



广义多尺度决策表最优尺度选择的快速算法

张晓燕, 黄雨阳

(西南大学人工智能学院, 重庆 400715)

摘要: 广义多尺度决策表的条件属性和决策属性都具有多个尺度. 最优尺度能够将较粗的条件属性与较细的决策属性相结合, 达到效率与精度的平衡. 然而现有的最优尺度选择算法计算效率较低. 为此, 提出一种最优尺度选择的快速算法. 首先探讨最优尺度的一些性质和边界域的变化情况, 给出判断边界域是否相等的条件以及最优尺度的等价定义; 然后一种快速算法. 最后通过数值实验表明其相较于现有算法速度更快, 能有效解决最优尺度选择问题, 该算法在计算效率方面取得了显著的提升, 实现了在较短的时间内得出最佳结果的目标.

关键词: 最优尺度选择; 粒计算; 粗糙集; 广义多尺度决策表

中图分类号: TP18 文献标志码: A 文章编号: 2095-7122(2024)02-0001-07

A fast algorithm for optimal scale selection of generalized multi-scale decision tables

ZHANG Xiaoyan, HUANG Yuyang

(School of Artificial Intelligence, Southwest University, Chongqing 400715, China)

Abstract: Each attribute of generalized multi-scale decision tables has multiple scales, whether it is a conditional attribute or a decision attribute. The optimal scale combines the coarser condition attributes with the finer decision attributes, so as to achieve a balance between efficiency and accuracy. However, the existing optimal scale selection algorithms lack computational efficiency. In this paper, a fast algorithm for optimal scale selection is proposed. First, some properties of the optimal scale and the variation of boundary region are discussed. Second, the conditions for judging the equality of boundary region and the equivalent definition of the optimal scale are given. Then a fast algorithm is proposed. Finally, numerical experiments show that the algorithm is faster than the existing algorithm and can effectively solve the optimal scale selection problem. The algorithm has achieved a significant improvement in computational efficiency. And it achieves the goal of obtaining the best results in a relatively short time.

Key words: optimal scale selection; granular computing; rough set; generalized multi-scale decision tables

粒计算(granular computing)是一种计算模型,旨在通过模拟人类思维和认知过程来解决复杂问题. 针对特定的应用背景,人们提出了不同的粒计算模型,如模糊集^[1]、三支决策^[2]和概念认

收稿日期: 2024-04-24

基金项目: 国家自然科学基金项目(12371465); 重庆市自然科学基金项目(CSTB2023NSCQ-MSX1063)

作者简介: 张晓燕(1979—),女,山西怀仁人,博士,教授.

知学习^[3]等.粗糙集理论^[4]在粒计算的研究中起着重要的作用.粗糙集的核心思想是基于信息的粗糙度,通过区分不同属性之间的重要性和冗余性,进行数据的约简和分类.粗糙集理论在机器学习、数据挖掘和模式识别等领域中得到了广泛应用.它可以用于特征选择^[5]、多属性决策^[6]、信息融合^[7]等任务,帮助人们理解和利用数据中的信息,支持决策和推理过程.

随着时代的进步和社会的发展,人们需要在不同尺度上处理和分析数据,传统的单尺度粗糙集模型已经远远不能满足实际应用的需要.为此,吴伟志等^[8]提出了多尺度粗糙集模型,并讨论了此模型下的最优尺度选择和规则提取问题.基于三支决策的思想,李金海等^[9]研究了动态多尺度决策表的最优尺度动态更新问题.考虑到吴伟志提出的多尺度决策表对每个条件属性要求相同的尺度数过于苛刻,胡宝清等^[9-10]提出了一种推广的多尺度粗糙集模型,研究了每个条件属性具有不同尺度个数的广义多尺度粗糙集模型.黄哲煌等^[11]进一步放宽了对决策属性的限制,提出了具有多尺度决策属性的广义多尺度粗糙集模型.在最优尺度选择中,从多个单尺度决策表中选择最优的单尺度决策表,以获得所考虑问题的最优结果.随着属性个数和每个属性的尺度数的增加,尺度呈爆炸式增长,这使得计算非常耗时.为了提高计算效率,张晓燕等^[12]提出了两种最优尺度选择算法,分别可以找到一个最优尺度和所有最优尺度.近年来,多尺度粗糙集的研究受到了广泛的关注,但这些文章对最优尺度选择问题的讨论主要集中在理论层面,或是给出了最优尺度的不同定义^[13],或是给出了最优尺度保持不变的充要条件^[14],缺乏可行的最优尺度选择算法.

针对现有的最优尺度选择算法,首先,深入探究了广义多尺度决策表的最优尺度的性质;其次,讨论了边界域的变化情况,并给出了判断边界域相等的等价条件和边界域不等的充分条件;最后,提出了一种最优尺度选择的快速算法,并通过数值实验比较算法的运行时间和分类精度,验证了算法的有效性.

1 预备知识

在本节中,给出了广义多尺度决策表的定义,并探讨了一些相关的性质.

定义 1^[12] 二元组 $S=(U, AT \cup D)$ 称为一个广义多尺度决策表,其中 $(U, AT) = (U, \{a_j^k | k=1, 2, \dots, I_j; j=1, 2, \dots, m\})$ 为一个多尺度信息表. $D = \{d\}$ 是一个非空有限决策属性集合且 d 有 n 个尺度,记作 $\{d^t | t=1, 2, \dots, n\}$.

为了方便表示,下文都使用 $S=(U, AT \cup D) = (U, \{a_j^k | k=1, 2, \dots, I_j; j=1, 2, \dots, m\} \cup \{d^t | t=1, 2, \dots, n\})$ 来表示广义多尺度决策表.

命题 1^[12] 设 $S=(U, AT \cup D) = (U, \{a_j^k | k=1, 2, \dots, I_j; j=1, 2, \dots, m\} \cup \{d^t | t=1, 2, \dots, n\})$ 为一个广义多尺度决策表.若 $D_t = \{d^t\}, L_1 = (l_1^1, l_1^2, \dots, l_1^m) \in \mathcal{L}, L_2 = (l_2^1, l_2^2, \dots, l_2^m) \in \mathcal{L}$, 则有以下性质成立:

- 1) $L_1 < L_2 \Rightarrow \text{BND}(AT^{L_1}, D_t) \subseteq \text{BND}(AT^{L_2}, D_t), t=1, 2, \dots, n;$
- 2) $t_1 < t_2 \Rightarrow \text{BND}(AT^{L}, D_{t_1}) \supseteq \text{BND}(AT^{L}, D_{t_2});$
- 3) 若令 $L_0 = (1, 1, \dots, 1)$, 则有 $\text{BND}(AT^{L_0}, D_n) \subseteq \text{BND}(AT^L, D_t), t=1, 2, \dots, n, L \in \mathcal{L};$

4)若令 $L' = (I_1, I_2, \dots, I_m)$, 则有 $\text{BND}(AT^{L'}, D_1) \supseteq \text{BND}(AT^L, D_t), t = 1, 2, \dots, n, L \in \mathcal{L}$.

2 最优尺度选择及其快速算法

本节探讨了最优尺度的一些性质,提出了一种最优尺度选择的快速算法.

定义2 设 $S = (U, AT \cup D) = (U, \{a_j^k | k = 1, 2, \dots, I_j; j = 1, 2, \dots, m\} \cup \{d^t | t = 1, 2, \dots, n\})$ 为一个广义多尺度决策表. 若满足以下两条性质: 1) $\text{BND}(AT^L, D_t) = \text{BND}(AT^{L_0}, D_n)$; 2) 对于任意的 $H = (h_1, h_2, \dots, h_m) \in \mathcal{L}$ 和 $t' = \{1, 2, \dots, n\}$, 当 $L < H$ 且 $t \leq t'$ 或 $L = H$ 且 $t < t'$ 时, 都有 $\text{BND}(AT^{L_0}, D_n) \subset \text{BND}(AT^{H'}, D_{t'})$. 则称 $Q = (L, t) = (\{l_1, l_2, \dots, l_m\}, t)$ 是广义多尺度决策表 S 的最优尺度.

边界域是随着尺度的变化而变化. 命题2给出了边界域的变化范围,并指出了两种特殊情况.

命题2 设 $S = (U, AT \cup D) = (U, \{a_j^k | k = 1, 2, \dots, I_j; j = 1, 2, \dots, m\} \cup \{d^t | t = 1, 2, \dots, n\})$ 为一个广义多尺度决策表. 令 $\text{UNC} = \text{BND}(AT^L, D_1) - \text{BND}(AT^{L_0}, D_n)$, 则有以下性质成立:

- 1) $\emptyset \subseteq \text{UNC} \subseteq U$;
- 2) 若 $\text{UNC} = U$, 则有 $\text{BND}(AT^L, D_1) = U, \text{BND}(AT^{L_0}, D_n) = \emptyset$;
- 3) 若 $\text{UNC} = \emptyset$, 则有 $\text{BND}(AT^L, D_1) = \text{BND}(AT^{L_0}, D_n)$.

证明 因为 $\emptyset \subseteq \text{BND}(AT^{L_0}, D_n) \subseteq U$ 且 $\emptyset \subseteq \text{BND}(AT^L, D_1) \subseteq U$, 所以 $\emptyset \subseteq \text{UNC} \subseteq U$.

由此可知, 当 $\text{UNC} = U$ 时, $\text{BND}(AT^L, D_1) = U, \text{BND}(AT^{L_0}, D_n) = \emptyset$; 当 $\text{UNC} = \emptyset$ 时, $\text{BND}(AT^L, D_1) = \text{BND}(AT^{L_0}, D_n)$.

随着尺度的变粗,论域 U 中的部分元素从正域落到边界域. 图1显示了尺度由细到粗时边界域的变化情况. 接下来的定理1和定理2给出了确定边界域是否相等的简单方法.

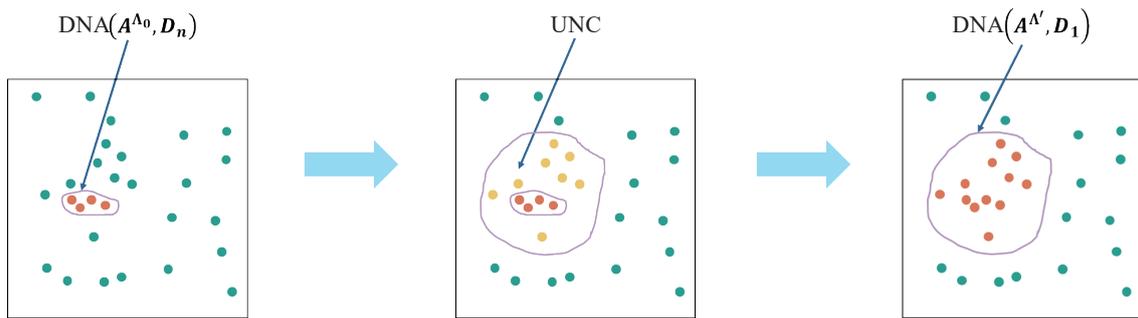


图1 边界域的变化

Fig.1 The change of the boundary region

定理1 设 $S = (U, AT \cup D) = (U, \{a_j^k | k = 1, 2, \dots, I_j; j = 1, 2, \dots, m\} \cup \{d^t | t = 1, 2, \dots, n\})$ 为一个广义多尺度决策表. $\text{BND}(AT^L, D_t) = \text{BND}(AT^{L_0}, D_n)$ 当且仅当对于任意的 $x \in \text{UNC}$ 都有

$$[x]_{AT^L} \subseteq [x]_{d^t}.$$

证明 若 $BND(AT^L, D_i) = BND(AT^{L_0}, D_n)$, 则对任意 $x \in UNC - BND(AT^{L_0}, D_n)$ 都有 $[x]_{AT^L} \subseteq [x]_{d^t}$.

由于 $BND(AT^L, D_1) \subseteq U$, 故有 $BND(AT^L, D_1) - BND(AT^{L_0}, D_n) \subseteq U - BND(AT^{L_0}, D_n)$.

因此, $UNC \subseteq U - BND(AT^{L_0}, D_n)$, 从而对任意 $x \in UNC$ 都有 $[x]_{AT^L} \subseteq [x]_{d^t}$.

由命题1可知, $BND(AT^{L_0}, D_n) \subseteq BND(AT^L, D_i) \subseteq BND(AT^L, D_1)$, 从而 $\emptyset \subseteq BND(AT^L, D_i) - BND(AT^{L_0}, D_n) \subseteq BND(AT^L, D_1) - BND(AT^{L_0}, D_n)$.

因为 $BND(AT^L, D_i) - BND(AT^{L_0}, D_n) = \emptyset$ 而且 $UNC = BND(AT^L, D_1) - BND(AT^{L_0}, D_n)$, 所以 $\emptyset \subseteq BND(AT^L, D_i) - BND(AT^{L_0}, D_n) \subseteq UNC$.

由此可知, 若对于任意的 $x \in UNC$ 都有 $[x]_{AT^L} \subseteq [x]_{d^t}$, 则对于任意的 $x \in BND(AT^L, D_i) - BND(AT^{L_0}, D_n)$ 都有 $[x]_{AT^L} \subseteq [x]_{d^t}$.

因此, $BND(AT^L, D_i) = BND(AT^{L_0}, D_n)$.

定理1给出了判定边界域相等的充要条件. 接下来给出判断边界域不相等的充分条件.

定理2 设 $S = (U, AT \cup D) = (U, \{a_j^k | k=1, 2, \dots, I_j; j=1, 2, \dots, m\} \cup \{d^t | t=1, 2, \dots, n\})$ 为一个广义多尺度决策表. 若存在 $x \in UNC$ 使得 $[x]_{AT^L} \not\subseteq [x]_{d^t}$, 则 $BND(AT^{L_0}, D_n) \subset BND(AT^L, D_i)$.

证明 若存在 $x \in UNC$ 使得 $[x]_{AT^L} \not\subseteq [x]_{d^t}$, 则 $x \notin BND(AT^{L_0}, D_n)$ 且 $x \in BND(AT^L, D_i)$. 因此, $BND(AT^{L_0}, D_n) \subset BND(AT^L, D_i)$.

根据上述命题和定理可以给出和定义2等价的定义3.

定义3 设 $S = (U, AT \cup D) = (U, \{a_j^k | k=1, 2, \dots, I_j; j=1, 2, \dots, m\} \cup \{d^t | t=1, 2, \dots, n\})$ 为一个广义多尺度决策表. 若满足以下两条性质: 1) 对于任意的 $x \in UNC$ 都有 $[x]_{AT^L} \subseteq [x]_{d^t}$; 2) 对于任意的 $H = (h_1, h_2, \dots, h_m) \in \mathcal{L}$ 和 $t' = \{1, 2, \dots, n\}$, 当 $L < H$ 且 $t \leq t'$ 或 $L = H$ 且 $t < t'$ 时, 存在 $x \in UNC$ 使得 $[x]_{AT^L} \not\subseteq [x]_{d^t}$. 则称 $Q = (L, t) = (\{l_1, l_2, \dots, l_m\}, t)$ 是广义多尺度决策表 S 的最优尺度.

算法1 判断 $BND(AT^{L_0}, D_n)$ 和 $BND(AT^L, D_1)$ 是否相等

Input: 一个广义多尺度决策系统 $S = (U, AT \cup D)$

Output: 判断结果 Judgement

1: 计算 $BND(AT^{L_0}, D_n)$ 和 $BND(AT^L, D_1)$;

2: 令 $UNC = BND(AT^L, D_1) - BND(AT^{L_0}, D_n)$, Judgement = True;

3: for each $x \in UNC$ do

4: if $[x]_{AT^L} \not\subseteq [x]_{d^t}$ then

5: Judgment = False;

6: break;

7: end

8: end

根据定义3,提出了一种判断 $\text{BND}(AT^L, D_n)$ 和 $\text{BND}(AT^U, D_1)$ 是否相等的算法,该算法可为最优尺度选择算法的内层算法,其最坏时间复杂度为 $O(|\text{UNC}|^2)$.文献[12]中提出的最优尺度选择算法的时间复杂度为 $O\left(\left(\prod_{j=1}^m I_j \times n\right) \times |U|^2\right)$,其内层算法的最坏时间复杂度为 $O(|U|^2)$.本文提出的算法1可以将最优尺度选择算法的时间复杂度降低为 $O\left(\left(\prod_{j=1}^m I_j \times n\right) \times |\text{UNC}|^2\right)$,得到一个运行结果相同但计算效率更高的最优尺度选择快速算法.

3 实验分析

本节选择来自UCI的8个公开数据集,验证上节提出的算法1的有效性,详细情况见表1.本节中采用文献[12]中构造广义多尺度决策表的方法将标准数据集进行转化.

表1 数据集说明
Tab.1 Description of datasets

| 编号 | 数据集 | 属性数 | 对象数 | 决策类 |
|----|------------------------------|-----|--------|-----|
| 1 | Iris | 4 | 150 | 3 |
| 2 | Wholesale customers(region) | 6 | 440 | 3 |
| 3 | Car Evaluation | 6 | 1 728 | 4 |
| 4 | Nursery | 8 | 12 960 | 5 |
| 5 | Abalone | 8 | 4 177 | 9 |
| 6 | Shill Bidding | 9 | 6 321 | 2 |
| 7 | Contraceptive Method Choice | 9 | 1 473 | 3 |
| 8 | Estimation of obesity levels | 16 | 2 111 | 7 |

接下来将算法1分别与文献[12]、文献[11]和文献[9]中的最优尺度选择算法进行运行时间的比较,结果如表2和图2所示.

表2 算法在8个数据集上的运行时间
Tab.2 The runtime of algorithms on 8 datasets

| 编号 | 数据集 | 运行时间/s | | | |
|----|------------------------------|--------------|--------|-----------|-----------|
| | | 算法1 | 文献[12] | 文献[11] | 文献[9] |
| 1 | Iris | 1.32 | 2.52 | 14.31 | 13.84 |
| 2 | Wholesale customers(region) | 0.21 | 0.60 | 140.86 | 142.89 |
| 3 | Car Evaluation | 7.24 | 10.35 | 82.68 | 80.56 |
| 4 | Nursery | 24.26 | 144.84 | 1 605.35 | 1 188.74 |
| 5 | Abalone | 1.01 | 4.08 | 3 164.57 | 3 368.16 |
| 6 | Shill Bidding | 13.48 | 250.04 | 1 185.04 | 1 097.40 |
| 7 | Contraceptive Method Choice | 0.50 | 2.05 | 200.78 | 229.64 |
| 8 | Estimation of obesity levels | 14.02 | 23.77 | 29 118.08 | 42 730.95 |

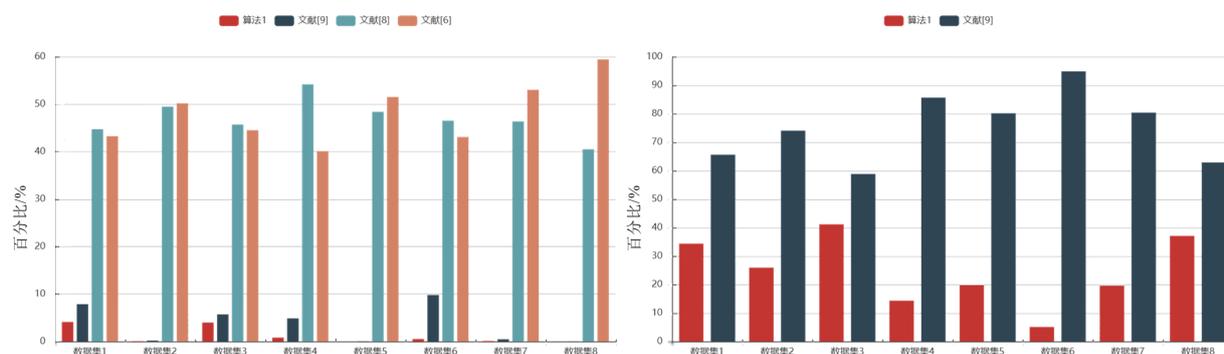


图2 运行时间的百分比柱状图
Fig.2 Percentage histogram of run time

从表2和图2可以看出算法1在所有数据集上的运行时间都小于其他的算法.相较于文献[11]和文献[9],算法1的计算效率提高了几十倍,同时比文献[12]的计算效率也有提升.在属性数或对象数较多的数据集上,算法1的优势比较突出.例如在数据集4、5、6、9上,算法1的运行时间都在30 s内.在运行时间最长的数据集8上,算法1的运行时间也不超过500 s.从实验结果来看,算法1的计算效率是最高的.

因为上述最优尺度选择算法所得到最优尺度结果相同,所以在接下来的实验中应用SVM分类器和CART分类器来比较最细尺度、最粗尺度和最优尺度的分类精度.将最细的条件属性尺度和最粗的决策属性尺度组合起来即为最细尺度,将最粗的条件属性尺度和最细的决策属性尺度组合起来即为最粗尺度.实验结果如表3和表4所示.

实验结果表明,绝大部分数据集最优尺度的分类精度比最粗尺度和最细尺度更高.这说明单尺度决策系统向多尺度决策系统的转换,可以在一定程度上提高性能,在较弱的条件下仍能取得较好的决策结果,最优尺度选择具有一定的意义.

表3 SVM分类器下8个数据集最优尺度的分类精度
Tab. 3 Classification accuracy of the optimal scale of 8 datasets with the SVM classifier

| 编号 | 数据集 | SVM/% | | |
|----|------------------------------|---------------------|--------------|---------------------|
| | | 最细尺度 | 最粗尺度 | 最优尺度 |
| 1 | Iris | 96.33 ± 0.41 | 89.50 ± 0.85 | 96.50 ± 0.62 |
| 2 | Wholesale customers(region) | 72.33 ± 0.52 | 71.36 ± 0.33 | 72.33 ± 0.52 |
| 3 | Car Evaluation | 99.77 ± 0.15 | 73.47 ± 0.24 | 97.55 ± 0.17 |
| 4 | Nursery | 99.95 ± 0.02 | 86.17 ± 0.25 | 99.97 ± 0.01 |
| 5 | Abalone | 88.16 ± 0.13 | 54.27 ± 0.31 | 88.16 ± 0.13 |
| 6 | Shill Bidding | 98.89 ± 0.08 | 98.58 ± 0.09 | 99.49 ± 0.08 |
| 7 | Contraceptive Method Choice | 55.43 ± 0.99 | 46.83 ± 1.22 | 55.43 ± 0.99 |
| 8 | Estimation of obesity levels | 92.50 ± 0.26 | 65.33 ± 0.40 | 92.65 ± 0.24 |

表4 KNN分类器下8个数据集最优尺度的分类精度
Tab. 4 Classification accuracy of the optimal scale of 8 datasets with the KNN classifier

| 编号 | 数据集 | KNN/% | | |
|----|------------------------------|---------------------|--------------|---------------------|
| | | 最细尺度 | 最粗尺度 | 最优尺度 |
| 1 | Iris | 96.00 ± 0.33 | 90.67 ± 1.11 | 96.17 ± 0.17 |
| 2 | Wholesale customers(region) | 70.80 ± 0.79 | 70.64 ± 3.33 | 70.80 ± 0.79 |
| 3 | Car Evaluation | 98.70 ± 0.18 | 72.66 ± 0.58 | 93.05 ± 0.56 |
| 4 | Nursery | 98.91 ± 0.05 | 84.95 ± 0.29 | 99.25 ± 0.09 |
| 5 | Abalone | 88.20 ± 0.46 | 49.90 ± 0.93 | 88.20 ± 0.46 |
| 6 | Shill Bidding | 96.86 ± 0.08 | 98.16 ± 0.09 | 98.38 ± 0.06 |
| 7 | Contraceptive Method Choice | 51.63 ± 0.94 | 43.51 ± 1.12 | 51.63 ± 0.94 |
| 8 | Estimation of obesity levels | 91.68 ± 0.25 | 61.59 ± 0.61 | 92.16 ± 0.37 |

4 结论

通过深入研究广义多尺度决策表的最优尺度性质,探讨了边界域的变化情况,给出了判断边界域是否相等的等价条件和充分条件,提出了最优尺度的等价定义及一种最优尺度选择快速算法.通过数值实验验证了最优尺度选择具有一定的意义.从实验结果来看,在属性数或对象数较多的数据集上算法的优势比较明显,但随着数据规模的增加,算法运行时间的增加不可避免.因此,最优尺度选择算法的计算效率有待于进一步提升.同时,如何将最优尺度选择方法推广到不完备的广义多尺度决策表,以及动态环境下的最优尺度更新问题是接下来需要研究的重要内容.

参考文献:

- [1] GUO D, XU W. Fuzzy-based concept-cognitive learning: An investigation of novel approach to tumor diagnosis analysis[J]. Information Sciences, 2023, 639: 118998.
- [2] HAO C, LI J, 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.
- [3] 郭豆豆,徐伟华. R-FCCL: 一种面向高维数据的模糊概念认知学习方法[J/OL]. 计算机研究与发展, 1-13. (2024-03-09)[2024-04-24]. <http://kns.cnki.net/kcms/detail/11.1777.TP.20240307.1525.002.html>.
- [4] PAWLAK Z. Rough sets[J]. International Journal of Computer & Information Sciences, 1982, 11: 341-356.
- [5] 盛茹雪,李红宇,姜春茂,等. 一种基于序贯三支决策的属性约简方法研究[J]. 模糊系统与数学, 2021, 35(6): 48-65.
- [6] 刘悦,姜春茂,郭豆豆. 一种基于区间模糊优势距离的多属性决策方法[J]. 山西大学学报(自然科学版), 2020, 43(4): 786-794.
- [7] 徐伟华,黄旭东,蔡可. 基于粒计算的多源信息融合方法综述[J]. 数据采集与处理, 2023, 38(2): 245-261.
- [8] WU W Z, LEUNG Y. Theory and applications of granular labelled partitions in multi-scale decision tables[J]. Information Sciences, 2011, 181(18): 3878-3897.
- [9] LI F, HU B Q. A new approach of optimal scale selection to multi-scale decision tables[J]. Information Sciences, 2017, 381: 193-208.
- [10] 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.
- [11] HUANG Z, LI J, DAI W, et al. Generalized multi-scale decision tables with multi-scale decision attributes[J]. International Journal of Approximate Reasoning, 2019, 115: 194-208.
- [12] ZHANG X, HUANG Y. Optimal scale selection and knowledge discovery in generalized multi-scale decision tables[J]. International Journal of Approximate Reasoning, 2023, 161: 108983.
- [13] ZHENG J W, WU W Z, BAO H, et al. Evidence theory based optimal scale selection for multi-scale ordered decision systems[J]. International Journal of Machine Learning and Cybernetics, 2022, 13(4): 1115-1129.
- [14] CHEN Y, LI J, LI J, et al. Sequential 3WD-based local optimal scale selection in dynamic multi-scale decision information systems[J]. International Journal of Approximate Reasoning, 2023, 152: 221-235.

[责任编辑:姜生有]