System Based on Valued Limited Tolerance Relation [J]. *International Journal of Database Theory and Application*, 2015, 8(6): 149-164.
- [6] STEFANOWSKI J, TSOUKIAS A. On the Extension of Rough Sets under Incomplete Information [C] //International Workshop on Rough Sets, Fuzzy Sets, Data Mining and Granular-Soft Computing. Springer, Berlin, Heidelberg, 1999: 73-81.
- [7] KRYSZKIEWICZ M. Rough Set Approach to Incomplete Information Systems [J]. *Information Sciences*, 1998, 112: 39-49.
- [8] GRZYMALA-BUSSE J W. Rough Set Strategies to Data with Missing Attribute Values [M] //Foundations and Novel Approaches in Data Mining. Springer, Berlin, Heidelberg, 2006: 197-212.
- [9] CLARK P G, GRZYMALA-BUSSE J W, RZASA W. Consistency of Incomplete Data [J]. *Information Sciences*, 2015, 322: 197-222.
- [10] CLARK P G, GAO C, GRZYMALA-BUSSE J W, et al. Characteristic Sets and Generalized Maximal Consistent Blocks in Mining Incomplete Data [J]. *Information Sciences*, 2018, 453: 66-79.
- [11] CHEN Y X, ZHU P. Extending Characteristic Relations on An Incomplete Data Set by the Three-way Decision Theory [J]. *International Journal of Approximate Reasoning*, 2020, 119: 108-121.
- [12] QIAN Y H, LIANG J Y, YAO Y Y, et al. MGRS: A Multi-Granulation Rough Set [J]. *Information Sciences*, 2010, 180: 949-970.
- [13] ZHU P F, HU Q H, ZUO W M, et al. Multi-Granularity Distance Metric Learning via Neighborhood Granule Margin Maximization [J]. *Information Sciences*, 2014, 282: 321-331.
- [14] WU W Z, LEUNG Y. Theory and Applications of Granular Labelled Partitions in Multi-Scale Decision Tables [J]. *Information Sciences*, 2011, 181: 3878-3897.
- [15] WU W Z, LEUNG Y. Optimal Scale Selection for Multi-Scale Decision Tables [J]. *International Journal of Approximate Reasoning*, 2013, 54(8): 1107-1129.
- [16] GU S M, WU W Z. On Knowledge Acquisition in Multi-Scale Decision Systems [J]. *International Journal of Machine Learning and Cybernetics*, 2013, 4(5): 477-486.
- [17] SHE Y H, LI J H, YANG H L. A Local Approach to Rule Induction in Multi-scale Decision Tables [J]. *Knowledge-Based Systems*, 2015, 89: 398-410.
- [18] CHEN D X, LI J J, LIN R D, et al. Information Entropy and Optimal Scale Combination in Multi-Scale Covering Decision Systems [J]. *IEEE Access*, 2020, 8: 182908-182917.

[19] 郑嘉文, 吴伟志, 包菡, 等. 基于熵的多尺度决策系统的最优尺度选择 [J]. 南京大学学报(自然科学版), 2021, 57(1): 130-140.

[20] ZHANG X Q, ZHANG Q H, CHENG Y L, et al. Optimal Scale Selection by Integrating Uncertainty and Cost-Sensitive Learning in Multi-Scale Decision Tables [J]. International Journal of Machine Learning and Cybernetics, 2020, 11(1): 1-20.

[21] SHE Y H, ZHAO Z J, HU M T, et al. On Selection of Optimal Cuts in Complete Multi-Scale Decision Tables [J]. Artificial Intelligence Review, 2021(11): 1-24.

[22] SHE Y H, QIAN Z H, HE X L, et al. On Generalization Reducts in Multi-Scale Decision Tables [J]. Information Sciences, 2021, 555: 104-124.

[23] WANG H R, LI W T, ZHAN T, et al. Multi-Granulation-Based Optimal Scale Selection in Multi-scale Information Systems [J]. Computers and Electrical Engineering, 2021, 92: 107107.

[24] LI F, HU B Q. A New Approach of Optimal Scale Selection to Multi-Scale Decision Tables [J]. Information Sciences, 2017, 381: 193-208.

[25] LI F, HU B Q, WANG J. Stepwise Optimal Scale Selection for Multi-scale Decision Tables via Attribute Significance [J]. Knowledge-Based Systems, 2017, 129: 4-16.

[26] WU W Z, LEUNG Y. A Comparison Study of Optimal Scale Combination Selection in Generalized Multi-Scale Decision Tables [J]. International Journal of Machine Learning and Cybernetics, 2019, 11(12): 961-972.

[27] 牛东蓓, 吴伟志, 李同军. 广义多尺度决策系统中基于可变精度的最优尺度组合 [J]. 模式识别与人工智能, 2019, 32(11): 965-974.

[28] 吴伟志, 庄宇斌, 谭安辉, 等. 不协调广义多尺度决策系统的尺度组合 [J]. 模式识别与人工智能, 2018, 31(6): 485-494.

[29] HAO C, LI J H, FAN M, et al. Optimal Scale Selection in Dynamic Multi-Scale Decision Tables Based on Sequential Three-Way Decisions [J]. Information Sciences, 2017, 415: 213-232.

[30] WU W Z, QIAN Y H, LI T J, et al. On Rule Acquisition in Incomplete Multi-Scale Decision Tables [J]. Information Sciences, 2017, 378: 282-302.

[31] 吴伟志, 陈颖, 徐优红, 等. 协调的不完备多粒度标记决策系统的最优粒度选择 [J]. 模式识别与人工智能, 2016, 29(2): 108-115.

[32] 吴伟志, 杨丽, 谭安辉, 等. 广义不完备多粒度标记决策系统的粒度选择 [J]. 计算机研究与发展, 2018, 55(6): 1263-1272.

[33] 顾沈明, 顾金燕, 吴伟志, 等. 不完备多粒度决策系统的局部最优粒度选择 [J]. 计算机研究与发展, 2017, 54(7): 1500-1509.

[34] WANG GY, GUAN L H, WU W Z, et al. Data-Driven Valued Tolerance Relation Based on the Extended Rough Set [J]. Fundamenta Informaticae, 2014, 132(3): 349-363.

责任编辑 包颖
崔玉洁