文章编号:1672-3961(2024)06-0038-11 DOI:10.6040/j.issn.1672-3961.0.2023.198

一种改进的悲观多粒度粗糙集粒度约简算法

谢立[1],叶军[1,2*],赖鹏飞[1],卢岚[1],周浩岩[1],李兆彬[1]
(1.南昌工程学院信息工程学院, 江西 南昌 330099; 2.江西省水信息协同感知与智能处理重点实验室(南昌工程学院), 江西 南昌 330099)

摘要:针对以粒度内部重要度和粒度外部重要度不能有效度量非核粒度的重要度,无法获得有效启发信息使约简过早收敛的问题,提出以正域变化度量核粒度的重要度、以边界集变化度量非核粒度的重要度。新的度量方法不仅能度量核粒度的重要度,而且能度量非核粒度的重要度。以新的粒度重要度为依据,提出一种改进的悲观多粒度约简算法,与样本选择的启发式属性约简算法、信息熵的模糊 ε-近似约简算法、粒度加速求解约简算法和邻域区分指数的特征选择算法相比,新算法可以减少迭代次数,能更有效地找到粒度约简子集。通过加州大学欧文分校(University of California Irvine, UCI)数据集进行试验,验证了算法的有效性和实用性。

关键词:多粒度粗糙集;粒度重要度;粒度空间;粒度约简;核粒度
中图分类号:TP18 **文献标志码**:A
引用格式:谢立,叶军,赖鹏飞,等. 一种改进的悲观多粒度粗糙集粒度约简算法[J].山东大学学报(工学版),2024,54(6):38-48.
XIE Li, YE Jun, LAI Pengfei, et al. An improved granular reduction algorithm for pessimistic multi-granularity rough sets[J]. Journal of Shandong University (Engineering Science), 2024, 54(6):38-48.

An improved granular reduction algorithm for pessimistic multi-granularity rough sets

XIE Li[1], YE Jun[1,2*], LAI Pengfei[1], LU Lan[1], ZHOU Haoyan[1], LI Zhaobin[1]
(1. College of Information Engineering, Nanchang Institute of Engineering, Nanchang 330099, Jiangxi, China; 2. Jiangxi Province Key Laboratory of Water Information Cooperative Sensing and Intelligent Processing(Nanchang Institute of Engineering), Nanchang 330099, Jiangxi, China)

Abstract: Aiming at the issue that non-kernel granularity significance couldnot be effectively measured by internal and external granularity significance, leading to premature convergence of reduction due to lack of effective heuristic information, a positive domain change was proposed to measure the significance of kernel granularity, and the change of boundary set was used to measure the significance of non-kernel granularity. The new measurement method could measure not only the significance of kernel granularity but also that of non-kernel granularity. Based on the new granularity significance, an improved pessimistic multi-granularity reduction algorithm was proposed. Compared with the heuristic attribute reduction algorithm of sample selection, the fuzzy ε-approximate reduction algorithm of information entropy, the reduction algorithm of granularity acceleration and the feature selection algorithm of neighborhood discrimination index, the new algorithm could reduce the number of iterations and found the granularity reduction subset more efficiently. Through the experimental analysisof University of California Irvine (UCI) data sets, it was proved that the algorithm was effective and practical.

Keywords: multi-granularity rough set; granularity significance; granular spaces; granularity reduction; kernel granular

0 引言

粗糙集理论是一种无需借助任何外部知识处理不确定数据的数学工具[1],广泛应用于属性约简、知识

收稿日期:2023-08-15
基金项目:江西省教育厅科技资助项目(GJJ211920);国家自然科学基金资助项目(61562061)
第一作者简介:谢立(1996—),女,江西赣州人,硕士研究生,主要研究方向为粗糙集与粒计算理论。E-mail:787719479@qq.com
*通信作者简介:叶军(1968—),男,江西万安人,教授,硕士生导师,硕士,主要研究方向为数据挖掘与知识发现、机器学习与人工智能、粗糙集与粒计算理论等。E-mail:2003992646@nit.edu.cn

发现和数据挖掘领域[2-4]。在粗糙集理论中,通常把一个属性集视为一个粒度空间,论域划分成知识粒度的集合[5]。在粒计算视角下,经典粗糙集属于单一粒度空间,只能从单一角度去近似目标概念。客观世界纷繁复杂,从多视角、多层次对问题进行刻画才能更加真实地揭示问题的本质。文献[6]将多粒度概念引入粗糙集中,提出多粒度粗糙集模型,通过多个粒度空间对目标进行描述,提高逼近目标概念的精度。随后,许多学者提出了多种拓展模型[7-10]。

与经典粗糙集属性约简一样,粒度约简是多粒度粗糙集中研究的重要问题[11-13]。近年来,研究者们在粒度约简方面取得了许多研究成果:文献[14-15]定义了粒度内部重要度和粒度外部重要度,设计了基于粒度内部重要度的启发式粒度约简算法;文献[16]将信息量引入悲观多粒度粗糙集,定义了一种基于信息熵的粒度内部重要度,将其作为启发因子提出多粒度约简算法;文献[17]在粒度熵的基础上,给出一种变精度悲观多粒度粗糙集粒度内部重要度的度量方法,提出基于粒度熵的悲观多粒度启发式粒度约简算法;文献[18]在定义决策类下近似布尔矩阵的基础上,给出基于该矩阵的计算决策粒度约简算法,通过布尔运算提升约简效率;文献[19]通过构造属性粒,利用属性粒在区分矩阵上的差异性定义粒度重要度,提出 2 种重要度在区分矩阵上的多粒度约简算法,获得更好的约简结果多样性;文献[20]定义了 2 种基于上下近似集的粒度重要性度量矩阵方法,以此设计了 2 种启发式粒度约简算法,在时间消耗上更少,约简效率较高。

文献[14-17]中,基于粒度内部重要度或粒度外部重要度为启发信息设计的约简算法虽然在一些决策表中取得较好的约简效果,但也存在一定的局限性,因为文献[14-17]中粒度重要度的度量方法延用了经典粗糙集方法,仍然是以正域依赖度的变化为依据,而能改变正域的只有核粒度,因此,只能度量核粒度的重要度,无法度量非核粒度的重要度,由于不能获得有效启发信息,约简过早收敛。为此,文献[21-22]提出改进方法:文献[21]以多数包含关系替代正域依赖度关系,给出一种粒度重要度的度量方法,但计算得到的结果会出现非核粒度重要度大于核粒度重要度的情况;文献[22]在正域依赖度基础上,加入剩余粒度对决策的间接作用,有效改善了粒度内部重要度全为 0 的情况,但由于每次迭代都要考虑剩余粒度的间接影响,增加了计算粒度重要度的工作量。针对此类问题,本研究借鉴上述众多研究成果,提出一种改进的粒度重要度度量方法,不仅定义了核粒度的重要度,而且能度量非核粒度的重要度,以非核粒度的重要度为启发因子设计粒度约简算法。实例分析表明,本研究方法可以有效找到粒度约简子集,为悲观多粒度粗糙集粒度空间约简提供一种新思路。

1　多粒度粗糙集基本知识

多粒度粗糙集把经典粗糙集的单一粒度空间拓展到多粒度、多层次空间,下面简要介绍相关知识。

在悲观多粒度粗糙集中[6],四元组 $S=(U,C\cup D,V,f)$ 是一个完备的决策信息系统,其中 U 为非空对象的集合,C 为条件属性集,D 为决策属性集,V 为全部对象在各个属性上的取值构成的集合,$f:U\times A\to V$ 表示一个信息函数,指定 U 中每个对象 x 的属性值,粒度集 $A=\{A_1,A_2,\cdots,A_m\}\subseteq C$,每个属性集 A_i 称为一个粒度,可对 U 基于等价关系 $\mathrm{IND}(A_i)$ 划分得到一个粒度空间,表示为 $U/\mathrm{IND}(A_i)=\{[x]_{A_i}:x\in U\}$,等价类 $[x]_{A_i}$ 称为知识粒,U 基于决策属性的划分 $U/D=\{Y_1,Y_2,\cdots,Y_r\}$,其中 $Y_1,Y_2,\cdots,Y_r$ 为由决策属性导出的分类子集。

定义 1　设 $S=(U,C\cup D,V,f)$ 是一个完备的决策信息系统,其中 $A=\{A_1,A_2,\cdots,A_m\}\subseteq C$,$X\subseteq U$ 为 U 的一个子集,X 的悲观多粒度的下近似、上近似分别为:

$$\underline{\sum_{i=1}^{m}A_i^P}(X)=(x:[x]_{A_1}\subseteq X\wedge\cdots\wedge[x]_{A_m}\subseteq X,\ x\in U),$$

$$\overline{\sum_{i=1}^{m}A_i^P}(X)=\sim(\underline{\sum_{i=1}^{m}A_i^P}(\sim X)),$$

式中 P 为悲观多粒度粗糙集。

根据以上定义可得悲观多粒度的边界域

$$B_{n\sum_{i=1}^{m}A_i^P(X)}=\overline{\sum_{i=1}^{m}A_i^P}(X)-\underline{\sum_{i=1}^{m}A_i^P}(X)。$$

定义 2　设 $S=(U,C\cup D,V,f)$ 是一个完备的决策信息系统,其中 $A=\{A_1,A_2,\cdots,A_m\}\subseteq C$,粒度集上的

任何一个子集 $B\subseteq A$, $U/D=\{Y_1,Y_2,\cdots,Y_r\}$,则悲观多粒度粗糙集下近似分布

$$\mu_A(U,D)=\left\{\underline{\sum_{A_i\in A}A_i^P(Y_1)},\ \underline{\sum_{A_i\in A}A_i^P(Y_2)},\cdots,\underline{\sum_{A_i\in A}A_i^P(Y_r)}\right\}。$$

若 $\mu_B(U,D)=\mu_A(U,D)$,则称 B 为 A 的粒度下近似分布一致集。若 $\mu_{A-\{A_i\}}(U,D)=\mu_A(U,D)$,则称 A_i 在粒度集 A 中是不必要的,否则 A_i 在粒度集 A 中是必要的。所有必要的粒度组成的集合为核粒度集,记为 $C_{\mathrm{ORE}}(A)$。

定义 3 设 $S=(U,C\cup D,V,f)$ 是一个完备的决策信息系统,$A=\{A_1,A_2,\cdots,A_m\}\subseteq C$, $U/D=\{Y_1,Y_2,\cdots,Y_r\}$。用粒度集 A 表示的 $Y_j(j=1,2,\cdots,r)$ 近似分布质量

$$\gamma_{\sum_{i=1}^m A_i}(D)=\left(\sum_{j=1}^r\frac{\left|\underline{\sum_{i=1}^m A_i(Y_j)}\right|}{|U|}\right)\Big/r,$$

式中 $|\cdot|$ 表示集合的基数。

定义 4 设 $S=(U,C\cup D,V,f)$ 是一个完备的决策信息系统,$A=\{A_1,A_2,\cdots,A_m\}\subseteq C$,若满足 $\gamma_A(D)=\gamma_C(D)$ 且对 $\forall A_i\in A$ 都有 $\gamma_{A-\{A_i\}}(D)\neq\gamma_C(D)$,则称 A 为 S 的多粒度下近似分布约简,所有约简集记为 R_{ED},则有 $\cap R_{\mathrm{ED}}=C_{\mathrm{ORE}}(A)$。

2 改进的粒度重要度度量方法

目前,现有的基于粒度内部或粒度外部重要度定义方法均延续了经典粗糙集属性重要度的定义方法,仍然是以正域依赖度变化为依据。例如,文献[14-15]从定性角度定义了粒度内部重要度,定义如下。

定义 5 设 $S=(U,C\cup D,V,f)$ 是一个完备的决策信息系统,$A=\{A_1,A_2,\cdots,A_m\}\subseteq C$, $U/D=\{Y_1,Y_2,\cdots,Y_r\}$,对 $\forall A_i\in A$,粒度 A_i 在粒度集 A 相对于 D 的重要性

$$s_{\mathrm{ig,\ in}}(A_i,A,D)=|\gamma_A(D)-\gamma_{A-\{A_i\}}(D)|。\tag{1}$$

文献[16-17]引入信息熵,从定量角度给出了粒度内部重要度的定义方法,具体定义如下。

定义 6 设 $S=(U,C\cup D,V,f)$ 是一个完备的决策信息系统,$A=\{A_1,A_2,\cdots,A_m\}\subseteq C$, $U/D=\{Y_1,Y_2,\cdots,Y_r\}$,对 $\forall A_i\in A$,粒度 A_i 在粒度集 A 相对于 D 的重要度

$$S_{\mathrm{GF}}(A_i,A)=I(A|D)-I(A-\{A_i\}|D),\tag{2}$$

式中 I 为信息量,$I(A|D)=1-\dfrac{1}{|U|^2}\left|\underline{\sum_{j=1}^r\cap_{A_i\in A}A_i(Y_j)}\right|^2$。

从式(1)(2)可以看出,两者定义都是将下近似即正域的变化作为依据,而影响下近似变化的只有核粒度,因此,该方法只能度量核粒度空间的重要度,无法有效度量非核粒度空间的重要度。

针对定义 5 和定义 6 存在的不足,文献[21]以多数包含关系为基础,提出新的知识依赖性度量公式,定义方法如下。

定义 7 设 $S=(U,C\cup D,V,f)$ 是一个完备的决策信息系统,$A=\{A_1,A_2,\cdots,A_m\}\subseteq C$, $U/D=\{Y_1,Y_2,\cdots,Y_r\}$, $X,Y\subseteq U$,粒度 A_i 在粒度集 A 相对于 D 的重要性

$$s'_{\mathrm{ig}}(A_i,A,D)=|\gamma'_A(D)-\tau_{A-\{A_i\}}\gamma'_{A-\{A_i\}}(D)|,\tag{3}$$

式中:τ 为可信系数,取值介于 0 ~1;A_i 与 D 满足多数包含关系, $\gamma'_A=\sum_{q=1}^m\left(\sum_{X_i\in U/\mathrm{IND}(A_q),\ Y_j\in U/D}[1-c(X_i,Y_j)|X_i]\right)/|U|^2$,其中 $c(X_i,Y_j)$ 为相对错误分类率。

定义 7 不是以下近似发生变化为依据度量粒度重要度,是以多数包关系替代正域依赖度度量粒度重要度。该定义方法虽然有效避免了定义 5 和定义 6 中的非核粒度重要度都为 0 的情况,但是会出现非核粒度重要度大于核粒度重要度的情况,与多粒度粗糙集理论中核粒度最重要的相关知识相矛盾。

针对此类问题,为了能够客观真实地反映各粒度空间相对于决策的作用,本研究提出一种新的粒度重要度度量方法,定义如下。

定义 8 设 $S=(U,C\cup D,V,f)$ 是一个完备的决策信息系统,$A=\{A_1,A_2,\cdots,A_m\}\subseteq C$, $U/D=\{Y_1,Y_2,\cdots,Y_r\}$, $\forall A_i\in A$,多粒度粗糙集依赖度

$$\gamma_A^P(D)=\frac{\left|\underline{\sum_{j=1}^{r}A_i(Y_j)}\right|}{|U|}。\tag{4}$$

定义 9　设 $S=(U,C\cup D,V,f)$ 是一个完备的决策信息系统,$A=\{A_1,A_2,\cdots,A_m\}\subseteq C$, $U/D=\{Y_1,Y_2,\cdots,Y_r\}$, $\forall A_i\in A$,粒度 A_i 在粒度集 A 关于 D 的粒度重要性

$$s_{\mathrm{ig}}(A_i,A,D)=\begin{cases}1+\left|\gamma_A^P(D)-\gamma_{A-\{A_i\}}^P(D)\right|, & \gamma_A^P(D)\neq\gamma_{A-\{A_i\}}^P(D)\\ \sum_{j=1}^{r}\dfrac{|Y_j\cap[x]_{A-\{A_i\}}|}{|U|}, & \gamma_A^P(D)=\gamma_{A-\{A_i\}}^P(D)\end{cases}。\tag{5}$$

由式(4)(5)可知,当 $\gamma_A^P(D)\neq\gamma_{A-\{A_i\}}^P(D)$ 时,表示从条件粒度集去掉某个粒度后,正域依赖度发生改变,即下近似发生变化,说明删除的是必要粒度即核粒度。由于删除该粒度后改变了正域,说明该粒度相对于决策的作用大,对应重要度也应该大。因此,本研究在原有的粒度内部重要度基础上加上 1 作为该粒度的重要度,既合理反映了该粒度相对于决策的作用,又突出了该粒度作为核粒度的关键作用。

当 $\gamma_A^P(D)=\gamma_{A-\{A_i\}}^P(D)$ 时,表示从条件粒度集中去掉某个粒度后,正域依赖度没有改变,即下近似不变,说明去掉的是非核粒度。虽然下近似没有变化,但上近似或边界集可能发生改变,表明删除的粒度在区分对象时也起了作用。本研究以删除非核粒度后,条件粒度子集导出的分类子集与决策类子集的交再与论域全集的比作为非核粒度的重要度,即边界集的变化量与全集的比,由于其没有改变正域,只是改变了边界域,对决策的作用比核粒度小。因此,以边界集的变化量与全集的比度量非核粒度的重要度,客观体现了非核粒度对决策的作用。

分析定义 9 可知,该度量方法不仅能度量核粒度的重要度,而且合理度量了非核粒度的重要度,是对定义 5 和定义 6 的拓展,准确反映了各粒度对决策的作用。

下面通过实例说明本研究定义 9 相较于其他几个定义的优势。

例 1　设 $S=(U,C\cup D,V,f)$ 是一个关于风险投资的决策信息系统,决策表如表 1 所示[10],其中 $A=\{R_1,R_2,R_3,R_4\}$ 为粒度空间,D 为决策属性。

表 1　风险投资决策信息系统表

Table 1　Table of venture capital decision information system

U	R_1	R_2	R_3	R_4	D
x_1	一般	高	良好	大	是
x_2	较差	低	差	小	否
x_3	较差	高	良好	大	是
x_4	较差	中等	中	小	否
x_5	良好	中等	中	大	是
x_6	较差	中等	中	中	否
x_7	较差	高	差	中	是
x_8	较差	低	中	中	否
x_9	较差	高	良好	中	是

由表 1 可得由决策属性导出的分类 $U/D=\{\{x_1,x_3,x_5,x_7,x_9\},\{x_2,x_4,x_6,x_8\}\}$。设 $Y_1=\{x_1,x_3,x_5,x_7,x_9\}$,$Y_2=\{x_2,x_4,x_6,x_8\}$。根据定义 1.1 得到各粒度的下近似 $\underline{\sum_{i=1}^{4}R_i}(Y_1)=\{x_1\}$,$\underline{\sum_{i=1}^{4}R_i}(Y_2)=\varnothing$。分别用定义 5、定义 6、定义 7 和定义 9 计算得到各粒度重要度,结果如表 2 所示。

表 2　各粒度重要度对比

Table 2　Comparison of importance of different particle sizes

定义	重要度			
	R_1	R_2	R_3	R_4
定义 5[14-15]	1/18	0	0	0
定义 6[16-17]	3/81	0	0	0
定义 7[21]	28/324	46/324	43/324	37/324
定义 9	10/9	8/9	8/9	7/9

从计算结果可知,决策表的核粒度集为$\{R_1\}$,非核粒度集为$\{R_2,R_3,R_4\}$。对比表 2 可知,定义 5 得到了核粒度 R_1 的重要度为 1/18,而非核粒度 R_2、R_3、R_4 的重要度均为 0;定义 6 计算得到核粒度 R_1 的重要度为 3/81,而非核粒度 R_2、R_3、R_4 的重要度同样均为 0。可以看出定义 5 和定义 6 无法区分各非核粒度之间的重要度。定义 7 得到了核粒度 R_1 的重要度为 28/324,非核粒度 R_2 的重要度为 46/324,R_3 的重要度为43/324,R_4 的重要度为 37/324。定义 7 虽然得到了所有粒度重要度,但非核粒度的重要度大于核粒度 R_1 的重要度,与多粒度粗糙集理论中核粒度重要度最大的观点矛盾。本研究定义 9 计算得到了所有粒度重要度,核粒度 R_1 的重要度为 10/9,大于非核粒度(R_2 的重要度为 8/9,R_3 的重要度为 8/9,R_4 的重要度为 7/9),准确反映了各粒度对决策的作用。

由定义 9 可以得到如下几个性质。

性质 1 设 $S=(U,C\cup D,V,f)$是一个完备的决策信息系统,$A=\{A_1,A_2,\cdots,A_m\}\subseteq C$, $U/D=\{Y_1,Y_2,\cdots,Y_r\}$,$C_{\mathrm{ORE}}(A)$为核粒度集。若$\forall A_i\in C_{\mathrm{ORE}}(A)$,则 $s_{\mathrm{ig}}(A_i,A,D)>1$。

证明 由定义 9 可知:若$\forall A_i\in C_{\mathrm{ORE}}(A)$,则由式(5)可知 $\gamma_A^P(D)\neq\gamma_{A-\{A_i\}}^P(D)$,即$|\gamma_A^P(D)-\gamma_{A-\{A_i\}}^P(D)|>0$,显然 $1+|\gamma_A^P(D)-\gamma_{A-\{A_i\}}^P(D)|>1$,因此,$s_{\mathrm{ig}}(A_i,A,D)>1$。

由性质 1 可知,核粒度的重要度都大于 1,反映了核粒度对于决策的作用最大,相应的重要度最大。本研究方法对决策表 1 计算得到核粒度 R_1 的重要度大于 1,与性质 1 相吻合。

性质 2 设 $S=(U,C\cup D,V,f)$是一个完备的决策信息系统,$A=\{A_1,A_2,\cdots,A_m\}\subseteq C$, $U/D=\{Y_1,Y_2,\cdots,Y_r\}$,若有$\forall A_i\in A$ 且$A_i\notin C_{\mathrm{ORE}}(A)$,则 $0\leqslant s_{\mathrm{ig}}(A_i,A,D)<1$。

证明 由于对 $\forall\ Y_j\in U/D$,都有 $\varnothing\subseteq Y_j\cap[x]_{A-\{A_i\}}\subseteq Y_j\subset U$,则 $0\leqslant\sum_{j=1}^{r}|Y_j\cap[x]_{A-\{A_i\}}|\leqslant\sum_{j=1}^{r}|Y_j|<|U|$,进而 $0\leqslant\sum_{j=1}^{r}|Y_j\cap[x]_{A-\{A_i\}}|/|U|<1$,因此,有 $0\leqslant s_{\mathrm{ig}}(A_i,A,D)<1$。

性质 2 表明非核粒度的重要度不一定都为 0,即只要在区分对象时起了分类作用,非核粒度的重要度就不为 0。性质 2 改善了定义 5 与定义 6 非核粒度的重要度全为 0 的局限性。

性质 3 设 $S=(U,C\cup D,V,f)$是一个完备的决策信息系统,$A=\{A_1,A_2,\cdots,A_m\}\subseteq C$,若对于$\forall A_i\in C-C_{\mathrm{ORE}}(A)$,$\forall A_j\in C_{\mathrm{ORE}}(A)$,则有 $s_{\mathrm{ig}}(A_i,A,D)<s_{\mathrm{ig}}(A_j,A,D)$。

证明 由定义 9 可直接得出。

性质 3 说明核粒度的重要度一定比非核粒度的重要度大,与多粒度粗糙集理论中的核粒度最重要等概念相吻合。性质 3 还有效改善了定义 7 中非核粒度重要度比核粒度重要度大的情况。从决策表 1 计算得到的结果印证了本研究定义的粒度重要度度量方法更为合理。

3 基于改进的粒度重要度约简算法

3.1 算法设计

本研究以粒度重要度为启发性信息,设计粒度约简算法。计算粒度集的核粒度,以非核粒度的重要度作为启发信息;逐个选择粒度重要度最大的粒度加入核粒度;将近似分布质量作为约简算法的终止条件,直至获取到一个最小粒度约简子集。

算法 1 改进的悲观多粒度启发式约简算法

输入 决策信息系统 $S=(U,C\cup D,V,f)$, $A=\{A_1,A_2,\cdots,A_m\}\subseteq C$。

输出 决策信息系统的一个粒度约简集 R_{ED}。

(1) 初始化 $C_{\mathrm{ORE}}(A)\leftarrow\varnothing$, $R_{\mathrm{ED}}\leftarrow\varnothing$, $B\leftarrow\varnothing$;

(2) $\forall A_i\in A$, 计算 U/A_i, U/D;

(3) $\forall A_i\in A$, $Y_j\in U/D$,计算$\underline{A_i}(Y_j)$;

(4) $\forall A_i\in A$, $Y_j\in U/D$,计算 $\underline{\sum_{j=1}^{r}A_i}(Y_j)$;

(5) $\forall A_i\in A$, 计算 $s_{\mathrm{ig}}(A_i,A,D)$;

if $s_{\mathrm{ig}}(A_i,A,D)>1$

$C_{\mathrm{ORE}}(A)=C_{\mathrm{ORE}}(A)\cup A_i$;

else

$B=B\cup A_i$;

end if

对 B 中各 A_i 按重要度大小降序排列;

(6) $R_{\mathrm{ED}}\leftarrow C_{\mathrm{ORE}}(A)$;

if $\gamma^P_{R_{\mathrm{ED}}}(D)=\gamma^P_C(D)$

执行步骤(8);

else

执行步骤(7);

(7) $R_{\mathrm{ED}}(A)=R_{\mathrm{ED}}(A)\cup A_i$, $A_i\in B$, $B=B-A_i$;

if $\gamma^P_{R_{\mathrm{ED}}}(D)=\gamma^P_C(D)$

执行步骤(8);

else

执行步骤(7);

(8) 输出一个粒度约简集 R_{ED},算法终止。

3.2 时间复杂度分析

算法计算工作量主要在步骤(2)~(5)。步骤(2)中,$|A|$ 个粒度空间对论域 U 进行划分的时间复杂度为 $o(|A||U|^2)$;步骤(3)中,计算 $|A|$ 个粒度空间相对决策类的下近似的时间复杂度为 $o(|A||U|^2)$;步骤(4)中,计算不同决策类 $Y_j\in U/D$ 下近似的交集,时间复杂度为 $o(k|A||U|^2)$;步骤(5)中,计算粒度重要度时需要计算一遍下近似和边界集相对于决策类对象的变化量,时间复杂度为 $o(k^2|A||U|^2)$。因此,算法的总时间复杂度为 $o(k^2|A||U|^2)$。

文献[14]为悲观多粒度的下近似粒度分布约简算法,算法总时间复杂度为 $o(|A||U|^3)$。文献[15]为 α-下近似分布粒度约简算法,算法总时间复杂度为 $o((k+1)|A|^2|U|^2+\sum_{i=1}^{|A|}|A|-(i+1)(k+1)|A||U|^2)$。文献[16]为基于信息量的悲观多粒度粗糙集粒度约简启发式算法,算法总时间复杂度为 $o(|A|^2|U|^2)$。文献[23]为基于样本选择的启发式属性约简算法,算法总时间复杂度为 $o(|A|^2|U|^2+kT|U|)$。文献[24]为基于信息熵的模糊 ε-近似约简算法,文献[25]为基于邻域区分指数的特征选择算法,算法总时间复杂度为 $o(|A|^2|U|^2)$。文献[26]为基于粒度加速求解约简算法,算法总时间复杂度为 $o(|U|^2(|A|+1)|A|)$。

上述算法中,T 为迭代次数,k 为决策类别数,$|A|$ 为粒度数,$|U|$ 为对象数。由于 $k\ll|U|$,本研究算法的时间复杂度要优于文献[14,15,23],与文献[25]的 $o(|U|^2(|A|+1)|A|)$ 相当。在小型数据集上,本研究算法的时间复杂度高于文献[16,24,26],但在大型数据集上,本研究算法的时间复杂度较文献[16,24,26]会更有优势。

4 试验数据分析

4.1 实例验证

某医疗诊断决策表如表 3 所示。中医专家通过患者疾病所呈现的不同阶段的症状初步判断患者是否患有关节炎。其中,粒度集为 $A=\{A_1,A_2,A_3,A_4,A_5\}=\{$畸形程度,舌脉,僵硬,疼痛肿胀,年龄$\}$。对于膝关节畸形程度的描述:0 表示关节间隙无变窄,可疑或微小骨赘;1 表示关节间隙可疑变窄,明显轻度骨赘;2 表示关节间隙明显狭窄,骨质硬化性改变,有中度多发骨赘形成。对于舌脉的描述:0 表示舌质淡,苔白腻;1 表示舌质红,苔黄腻;2 表示舌质紫暗或有瘀斑。对于僵硬的描述:0 表示不严重,1 表示有点严重,2 表示比较严重。对于疼痛肿胀的描述:0 表示无明显不适;1 表示轻微疼痛无肿胀感;2 表示轻度疼痛局部肿胀;3 表示中度疼痛肿胀,关节活动受限。对于患者年龄的描述:1 表示 20~30 岁,2 表示 31~40 岁,3 表示 41~50 岁,4 表示 51~60 岁。d 为是否患有关节炎,0 表示没有患关节炎,1 表示患有关节炎。患者集合 $O=\{o_1,o_2,o_3,$

$o_4,o_5,o_6,o_7\}$,其中 o 表示患者个体。本研究分别用文献[14]的算法、文献[16]的算法和本研究算法寻找粒度约简子集。

表 3 医疗诊断决策系统

Table 3 Medical diagnosis decision-making system

O	A_1	A_2	A_3	A_4	A_5	d
o_1	1	0	0	2	3	0
o_2	0	0	1	1	3	0
o_3	0	1	1	1	1	1
o_4	2	1	2	1	2	1
o_5	0	2	2	3	4	1
o_6	0	2	2	3	4	1
o_7	2	1	1	0	1	1

论域 O 在决策属性集 D 下的划分为 $O/D=\{\{o_1,o_2\},\{o_3,o_4,o_5,o_6,o_7\}\}$。取 $Y_1=\{o_1,o_2\}$,$Y_2=\{o_3,o_4,o_5,o_6,o_7\}$。下面分别计算粒度重要度并进行约简。

(1)文献[14]的方法

由定义 5 中式(1)可得 $s_{\mathrm{ig,in}}(A_1,A,D)=|\gamma_A(D)-\gamma_{A-\{A_1\}}(D)|=\left|\left(\sum_{j=1}^{2}\frac{|\underline{\sum_{i=1}^{5}A_i(Y_j)}|}{|U|}\right)/r-\left(\sum_{j=1}^{2}\frac{|\underline{\sum_{i=2}^{5}A_i(Y_j)}|}{|U|}\right)/r\right|=|(1+0)-(1+2)|/14=0.142\ 9$。同理可得 $s_{\mathrm{ig,in}}(A_2,A,D)=0$,$s_{\mathrm{ig,in}}(A_3,A,D)=s_{\mathrm{ig,in}}(A_4,A,D)=0.071\ 4$,$s_{\mathrm{ig,in}}(A_5,A,D)=0$。

由上述计算结果可知,该决策表的核粒度集 $C_{\mathrm{ORE}}(A)=\{A_1,A_3,A_4\}$,非核粒度集为 $=\{A_2,A_5\}$。由于非核粒度 A_2、A_5 的重要度都为 0,找不到有效启发信息,因此无法寻找到粒度子集。

(2)文献[16]的方法

由定义 6 中式(2)计算粒度 A_1 的重要度 $S_{\mathrm{GF}}(A_1,A)=I(A|D)-I(A-\{A_1\}|D)=1-\frac{1}{|U|^2}|\underline{\sum_{j=1}^{2}\cap_{A_i\in A}A_i(Y_j)}|^2-(1-\frac{1}{|U|^2}|\underline{\sum_{j=1}^{2}\cap_{A_i\in A,i\neq 1}A_i(Y_j)}|^2)=|(1^2+0^2)-(1^2+2^2)|/9^2=0.049\ 4$。同理可得 $S_{\mathrm{GF}}(A_2,A)=0$,$S_{\mathrm{GF}}(A_3,A)=S_{\mathrm{GF}}(A_4,A)=0.012\ 3$,$S_{\mathrm{GF}}(A_5,A)=0$。

由上述计算结果可知,该决策表的核粒度集 $C_{\mathrm{ORE}}(A)=\{A_1,A_3,A_4\}$,非核粒度集为 $=\{A_2,A_5\}$。同样,由于非核粒度 A_2、A_5 的重要度都为 0,找不到有效启发信息,无法得到粒度约简子集。

(3)本研究方法

由定义 8 中式(4)和定义 9 中式(5)可得 $s_{\mathrm{ig}}(A_1,A,D)=1+|\gamma_A^P(D)-\gamma_{A-\{A_1\}}^P(D)|=1+\sum_{j=1}^{2}\left|\frac{|\underline{\sum_{i=1}^{5}A_i(Y_j)}|}{|U|}-\frac{|\underline{\sum_{i=2}^{5}A_i(Y_j)}|}{|U|}\right|=1+\left|\frac{1+0}{7}-\frac{1+2}{7}\right|=1.285\ 7$,$s_{\mathrm{ig}}(A_2,A,D)=\sum_{j=1}^{2}\frac{|Y_j\cap[x]_{A-\{A_1\}}|}{|U|}=\left|\frac{2+4}{7}\right|=0.857\ 1$。同理可得 $s_{\mathrm{ig}}(A_3,A,D)=s_{\mathrm{ig}}(A_4,A,D)=1.142\ 9$,$s_{\mathrm{ig}}(A_5,A,D)=0.857\ 1$。

由计算结果可以看出,只有粒度 A_1、A_3、A_4 的重要度大于 1,因此,得到核粒度集 $C_{\mathrm{ORE}}(A)=\{A_1,A_3,A_4\}$,非核粒度集为 $=\{A_2,A_5\}$。以核粒度集 $C_{\mathrm{ORE}}(A)=\{A_1,A_3,A_4\}$ 为起点,依次将重要度最大的粒度 A_2 加入核集,同时由于 $\gamma_{\{A_1,A_2,A_3,A_4\}}^P(D)=\gamma_A^P(D)$,得到的粒度约简子集 $R_{\mathrm{ED}}=\{A_1,A_2,A_3,A_4\}$。

从上述计算结果可以看出,本研究方法计算得到的各粒度重要度结果与 3 个性质完全吻合,反映了本研究定义的粒度重要度度量方法更合理,能够有效找到粒度约简子集。

4.2 UCI 数据集试验与分析

为了进一步验证本研究提出的基于改进的粒度重要度粒度约简算法的有效性和实用性,选取加州大学

欧文分校(University of California Irvine, UCI)数据集中的 10 个常用数据集进行试验测试。为了便于对比,将 10 个数据集作为单粒度决策表进行试验,连续型数据决策表试验前作了离散化处理。数据集具体情况如表 4 所示。从表 4 可以看出,本研究选取了不同规模的数据集进行试验,样本数从 32 到 19 020,属性维数从 4 到 256。将测试结果与文献[23-26]算法进行对比。其中,文献[23]算法是用 *K*-means 选取出的样本构建新的决策系统,再执行传统启发式属性约简算法;文献[24]算法是在信息熵的基础上引入模糊计算的概念作为约简标准;文献[25]算法是以粒度大小为标准,剔除对应较粗粒化结果的属性,压缩候选属性的搜索空间求约简;文献[26]算法是用邻域辨别指数表示邻域关系的区分信息求约简。

表 4　试验数据集信息

Table 4　Experimental data set information

编号	数据集	实例数	属性维数	类别数
1	Lung cancer	32	56	3
2	Molecular biology	106	57	2
3	Breast tissue	106	9	6
4	LSVT	126	256	2
5	Iris	150	4	3
6	Seeds	210	7	3
7	Column	310	6	3
8	Climate model	540	20	2
9	Banknote	1 372	5	2
10	Magic telescope	19 020	11	2

试验环境是在 12th Gen Intel (R) Core (TM) i5-12400 CPU 和 8GB 内存 PC 机上,采用 Window10 环境下的 MATLAB 2021a。对 10 个数据集进行多次重复试验,排除试验过程中因试验次数少而出现的偶然性。从时间消耗、分类准确率、约简数、约简率等 4 个方面进行比较。

4.2.1　时间消耗比较

在 10 个数据集上运行 10 次,取平均消耗时间。为便于比对,对平均消耗时间 t 取对数,如图 1 所示。

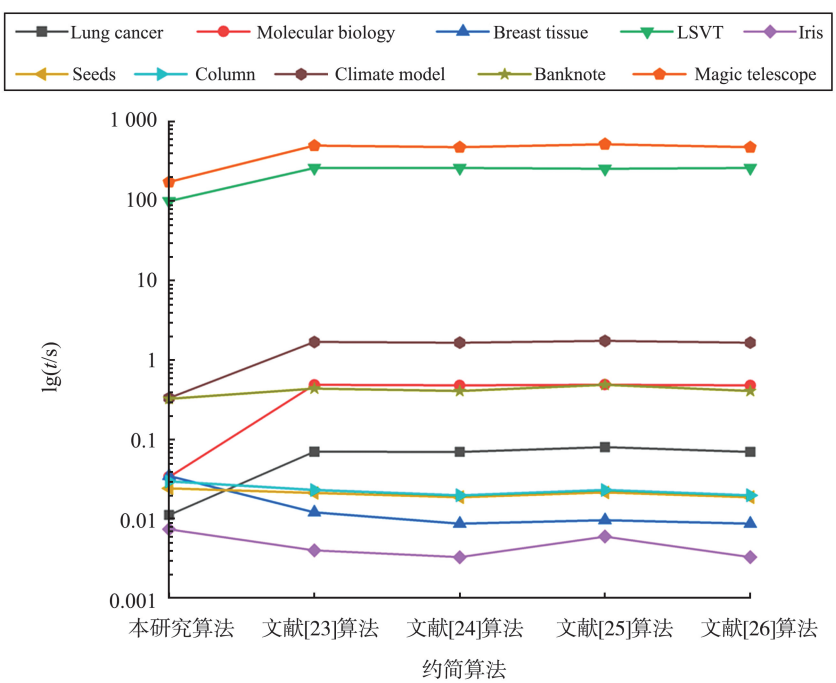

图 1　时间消耗对比

Fig.1　Comparison of time consumption

由图 1 可知:在样本数小和维度较少的 Breast tissue、Iris、Seeds、Column 数据集上,5 种约简算法运行所消耗的时间相近,差距并不明显;在样本多的数据集 Climate model、Banknote、Magic telescope 及高维数据集 Lung cancer、Molecular biology、LSVT 上,本研究算法消耗的时间比其他 4 种算法少;随着数据集的样本数和

维数的增加,本研究算法的优势更明显。由 3.2 节中算法时间复杂度分析可知,本研究算法的优势在于计算粒度重要度时只需要计算一遍下近似和边界集相对变化量,不需要计算与样本相同数量的个数。本研究算法的时间复杂度为 $o(k^2|A||U|^2)$,而其他 4 个算法的时间复杂度等于或大于 $o(|A|^2|U|^2)$。本研究算法有效减少了迭代次数,故整体约简所耗时间减少。

4.2.2 分类准确率比较

分别用 5 种算法获得的粒度约简子集对 10 个数据集进行分类,取运行 10 次得到的最优准确率对比结果,分类准确率结果如表 5 所示,其中加粗数据表示每个数据集上取得的最好结果。

表 5 分类准确率结果比较

Table 5 Comparison of classification accuracy results

编号	分类准确率				
	本研究算法	文献[23]算法	文献[24]算法	文献[25]算法	文献[26]算法
1	**0.948 2**	0.944 8	0.913 6	0.921 1	0.900 4
2	0.892 2	0.934 1	**0.941 8**	0.861 0	0.897 1
3	**0.682 5**	0.576 8	0.471 7	0.423 8	0.421 0
4	**0.808 4**	0.804 7	0.774 5	0.795 6	0.800 2
5	**0.988 7**	0.924 7	0.963 2	0.956 3	0.984 2
6	0.914 3	0.891 4	0.900 0	0.904 2	**0.964 2**
7	**0.832 3**	0.760 0	0.819 4	0.790 0	0.800 0
8	0.914 4	0.907 4	0.929 6	**0.940 7**	0.922 2
9	0.978 7	0.924 7	0.963 2	**0.987 5**	0.987 3
10	**0.851 2**	0.824 7	0.724 7	0.834 0	0.708 8

由表 5 可知:本研究算法在 Lung cancer、Breast tissue、LSVT、Iris、Column、Magic telescope 数据集上具有较好的分类准确率,文献[25-26]算法表现次之;文献[23]算法在 Climate model、Banknote 数据集上具有较好的分类准确率;文献[26]算法在 Seeds 数据集上具有较好的分类准确率。上述结果验证了本研究算法的可行性和实用性。

4.2.3 约简数比较

将本研究算法求得的约简数与其他 4 个算法进行比较,结果如表 6 所示。

表 6 约简数的对比

Table 6 Comparison of numbers freduction

编号	约简数				
	本研究算法	文献[23]算法	文献[24]算法	文献[25]算法	文献[26]算法
1	8	10	9	12	9
2	10	9	6	10	11
3	2	3	4	6	7
4	56	68	82	74	59
5	1	3	3	3	2
6	3	5	6	4	3
7	2	5	3	2	3
8	15	12	17	16	12
9	2	3	2	2	2
10	5	6	8	5	9

由表 6 可知:本研究算法在 Lung cancer、Breast tissue、LSVT、Iris、Column、Banknote、Magic telescope 数据集上约简掉的粒度个数更多;在 Seeds、Banknote 数据集上,文献[26]算法与本研究算法有同样的约简数;在 Climate model 数据集上,文献[23,26]算法有更好的约简结果。总体可见,本研究算法对原始数据集具有较好的约简效果。

4.2.4 约简率比较

将 5 种算法获得的粒度约简结果用约简率表示,通常约简率越高,经过粒度约简后,去除冗余信息的能力越强。约简率

$$R=[(|A|-|R_{ED}|)/|A|]\times 100\%。$$

5 种算法在 8 个数据集上约简后的约简率结果如图 2 所示。由图 2 可知:在 Lung cancer、Breast tissue、LSVT、Iris 数据集上,本研究算法的粒度约简率最高,分别为 85%、78%、78%、75%;在 Seeds 数据集上,本研究算法与文献[26]算法相同,获得了最高的约简率;在 Molecular biology、Column、Magic telescope 数据集上,本研究算法与文献[25]算法相同,获得了较好的约简率;在 Banknote 数据集上,本研究算法和文献[23,25-26]算法一样,同样也取得了较好的约简率。综上表明,本研究算法在多数数据集进行粒度约简能有效去除冗余信息,所得结果更为精简,具有明显优势。

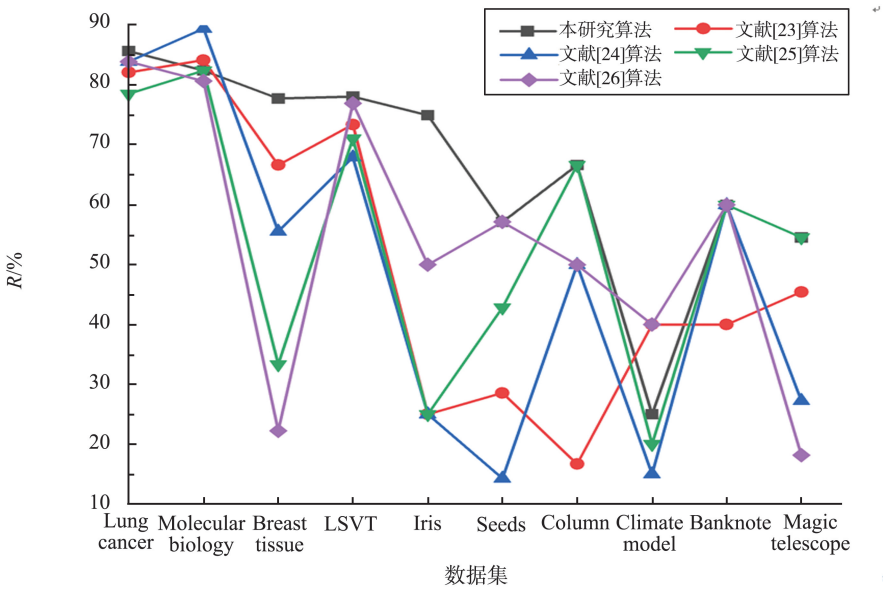

图 2　约简率比较

Fig.2　Comparison of reduction rates

5　结论

针对目前已有的以粒度内部重要度和粒度外部重要度为启发因子的悲观粒度约简方法存在的不足,本研究给出一种悲观多粒度粗糙集粒度重要度的定义方法,突出核粒度的重要度,且能度量非核粒度的重要度,合理反映各粒度对决策的作用。以新的粒度重要度为启发因子,设计一种改进的悲观多粒度粗糙集粒度约简算法,能更有效地找到粒度约简子集,加快收敛速度,降低时间复杂度,通过实例验证了该算法的有效性和实用性。本研究存在以下不足:本研究方法只适合于完备多粒度决策信息系统下的粒度约简问题,对非完备系统的多粒度约简是后续要研究的工作;本研究粒度重要度度量方法有效区分了核粒度和非核粒度 2 类粒度的重要度,但如何区分非核粒度集中冗余粒度是今后进一步研究的重点。

参考文献:

[1] SONG S M, REN X J, HE J, et al. An optimal hierarchical approach for oral cancer diagnosis using rough set theory and an amended version of the competitive search algorithm[J]. Diagnostics, 2023, 13(14):2454.

[2] PEDRYCZ W. Granular computing for data analytics: a manifesto of human-centric computing[J]. IEEE/CAA Journal of Automatica Sinica, 2018, 5(6): 1025-1034.

[3] ALFEO A L, CIMINO M G C A, GAGLIARDI G. Concept-wise granular computing for explainable artificial intelligence[J]. Granular Computing, 2022, 8(4): 827-838.

[4] 高天宇, 王庆荣, 杨磊. 粗糙集属性依赖度强化的应急数据挖掘模型[J]. 计算机工程与应用, 2021, 57(3): 87-93.
GAO Tianyu, WANG Qingrong, YANG Lei. Data mining model based on attribute dependability enhancement of rough set [J]. Computer Engineering and Applications, 2021, 57(3): 87-93.

[5] WANG H, GUAN J T. A dynamic framework for updating approximations with increasing or decreasing objects in multi-granulation rough sets[J]. Soft Computing, 2023, 27(9): 5257-5276.

[6] QIAN Y H, LIANG J Y, YAO Y Y, et al. MGRS: a multi-granulation rough set[J]. Information Sciences, 2010, 180(6): 949-970.

[7] SUN B Z, ZHANG X R, QI C, et al. Neighborhood relation-based variable precision multigranulation Pythagorean fuzzy rough set approach for multi-attribute group decision making[J]. International Journal of Approximate Reasoning, 2022, 151: 1-20.

[8] ZHANG X Y, JIANG J F. Measurement, modeling, reduction of decision-theoretic multigranulation fuzzy rough sets based on three-way decisions[J]. Information Sciences, 2022, 607: 1500-1582.

[9] ATEF M, ATIK E F E A. Some extensions of covering-based multigranulation fuzzy rough sets from new perspectives[J]. Soft Computing, 2021, 25(8): 6633-6651.

[10] 张明, 程科, 杨习贝, 等. 基于加权粒度的多粒度粗糙集[J]. 控制与决策, 2015, 30(2): 222-228.
ZHANG Ming, CHENG Ke, YANG Xibei, et al. Multigranulation rough set based on weighted granulations[J]. Control and Decision, 2015, 30(2): 222-228.

[11] 史进玲, 张倩倩, 徐久成. 多粒度决策系统属性约简的最优粒度选择[J]. 计算机科学, 2018, 45(2), 152-156.
SHI Jinling, ZHANG Qianqian, XU Jiucheng. Optimal granularity selection of attribute reductions in multi-granularity decision system[J]. Computer Science, 2018, 45(2), 152-156.

[12] 吕萍, 常玉慧, 钱进. 知识粒度框架下并行知识约简算法研究[J]. 南京大学学报(自然科学), 2022, 58(4): 594-603.
LÜ Ping, CHANG Yuhui, QIAN Jin. Parallel knowledge reduction algorithm using knowledge granularity[J]. Journal of Nanjing University (Natural Science), 2022, 58(4): 594-603.

[13] 李金海, 周新然. 多粒度决策形式背景的属性约简[J]. 模式识别与人工智能, 2022, 35(5): 387-400.
LI Jinhai, ZHOU Xinran. Attribute reduction in multi-granularity formal decision contexts[J]. Pattern Recognition and Artificial Intelligence, 2022, 35(5): 387-400.

[14] 桑妍丽, 钱宇华. 一种悲观多粒度粗糙集中的粒度约简算法[J]. 模式识别与人工智能, 2012, 25(3): 361-366.
SANG Yanli, QIAN Yuhua. A granular space reduction approach to pessimistic multi-granulation rough sets[J]. Pattern Recognition and Artificial Intelligence, 2012, 25(3): 361-366.

[15] 桑妍丽, 钱宇华. 多粒度决策粗糙集中的粒度约简方法[J]. 计算机科学, 2017, 44(5): 199-205.
SANG Yanli, QIAN Yuhua. Granular structure reduction approach to multigranulation decision-theoretic rough sets[J]. Computer Science, 2017, 44(5): 199-205.

[16] 孟慧丽, 马媛媛, 徐久成. 基于信息量的悲观多粒度粗糙集粒度约简[J]. 南京大学学报(自然科学), 2015, 51(2): 343-348.
MENG Huili, MA Yuanyuan, XU Jiucheng. The granularity reduction of pessimistic multi-granulation rough set based on the information quantity[J]. Journal of Nanjing University (Natural Science), 2015, 51(2): 343-348.

[17] 孟慧丽, 马媛媛, 徐久成. 基于下近似分布粒度熵的变精度悲观多粒度粗糙集粒度约简[J]. 计算机科学, 2016, 43(2): 83-85.
MENG Huili, MA Yuanyuan, XU Jiucheng. Granularity reduction of variable precision pessimistic multi-granulation rough set based on granularity entropy of lower approximate distribution[J]. Computer Science, 2016, 43(2): 83-85.

[18] 胡善忠, 徐怡, 何明慧, 等. 多粒度粗糙集粒度约简的高效算法[J]. 计算机应用, 2017, 37(12): 3391-3396.
HU Shanzhong, XU Yi, HE Minghui, et al. Effective algorithm for granulation reduction of multi-granulation rough set[J]. Journal of Computer Applications, 2017, 37(12): 3391-3396.

[19] 翁冉, 王俊红, 魏巍, 等. 基于区分矩阵的多粒度属性约简[J]. 南京航空航天大学学报, 2019, 51(5): 636-641.
WENG Ran, WANG Junhong, WEI Wei, et al. Multi-granulation attribute reduction based on discernibility matrix[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2019, 51(5): 636-641.

[20] 郑文彬, 李进金, 张燕兰, 等. 基于矩阵的多粒度粗糙集粒度约简方法[J]. 南京大学学报(自然科学), 2021, 51(1): 141-149.
ZHENG Wenbin, LI Jinjin, ZHANG Yanlan, et al. Matrix-based granulation reduction method for multi-granulation rough sets[J]. Journal of Nanjing University (Natural Science), 2021, 51(1): 141-149.

[21] 卢加学, 汪小燕. 关于粒度重要性公式的改进[J]. 苏州科技大学学报(自然科学版), 2021, 38(4): 79-84.
LU Jiaxue, WANG Xiaoyan. Improvements to the granularity importance formula[J]. Journal of Suzhou University of Science and Technology (Natural Science Edition), 2021, 38(4): 79-84.

[22] 薛占熬, 韩丹杰, 吕敏杰, 等. 一种新的基于粒度重要度的三支决策模型[J]. 计算机科学, 2019, 46(2): 236-241.
XUE Zhanao, HAN Danjie, LÜ Minjie, et al. New three-way decisions model based on granularity importance degree[J]. Computer Science, 2019, 46(2): 236-241.

(下转第 56 页)