



· 粒计算与概念知识获取 ·

多粒度粒球粗糙集模型

蒋珊珊¹, 林国平^{1,2,3}, 林艺东¹, 寇毅¹

(1. 闽南师范大学 数学与统计学院, 福建 漳州 363000; 2. 数字福建气象大数据研究所, 福建 漳州 363000;
3. 福建省粒计算及其应用重点实验室, 福建 漳州 363000)

摘要 基于粒球计算的粗糙集理论作为知识发现和数据挖掘的重要工具之一, 已成功地应用于标记预测、属性约简等。而现有的粒球粗糙集模型仅仅是从单粒度出发, 无法从多粒度角度对数据进行分析 and 处理, 实际生活中仍有很多应用场景需从多粒度角度进行思考。将粒球计算思想结合到多粒度粗糙集模型, 提出了多粒度粒球粗糙集模型, 并讨论了该模型的相关性质。该模型通过纯度的设定对数据进行粒球划分, 能够有效地刻画数据之间的内在联系, 以此设计多粒度粒球粗糙集的正域生成算法。实验分析表明该模型的可行性和有效性。

关键词 粒球计算; 粒球粗糙集; 多粒度粗糙集; 纯度

中图分类号: TP391 **DOI:** 10.16152/j.cnki.xdxbr.2024-02-006

Multi-granulation rough set model based on granular-ball computing

JIANG Shanshan¹, LIN Guoping^{1,2,3}, LIN Yidong¹, KOU Yi¹

(1. School of Mathematics and Statistics, Minnan Normal University, Zhangzhou 363000, China;

2. Institute of Meteorological Big Data-Digital Fujian, Zhangzhou 363000, China;

3. Fujian Key Laboratory of Granular Computing and Applications, Zhangzhou 363000, China)

Abstract As one of the important tools for knowledge discovery and data mining, rough set theory based on granular-ball computing has been successfully applied to label prediction and attribute reduction. However, the existing granular-ball rough set models only consider a single granulation, and cannot analyze and process data from a multi-granulation, and there are still many application scenarios that need to be considered from the perspective of multi-granulation. Based on this, this paper proposes a multi-granulation rough set based on granular-ball computing by embedding the idea of granular-ball in the multi-granulation rough set model, and discusses the relevant properties of the model. The model divides the data by setting the purity, which can effectively depict the internal relationship between the data, and thus design a position region generation algorithm for multi-granulation granular-ball rough set. Experimental analysis shows the feasibility and effectiveness of this model.

Keywords granular-ball computing; granular-ball rough set; multi-granulation rough set; purity

收稿日期: 2023-09-15

基金项目: 国家自然科学基金(11871259, 12101289, 12201284); 福建省自然科学基金(2021J01983, 2021J01979)。

第一作者: 蒋珊珊, 女, 从事粒计算与粗糙集理论及其应用研究, J_SShan0103@163.com。

通信作者: 林国平, 女, 博士, 教授, 从事粒计算与人工智能研究, guoplin@163.com。

粗糙集理论是 Pawlak^[1] 于 1982 年提出的一种能够有效分析和处理不精确、不一致、不完整信息的数学方法。粗糙集理论^[2-7] 已经广泛地应用于模式识别、数据挖掘、机器学习、决策支持系统等领域。为了处理不同类型的信息系统,许多学者将 Pawlak 粗糙集模型扩充为容差关系、相似关系、限制容差关系、优势关系和模糊关系粗糙集等^[8-11]。然而,经典粗糙集模型基于单个不可分辨二元关系的单一粒度框架,无法从多粒度、多层次的角度对数据进行分析 and 处理,单一粒度框架下的数据处理方法已经不能满足实际应用的需求。基于此,Qian 等人从粒计算的角度出发,考虑多个二元关系,将单粒度粗糙集模型拓展至多粒度结构,提出多粒度粗糙集思想,建立基于“求同存异”思想的乐观多粒度粗糙集和基于“求同排异”思想的悲观多粒度粗糙集^[12-14]。

此外,传统的粗糙集模型只能处理离散数据,而现实中的数据多为连续数据,离散化不可避免地造成信息的丢失。为了解决这一问题,Hu 等提出了邻域粗糙集,利用邻域来描述样本之间的关系,能够有效地处理连续型数据^[15]。基于它的诸多优势,许多学者对其进行了相关的研究和改进。李和谢提出了一种基于邻域粗糙集的特征子集增量式更新 NRS 加速方法^[16]。胡和赵等根据样本的分布提出了基于不确定性和邻域关系粗糙集的增量属性约简方法^[17]。彭和刘等设计了一个适应度函数,它结合了数据集和分类器的属性,从给定的邻域半径区间中选择最优邻域半径^[18]。

然而,邻域粗糙集的上下近似是由样本点组成,而不是等价类,因此使邻域粗糙集失去了可解释性。基于此,Xia 等人提出了一种基于粒球计算^[19]的粒球粗糙集^[20],通过引入粒球计算来表示邻域,用等价类来表示上下近似,从而实现 Pawlak 粗糙集和邻域粗糙集的统一。Xia 等提出的粒球计算是一种基于颗粒认知计算的新型、高效、鲁棒的粒计算方法,其核心思想是利用“粒球”覆盖或部分覆盖样本空间^[19]。此外,Xia 等还将粒球计算进行改进和发展,提出粒球分类器^[19]、粒球聚类^[21]、粒球邻域粗糙集^[22]和粒球采样方法^[23]。其中,粒球邻域粗糙集可以自动优化邻域半径。粒球计算还拓展到基于伪标签粒球粗糙集的约简^[24]、粒球生成树聚类算法^[25]等研究。

受文献[12]、[20]的启发,本文借鉴粒球计算的思想,结合多粒度粗糙集模型,提出“多粒度

粒球粗糙集模型”,将粒球粗糙集从单一粒度拓展为多粒度。此外,讨论了该模型的重要性质,给出了多粒度粒球粗糙集正域的生成算法,并通过实验验证该模型的可行性和有效性。

1 相关知识

本节主要回顾多粒度粗糙集、粒球粗糙集的相关知识。

1.1 多粒度粗糙集

Qian 等将 Pawlak 粗糙集模型扩展为多粒粗糙集模型,该模型通过论域上的多重等价关系定义集合近似^[12]。

定义 1^[12] 设 $DS = \langle U, AT, V, f \rangle$ 是一个完备的信息系统,任意 $B \in 2^{AT}, B = \{B_1, B_2, \dots, B_m\}, B_i \subseteq AT, i = 1, 2, \dots, m$ 。对于任意 $X \subseteq U$,则 X 关于 B 的上、下近似表示为

$$\sum_{i=1}^m \overline{B_i(X)} = \{x: [x]_{B_1} \subseteq X \wedge [x]_{B_1} \subseteq X \wedge \dots \wedge [x]_{B_m} \subseteq X, x \in U\} \quad (1)$$

$$\sum_{i=1}^m \underline{B_i(X)} = \sim (\sim \sum_{i=1}^m \overline{B_i(X)}) \quad (2)$$

称 $(\sum_{i=1}^m \overline{B_i(X)}, \sum_{i=1}^m \underline{B_i(X)})$ 为多粒度粗糙集模型。

1.2 粒球粗糙集

Xia 等提出粒球计算方法,此方法能够在信息粒化过程中,自适应地生成粒球信息粒^[19]。进一步提出粒球粗糙集,从而实现了 Pawlak 粗糙集和邻域粗糙集的统一表示。

定义 2^[20] 设 $GB = \{x_i, i = 1, 2, \dots, N\}$ 为粒球, x_i 表示粒球 GB 内的样本, N 为粒球 GB 中样本的个数。粒球 GB 的中心 C 和半径 r 分别定义为

$$C = \frac{1}{N} \sum_{i=1}^N x_i \quad (3)$$

$$r = \frac{1}{N} \sum_{i=1}^N |x_i - C| \quad (4)$$

定义 3^[20] 设粒球 $GB = \{x_i, i = 1, 2, \dots, N\}$, x_i 表示粒球 GB 的样本, N 为粒球 GB 中样本的个数。设 M 为粒球 GB 样本标签占比最大的样本数,则可定义粒球 GB 的纯度为

$$\text{Purity}(GB) = \frac{M}{N} \quad (5)$$

在粒球的生成过程中,首先,将整个数据集视为一个粒球;其次,计算粒球纯度,纯度不满足时将粒球均分为 2 个子球,依次进行迭代,直到所有

粒球的纯度满足要求时,边界最清晰且算法收敛。其主要步骤如下:

- 1) 假设 m 表示当前粒球的数量,将论域 U 初始化为一个粒球,令 $m = 1$;
- 2) 利用 k -means 聚类算法对每个粒球进行聚类,令 $k = 2$,则每个粒球分裂为两个子粒球,此时粒球数量为 $2m$;
- 3) 计算所有的子粒球的纯度 Purity,若所有的子粒球纯度达到要求或粒球半径 r 达到指定的阈值,则算法结束;否则,则返回步骤 2)。

定义 4^[18] 设 $DS = \langle U, AT, V, f \rangle$ 是一个完备的信息系统,任意 $x_i \in U$,其中 c_j 和 r_j 分别表示粒球 GB_j 的中心和半径,则粒球 GB_j 定义为

$$GB_j = \{x \mid x \in U, \Delta(x, c_j) \leq r_j\} \quad (6)$$

其中: $\Delta(x, c_j)$ 表示任意的对象 $x \in U$ 与中心 c_j 的距离度量。本文中 $\Delta(x, c_j) = \sum_{a \in B} (f(x, a) - \frac{1}{|U'|} \sum_{x_i \in U'} f(x_i, a))^2$, $f(x, a)$ 表示对象在属性 a 下的属性值, U' 表示为 k -means 聚类算法每次迭代的类样本。

定义 5^[20] 设 $DS = \langle U, AT, V, f \rangle$ 是一个完备的信息系统,任意 $x, y \in U$ 且 $B \subseteq AT$,粒球 GB 基于属性集 B 下的不可分辨关系定义为 $INDGB(B) = \{(x, y) \in U^2 \mid f(x, a) = f(y, a) = GB, \forall a \in B\}$ 。

根据定义 5,存在 $(x, y) \in INDGB(B)$,则 x 与 y 等价。论域 U 在 GB 下的划分表示为 $U/GB(B)$,粒球 GB 在不可分辨关系 $INDGB(B)$ 下的等价类表示为 $[x]_{GB(B)} = \{y \in U \mid (x, y) \in INDGB(B)\}$, $[x]_{GB(B)}$ 是 $U/GB(B)$ 的元素。

定义 6^[20] 设 $DS = \langle U, AT \cup D, V, f \rangle$ 是一个完备的决策信息系统, U 是非空有限集合, AT 是属性集,决策 D 将论域 U 划分为 L 个等价类,表示为 $U/D = \{X_1, X_2, \dots, X_L\}$,任意 $B \subseteq AT$,在 U 上存在着相应的等价关系 GBR_B 。 D 关于属性集 B 的上、下近似分别定义为

$$\overline{GBR_B}D = \bigcup_{i=1}^L \overline{GBR_B}X_i \quad (7)$$

$$\underline{GBR_B}D = \bigcup_{i=1}^L \underline{GBR_B}X_i \quad (8)$$

其中,对于任意 $x \in U$,

$$\overline{GBR_B}X_i = \bigcup \{[x]_B \in U/GB(B) \mid$$

$$[x]_{GB(B)} \cap X_i \neq \emptyset\};$$

$$\underline{GBR_B}X_i = \bigcap \{[x]_B \in U/GB(B) \mid$$

$$[x]_{GB(B)} \subseteq X_i\}.$$

定义 7^[20] 设 $DS = \langle U, AT \cup D, V, f \rangle$ 是一个完备的决策信息系统,任意 $B \subseteq AT$, DS 关于属性集 B 的正域和边界域定义为

$$POS_B(D) = \overline{GBR_B}D \quad (9)$$

$$BN_B(D) = \overline{GBR_B}D - \underline{GBR_B}D \quad (10)$$

2 多粒度粒球粗糙集模型

在本节内容中,借鉴粒球计算的思想,结合多粒度粗糙集模型,构造多粒度粒球粗糙集模型。

定义 8 设 $DS = \langle U, AT, V, f \rangle$ 是一个完备的信息系统, U 是非空有限集合, AT 是属性集。任意 $B \in 2^{AT}$, $B = \{B_1, B_2, \dots, B_m\}$, $B_i \subseteq AT$, $i = 1, 2, \dots, m$ 。在 U 上存在基于 B_i 的粒球 GB_{B_i} 相应的等价关系 GBR_{B_i} 。对于任意 $X \subseteq U$, X 关于 B 的上、下近似定义为

$$\begin{aligned} \sum_{i=1}^m \overline{GBR_{B_i}}X &= \bigcup \{[x]_{B_i} \in U/GB(B_i) \mid \\ &[x]_{GB(B_1)} \cap X \neq \emptyset \wedge \\ &[x]_{GB(B_2)} \cap X \neq \emptyset \wedge \dots \wedge \\ &[x]_{GB(B_i)} \cap X \neq \emptyset\} \end{aligned} \quad (11)$$

$$\begin{aligned} \sum_{i=1}^m \underline{GBR_{B_i}}X &= \bigcup \{[x]_{B_i} \in U/GB(B_i) \mid \\ &[x]_{GB(B_1)} \subseteq X \vee [x]_{GB(B_2)} \subseteq \\ &X \vee \dots \vee [x]_{GB(B_i)} \subseteq X\} \end{aligned} \quad (12)$$

称 $(\sum_{i=1}^m \overline{GBR_{B_i}}X, \sum_{i=1}^m \underline{GBR_{B_i}}X)$ 为多粒度粒球粗糙集,并称 $POS_B(X) = \overline{GBR_B}(X)$ 为 X 关于 B 的正域, $BN_B(X) = \overline{GBR_B}(X) - \underline{GBR_B}(X)$ 为 X 关于 B 的边界域。

性质 1 设 $DS = \langle U, AT, V, f \rangle$ 是一个完备的信息系统, U 是非空有限集合, AT 是属性集。任意 $B \in 2^{AT}$, $B = \{B_1, B_2, \dots, B_m\}$, $B_i \in B$,在 U 上存在基于 B 和 B_i 的粒球 GB 和 GB_{B_i} ,其相应的等价关系为 GBR_B, GBR_{B_i} 。对任意 $X \subseteq U$, X 关于 B 的上、下近似有以下性质:

$$1) \sum_{i=1}^m \underline{GBR_{B_i}}X \subseteq \underline{GBR_B}X;$$

$$2) \sum_{i=1}^m \overline{GBR_{B_i}}X \supseteq \overline{GBR_B}X.$$

证明 1) 由定义 8 得,任意 $[x]_{B_i} \subseteq \sum_{i=1}^m \underline{GBR_{B_i}}X$,

则至少存在 $i = j (j \leq m)$, 使得 $[x]_{B_i} \subseteq \sum_{i=1}^m \underline{GBR}_{B_i} X$ 。

又因 $X/GB(B) = \{X'_1, X'_2, \dots, X'_b\}$, 则任意 $X'_i (i \leq b) \subseteq GB(B)$, 存在 $X_j (j \leq a) \subseteq GB(B_i)$, 使得

$X'_i \subseteq X_j$, 则 $GB(B) \subseteq \underline{GBR}_B X$ 。综上, 有 $\sum_{i=1}^m \underline{GBR}_{B_i} X \subseteq \underline{GBR}_B X$ 。

2) 由 Pawlak 粗糙集理论的相关性质可得 $\underline{GBR}_B X = \sim (\sim \underline{GBR}_B) X$, 又根据性质 (1) 有

$$\begin{aligned} \underline{GBR}_B X &= \\ &\sim (\sim \underline{GBR}_B) X \subseteq \sim (\sim \sum_{i=1}^m \underline{GBR}_{B_i}) X = \\ &\sum_{i=1}^m \underline{GBR}_{B_i} X. \end{aligned}$$

性质 2 设 $DS = \langle U, AT, V, f \rangle$ 是一个完备的信息系统, U 是非空有限集合, AT 是属性集。任意 $B \in 2^{AT}$, $B = \{B_1, B_2, \dots, B_m\}$, 在 U 上存在基于 B_i 的粒球 GB_{B_i} 相应的等价关系 GBR_{B_i} 。假设有 $X \subseteq U$, 则 X 关于属 B 的上、下近似的性质有

$$\begin{aligned} 1) & \sum_{i=1}^m \underline{GBR}_{B_i} X \subseteq X \subseteq \sum_{i=1}^m \overline{GBR}_{B_i} X; \\ 2) & \sum_{i=1}^m \underline{GBR}_{B_i} \emptyset = \emptyset = \sum_{i=1}^m \overline{GBR}_{B_i} \emptyset; \\ & \sum_{i=1}^m \underline{GBR}_{B_i} U = U = \sum_{i=1}^m \overline{GBR}_{B_i} U; \\ 3) & \sum_{i=1}^m \underline{GBR}_{B_i} (\sim X) = \sim \sum_{i=1}^m \overline{GBR}_{B_i} (X); \\ & \sum_{i=1}^m \overline{GBR}_{B_i} (\sim X) = \sim \sum_{i=1}^m \underline{GBR}_{B_i} (X); \\ 4) & \sum_{i=1}^m \underline{GBR}_{B_i} X = \underline{GBR}_{B_1} X \cup \dots \cup \underline{GBR}_{B_m} X; \\ 5) & \sum_{i=1}^m \overline{GBR}_{B_i} (\sim X) = \overline{GBR}_{B_1} X \cap \dots \cap \\ & \overline{GBR}_{B_m} X. \end{aligned}$$

证明

1a) 令 $x, y \in \sum_{i=1}^m \underline{GBR}_{B_i} X$, 且有 $X/GB(B_i) \subseteq X$, 则至少存在一个划分 $X/GB(B_i)$, 使得 $x \in X/GB(B_i)$, 且 $y \in X/GB(B_i)$; 又因 $x, y \in X$, 则 $\sum_{i=1}^m \underline{GBR}_{B_i} X \subseteq X$ 。

1b) 令 $x, y \in X$, 则有 $x \in X/GB(B_i) \cap X$, 且 $y \in X/GB(B_i) \cap X$; 又 $X/GB(B_i) \cap X = \emptyset$, 所以 $x, y \in \sum_{i=1}^m \overline{GBR}_{B_i} X$, $X \subseteq \sum_{i=1}^m \overline{GBR}_{B_i} X$ 。

2a) 根据 1) 可得 $\sum_{i=1}^m \underline{GBR}_{B_i} \emptyset \subseteq \emptyset$, 又有 $\emptyset \subseteq \sum_{i=1}^m \underline{GBR}_{B_i} \emptyset$ (空集是任何集合的子集), 则有 $\emptyset = \sum_{i=1}^m \underline{GBR}_{B_i} \emptyset$; 假设 $U = \sum_{i=1}^m \underline{GBR}_{B_i} U$, 据定义有 $X/\underline{GBR}_{B_i} \cap \emptyset \neq \emptyset$, 而 $X/\underline{GBR}_{B_i} \cap \emptyset = \emptyset$, 这与假设矛盾, 因此有 $\sum_{i=1}^m \overline{GBR}_{B_i} \emptyset = \emptyset$ 。

2b) 根据 1) 可得 $\sum_{i=1}^m \underline{GBR}_{B_i} U \subseteq U$, 若 $x \in U$, $x/\underline{GBR}_{B_i} \subseteq U$ 。则 $x \in \sum_{i=1}^m \underline{GBR}_{B_i} U$, $U \subseteq \sum_{i=1}^m \underline{GBR}_{B_i} U$, 因此, $U = \sum_{i=1}^m \underline{GBR}_{B_i} U$ 。

$$3a) \sum_{i=1}^m \underline{GBR}_{B_i} (\sim X) = \sim \sum_{i=1}^m \overline{GBR}_{B_i} (X)。$$

3b) 根据 3a), 令 $X = \sim X$, 则

$$\begin{aligned} \sum_{i=1}^m \overline{GBR}_{B_i} (\sim X) &= \sum_{i=1}^m \overline{GBR}_{B_i} (\sim (\sim X)) = \\ &\sim \sum_{i=1}^m \underline{GBR}_{B_i} (\sim (\sim X)) = \\ &\sim \sum_{i=1}^m \underline{GBR}_{B_i} (X); \end{aligned}$$

4) 任意 $x \subseteq U$, 如果有 $X/GB(B_i) \subseteq X$, 则 $x \in \sum_{i=1}^m \underline{GBR}_{B_i} X$, 若存在 y , 使

$$\begin{aligned} y \in & \sum_{i=1}^m \underline{GBR}_{B_i} X - \cup_{x \in U} X/GB(B_1) - \\ & \cup_{x \in U} X/GB(B_2) - \dots - \\ & \cup_{x \in U} X/GB(B_m) = \emptyset。 \end{aligned}$$

则 $X/GB(B_i) = \emptyset$, 所以

$$\sum_{i=1}^m \underline{GBR}_{B_i} X = \underline{GBR}_{B_1} X \cup \underline{GBR}_{B_2} X \cup \dots \cup \underline{GBR}_{B_m} X。$$

5) 根据性质 4) 可得

$$\begin{aligned} \sum_{i=1}^m \overline{GBR}_{B_i} (\sim X) &= \sim \sum_{i=1}^m \underline{GBR}_{B_i} (\sim X) = \\ &\sim (\underline{GBR}_{B_1} (\sim X) \cup \dots \cup \underline{GBR}_{B_m} (\sim X)) = \\ &\sim (\sim \overline{GBR}_{B_1} X \cup \dots \cup \sim \overline{GBR}_{B_m} X) = \\ &\overline{GBR}_{B_1} X \cap \overline{GBR}_{B_2} X \cap \dots \cap \overline{GBR}_{B_m} X。 \end{aligned}$$

性质 3 设 $DS = \langle U, AT, V, f \rangle$ 是一个完备的信息系统, U 是非空有限集合, AT 是属性集。任意 $B \in 2^{AT}$, $B = \{B_1, B_2, \dots, B_m\}$, 在 U 上存在基于 B_i 的粒球 GB_{B_i} 相应的等价关系 GBR_{B_i} 。假设有 X, Y

$\subseteq U$, 则 X, Y 关于 B 的上、下近似的性质有

- 1) $\sum_{i=1}^m \overline{GBR_{B_i}}(X \cap Y) = (\overline{GBR_{B_1}} \cap \overline{GBR_{B_1}}Y) \cup \cdots \cup (\overline{GBR_{B_m}} \cap \overline{GBR_{B_m}}Y);$
- 2) $\sum_{i=1}^m \overline{GBR_B}(X \cap Y) = (\overline{GBR_{B_1}}X \cap \overline{GBR_{B_1}}Y) \cup \cdots \cup (\overline{GBR_{B_m}}X \cap \overline{GBR_{B_m}}Y);$
- 3) $\sum_{i=1}^m \overline{GBR_{B_i}}(X \cap Y) \subseteq \sum_{i=1}^m \overline{GBR_{B_i}}X \cap \sum_{i=1}^m \overline{GBR_{B_i}}Y;$
- 4) $\sum_{i=1}^m \overline{GBR_{B_i}}(X \cap Y) \supseteq \sum_{i=1}^m \overline{GBR_{B_i}}X \cap \sum_{i=1}^m \overline{GBR_{B_i}}Y;$
- 5) $X \subseteq Y \Rightarrow \sum_{i=1}^m \overline{GBR_{B_i}}X \subseteq \sum_{i=1}^m \overline{GBR_{B_i}}Y;$
- 6) $X \subseteq Y \Rightarrow \sum_{i=1}^m \overline{GBR_{B_i}}X \subseteq \sum_{i=1}^m \overline{GBR_{B_i}}Y;$
- 7) $\sum_{i=1}^m \overline{GBR_{B_i}}(X \cup Y) \supseteq \sum_{i=1}^m \overline{GBR_{B_i}}X \cup \sum_{i=1}^m \overline{GBR_{B_i}}Y;$
- 8) $\sum_{i=1}^m \overline{GBR_{B_i}}(X \cap Y) \subseteq \sum_{i=1}^m \overline{GBR_{B_i}}X \cap \sum_{i=1}^m \overline{GBR_{B_i}}Y.$

证明 先证明 1)。

$$\begin{aligned} \sum_{i=1}^m \overline{GBR_{B_i}}(X \cap Y) &= \overline{GBR_{B_1}}(X \cap Y) \cup \overline{GBR_{B_2}}(X \cap Y) \cup \cdots \cup \overline{GBR_{B_i}}(X \cap Y) = \\ &= (\overline{GBR_{B_1}}X \cap \overline{GBR_{B_1}}Y) \cup \cdots \cup (\overline{GBR_{B_m}}X \cap \overline{GBR_{B_m}}Y). \end{aligned}$$

2) 与 1) 类似可以证得。

$$\begin{aligned} \sum_{i=1}^m \overline{GBR_{B_i}}(X \cap Y) &= \overline{GBR_{B_1}}(X \cap Y) \cup \overline{GBR_{B_2}}(X \cap Y) \cup \cdots \cup \overline{GBR_{B_m}}(X \cap Y) = \\ &= (\overline{GBR_{B_1}}X \cap \overline{GBR_{B_1}}Y) \cup \cdots \cup (\overline{GBR_{B_m}}X \cap \overline{GBR_{B_m}}Y). \end{aligned}$$

由 1) 可以得出 3) 的证明。

$$\begin{aligned} \sum_{i=1}^m \overline{GBR_{B_i}}(X \cap Y) &= \\ &= (\overline{GBR_{B_1}}X \cap \overline{GBR_{B_1}}Y) \cup \cdots \cup \end{aligned}$$

$$\begin{aligned} &(\overline{GBR_{B_m}}X \cap \overline{GBR_{B_m}}Y) = \\ &= ((\overline{GBR_{B_1}}X \cap \overline{GBR_{B_1}}Y) \cup \overline{GBR_{B_2}}X \cup \cdots \cup \overline{GBR_{B_m}}X) \cap \\ &= ((\overline{GBR_{B_1}}X \cap \overline{GBR_{B_1}}Y) \cup \overline{GBR_{B_2}}Y \cup \cdots \cup \overline{GBR_{B_m}}Y) = \\ &= (\overline{GBR_{B_1}}X \cup \overline{GBR_{B_2}}X \cup \cdots \cup \overline{GBR_{B_m}}X) \cap \\ &= (\overline{GBR_{B_1}}Y \cup \overline{GBR_{B_2}}X \cup \cdots \cup \overline{GBR_{B_m}}X) \cap \\ &= (\overline{GBR_{B_1}}X \cup \overline{GBR_{B_2}}Y \cup \cdots \cup \overline{GBR_{B_m}}Y) \cap \\ &= (\overline{GBR_{B_1}}Y \cup \overline{GBR_{B_2}}Y \cup \cdots \cup \overline{GBR_{B_m}}Y) = \\ &= \sum_{i=1}^m \overline{GBR_{B_i}}X \cap \sum_{i=1}^m \overline{GBR_{B_i}}Y \cap \\ &= (\overline{GBR_{B_1}}Y \cup \sum_{i=2}^m \overline{GBR_{B_i}}X) \cap \\ &= (\overline{GBR_{B_1}}X \cup \sum_{i=2}^m \overline{GBR_{B_i}}Y) \subseteq \\ &= \sum_{i=1}^m \overline{GBR_{B_i}}X \cap \sum_{i=1}^m \overline{GBR_{B_i}}Y. \end{aligned}$$

由 2) 可得出 4) 的证明。

$$\begin{aligned} \sum_{i=1}^m \overline{GBR_{B_i}}(X \cap Y) &= \\ &= (\overline{GBR_{B_1}}X \cap \overline{GBR_{B_1}}Y) \cup (\overline{GBR_{B_2}}X \cap \overline{GBR_{B_2}}Y) \cup \cdots \cup \\ &= (\overline{GBR_{B_m}}X \cap \overline{GBR_{B_m}}Y) = \\ &= ((\overline{GBR_{B_1}}X \cap \overline{GBR_{B_1}}Y) \cap \overline{GBR_{B_2}}X \cap \cdots \cap \overline{GBR_{B_m}}X) \cup \\ &= ((\overline{GBR_{B_1}}X \cap \overline{GBR_{B_1}}Y) \cap \overline{GBR_{B_2}}Y \cap \cdots \cap \overline{GBR_{B_m}}Y) = \\ &= (\overline{GBR_{B_1}}X \cap \overline{GBR_{B_2}}X \cap \cdots \cap \overline{GBR_{B_i}}X) \cup \\ &= (\overline{GBR_{B_1}}Y \cap \overline{GBR_{B_2}}X \cap \cdots \cap \overline{GBR_{B_m}}X) \cup \\ &= (\overline{GBR_{B_1}}X \cap \overline{GBR_{B_2}}Y \cap \cdots \cap \overline{GBR_{B_m}}Y) \cup \\ &= (\overline{GBR_{B_1}}Y \cap \overline{GBR_{B_2}}Y \cap \cdots \cap \overline{GBR_{B_m}}Y) = \\ &= \sum_{i=1}^m \overline{GBR_{B_i}}X \cup \sum_{i=1}^m \overline{GBR_{B_i}}Y \cup \\ &= (\overline{GBR_{B_1}}Y \cap \sum_{i=2}^m \overline{GBR_{B_i}}X) \cup \\ &= (\overline{GBR_{B_1}}X \cap \sum_{i=2}^m \overline{GBR_{B_i}}Y). \end{aligned}$$

5) 如果 $X \subseteq Y$, $X \cap Y = X$, 则根据性质 3)

可得

$$\begin{aligned} \sum_{i=1}^m \overline{GBR_{B_i}}(X \cap Y) &= \\ &= \sum_{i=1}^m \overline{GBR_{B_i}}X \subseteq \sum_{i=1}^m \overline{GBR_{B_i}}X \cap \sum_{i=1}^m \overline{GBR_{B_i}}Y, \end{aligned}$$

则有 $\sum_{i=1}^m \overline{GBR_{B_i}}X \subseteq \sum_{i=1}^m \overline{GBR_{B_i}}Y$ 。

6) 若 $X \subseteq Y, X \cup Y = Y$, 则根据性质 4) 可得

$$\begin{aligned} \sum_{i=1}^m \overline{GBR_{B_i}}(X \cup Y) &= \\ \sum_{i=1}^m \overline{GBR_{B_i}}Y &\supseteq \sum_{i=1}^m \overline{GBR_{B_i}}X \cup \\ \sum_{i=1}^m \overline{GBR_{B_i}}Y, \end{aligned}$$

则有 $\sum_{i=1}^m \overline{GBR_{B_i}}X \subseteq \sum_{i=1}^m \overline{GBR_{B_i}}Y$ 。

7) $X \subseteq X \cup Y$ 且 $Y \subseteq X \cup Y$ 。则根据性质

5), 满足 $\sum_{i=1}^m \overline{GBR_{B_i}}X \subseteq \sum_{i=1}^m \overline{GBR_{B_i}}(X \cup Y)$, 且有

$$\sum_{i=1}^m \overline{GBR_{B_i}}Y \subseteq \sum_{i=1}^m \overline{GBR_{B_i}}(X \cup Y), \text{ 则有}$$

$$\sum_{i=1}^m \overline{GBR_{B_i}}X \cup \sum_{i=1}^m \overline{GBR_{B_i}}Y \subseteq \sum_{i=1}^m \overline{GBR_{B_i}}(X \cup Y)。$$

8) 根据性质 7), 显然 $X \cap Y \subseteq X, X \cap Y \subseteq Y$,

根据 $X \cap Y \subseteq X$ 则有

$$\begin{aligned} \sum_{i=1}^m \overline{GBR_{B_i}}X &\supseteq \sum_{i=1}^m \overline{GBR_{B_i}}(X \cap Y)X \cap Y \subseteq Y, \\ \sum_{i=1}^m \overline{GBR_{B_i}}Y &\supseteq \sum_{i=1}^m \overline{GBR_{B_i}}(X \cap Y)。 \end{aligned}$$

综上所述,

$$\begin{aligned} \sum_{i=1}^m \overline{GBR_{B_i}}(X \cap Y) &\subseteq \\ \sum_{i=1}^m \overline{GBR_{B_i}}X \cap \sum_{i=1}^m \overline{GBR_{B_i}}Y. \end{aligned}$$

基于以上性质, 将粒球粗糙集模型拓展为多粒度粒球粗糙集模型。其中, 该模型的集合近似通过论域上的多个等价关系来定义。

定义 9 设 $DS = \langle U, AT \cup D, V, f \rangle$ 是一个完备的决策信息系统, 决策属性 D 将论域 U 划分为 L 个等价类, 表示为 $U/D = \{X_1, X_2, \dots, X_L\}$ 。任意 $B \in 2^{AT}, B = \{B_1, B_2, \dots, B_m\}$, 在 U 上存在基于 B_i 的粒球 GB_{B_i} 相应的等价关系 $GBR_{B_i} \circ D$ 关于 B 的上下近似定义为

$$\overline{GBR_B}D = \bigcup_{i=1}^L \overline{GBR_{B_i}}X_i \quad (13)$$

$$\underline{GBR_B}D = \bigcup_{i=1}^L \underline{GBR_{B_i}}X_i \quad (14)$$

与粒球粗糙集相似, 可以根据经典粗糙集理论得出粒球粗糙集的正域、边界域、边界域的定义。

定义 10 设 $DS = \langle U, AT \cup D, V, f \rangle$ 是一个完备的决策信息系统, $\forall B \subseteq AT, D$ 关于 B 的正域

和边界域定义为

$$POS_B(D) = \overline{GBR_B}(D) \quad (15)$$

$$BN_B(D) = \overline{GBR_B}(D) - \underline{GBR_B}(D) \quad (16)$$

根据上述定义和性质, 基于 Xia 等的粒球粗糙集模型计算正域的算法^[20], 结合多粒度粗糙集的理论, 设计出多粒度粒球粗糙集计算正域的算法如下。

算法 1 多粒度粒球粗糙集在第 i 个粒度下正域生成

输入 $DS = \langle U, AT \cup \{d\}, f \rangle / *$ 完备决策信息系统 $*/$;

$k / *$ DS 的类别个数 $*/$;

$GBS / *$ 用于临时储存 GB_n $*/$;

$size(GB) / *$ 最小粒径 $*/$;

$Purity / *$ 纯度阈值 $*/$ 。

输出 $GBRS_i / *$ 存储满足条件的 GB $*/$ 。

步骤 1 初始化。

步骤 2 对数据集 $DS = \langle U, AT \cup D, V, f \rangle$ 进行划分, 设 $AT = \{AT_1, AT_2, \dots, AT_m\}, DS_i = \langle U, AT_i \cup D, V, f \rangle$ 为 DS 的第 i 个子信息系统。

步骤 3 对 DS_i 进行 k -means 聚类分析。计算各簇心 C_n 和粒球半径 r_n , 绘制粒球 $GB_n \circ / *$ 将 GB_n 保存在 GBS 内 $*/$ 。

步骤 4 if $\exists GB_n \in GBS, Purity < 1$, do

1) 生成粒球 GB_n 内的数据集 DS'_n ;

2) 返回步骤 3。

until $\forall GB_n \in GBS, Purity = 1$ or

$size(GB_i) < size(GB)$ 。

$GBRS_i \leftarrow GB_n$ 。

end

步骤 5 输出 $GBRS_i (i = 1, 2, \dots, m)$ 。

在算法 1 中, 假设论域 U 被分解为 $|GBS|$ 个粒球, 迭代次数最多为 $(|U| \times |GBS|)$ 。因此, 算法 1 的时间复杂度为 $O(|U| \times |GBS|)$ 。

算法 2 多粒度粒球粗糙集正域生成

输入 $POS_GBRS_i (i = 1, 2, \dots, m)$ 。

输出 $POS_MGBRST / *$ 多粒度粒球粗糙集正域 $*/$ 。

步骤 1 for $i = 1 : m$

if $\exists GB \in GBRS_i, Purity < 1$, do

$POS_GBRS_i = GBRS_i \setminus GB$

end

end

步骤 2 对 $\forall POS_GBRS_i$, do

$$POS_MGBRS = POS_GBRS_1 \cup POS_GBRS_2 \cup \cdots \cup POS_GBRS_m。$$

步骤3 输出 $POS_MGBRS。$

由算法1可知,在最坏的情况下,多粒度粒球粗糙集在第*i*个粒度下正域生成所需的迭代次数为 $O(|U| \times |GBS|)$ 。此外,假定在粒度划分时,粒度划分为 $AT = \{AT_1, AT_2, \cdots, AT_m\}$ 。结合算法1可得算法2的时间复杂度为 $O(|U| \times |GBS| \times |m|)$ 。

3 案例的说明与比较分析

例1 表1是一个多专家面试学生的决策信息表。其中: $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ 表示6位不同的学生;条件属性集 $B = \{a_1, a_2, a_3, a_4, a_5, a_6\}$ 表示生物、管理领域的专家团队对学生是否通过面试的赞成率;决策属性 $D = \{d\}$ 表示学生能否通过面试。其中:“1”表示通过面试,“0”则表示不通过面试。结合相关性质,可将属性集 B 分成 B_1 和 $B_2。$ 其中 B_1 表示生物领域, B_2 表示管理领域,即 $B = \{B_1, B_2\} = \{\{a_1, a_2, a_3\}, \{a_4, a_5, a_6\}\}$,且设最小粒径 $size(GB) = 2$ (决定粒球 GB 是否继续迭代的最小样本数)。

表1 多专家面试学生评估表

Tab.1 Multi-expert interview student assessment form

	a_1	a_2	a_3	a_4	a_5	a_6	d
x_1	0.6	0.7	0.7	0.7	0.6	0.7	0
x_2	0.8	0.8	0.7	0.6	0.6	0.6	1
x_3	0.7	0.6	0.6	0.7	0.9	0.7	0
x_4	0.8	0.7	0.7	0.8	0.7	0.8	1
x_5	0.6	0.8	0.8	0.6	0.8	0.7	0
x_6	0.6	0.8	0.8	0.6	0.7	0.8	0

根据定义2可求得在粒度 B_1 和 B_2 下,粒球 GB_i 的球心 c_i 、半径 $r_i。$ 各样本到球心 C_i 的距离 d 如表2所示。

$C_1 = (0.68, 0.73, 0.68);$
 $C_2 = (0.67, 0.72, 0.72);$
 $r_1 = 0.0197; r_2 = 0.0217。$

则根据定义4和表2可以求得粒球 GB_i 的样本集如下。

$GB_1 = \{x_1, x_2, x_4\};$
 $GB_2 = \{x_1, x_5, x_6\}。$

表2 各样本到中心的距离

Tab.2 Distance from each sample to center

	c_1	c_2
x_1	0.0077	0.0157
x_2	0.0197	0.0337
x_3	0.0237	0.0337
x_4	0.0157	0.0237
x_5	0.0257	0.0117
x_6	0.0257	0.0117

根据定义3计算粒球 $GB_1、GB_2$ 的纯度为

$Purity(GB_1) = 0.6667,$
 $Purity(GB_2) = 1。$

又因为 $Purity(GB_1) = 0.6667 < 1;$ 且 $size(GB_1) \geq 2,$ 因此,对 GB_1 进行第二次迭代,根据定义2计算簇心和半径可得

$C_{11} = (0.73, 0.73, 0.70), r_{11} = 0.01。$

同理根据定义4可得

$GB_{11} = \{x_2, x_4\},$
 $Purity(GB_{11}) = 1。$

根据定义8,可以求得论域 U 在属性集 B 下近似为

$GBR_B D = \{x_1, x_2, x_4, x_5, x_6\},$

进而,可以得到 D 关于属性集 B 正域为

$POS_B(D) = \{x_1, x_2, x_4, x_5, x_6\}。$

注1 例1中粒球 GB_i 使用的最小粒径是粒球内样本个数为2(样本分为2类),在实际计算中,最小粒径的取值可以大于类别个数,否则会出现样本都是正域的情况,不利于正确地分类。

注2 在 k -means 聚类算法中,使用的距离度量是 Squared Euclidean Distance。因此,例1使用相同的度量。

注3 例1中计算簇心的方法是利用对所有 $f(x, c_i), i = 1, 2 \cdots, m$ 取均值。实验中可以根据样本的类别标签,分别对所分的簇求均值从而得到簇心和半径。

采用多粒度粗糙集模型进行计算过程如表3所示,所得正域的结果与多粒度粒球粗糙集进行对比,结果如表4所示。

通过对比和分析例1在粒球粗糙集模型(MGRST)和多粒度粒球粗糙集模型(MGBRST)正域生成的结果(见图1),根据表3,可以得到 $POS_B(D) _ MGBRST \subseteq POS_B(D) _ MGRST。$ 这是因为MGBRST模型结合了粒球粗糙集更细致的描述能力,更好地刻画了数据之间的内在联系,减少部分目标区域的样本,便于进一步决策。

表 3 例 1 在多粒度粗糙集模型下的计算结果

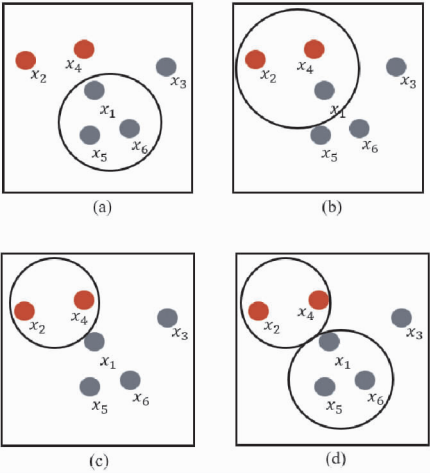
Tab.3 Calculation results of example 1 under multi-granulation rough set model

	在各属性下的划分	下近似
a_1	$\{\{x_1, x_5, x_6\}, \{x_2, x_4\}, \{x_3\}\}$	$\{x_3\}$
a_2	$\{\{x_1, x_4\}, \{x_2, x_5, x_6\}, \{x_3\}\}$	$\{x_4\}$
a_3	$\{\{x_1, x_2, x_4\}, \{x_3\}, \{x_5, x_6\}\}$	$\{x_3, x_5, x_6\}$
a_4	$\{\{x_1, x_3\}, \{x_2, x_5, x_6\}, \{x_4\}\}$	$\{x_1, x_3, x_4\}$
a_5	$\{\{x_1, x_2\}, \{x_3\}, \{x_4, x_6\}, \{x_5\}\}$	$\{x_3, x_5\}$
a_6	$\{\{x_1, x_3, x_5\}, \{x_2\}, \{x_4, x_6\}\}$	$\{x_1, x_2, x_3, x_5\}$
d	$\{\{x_1, x_3, x_5, x_6\}, \{x_2, x_4\}\}$	$\{x_1, x_2, x_3, x_4, x_5, x_6\}$

表 4 例 1 在两个模型下计算正域的结果

Tab.4 Positive region calculating results of example 1 under two models

模型	$POS_B(D)$
MRST	$\{x_1, x_2, x_3, x_4, x_5, x_6\}$
MGBRST	$\{x_1, x_2, x_4, x_5, x_6\}$



注:红色表示决策属性为“1”;蓝色表示决策属性为“0”。(a)、(b)表示论域 U 分别在粒度 B_1 和 B_2 下的划分。根据纯度的定义,可以得到(a)中的粒球纯度为 1,则表示正域,若粒球纯度不为 1,如(b),则表示边界域。(c)表示对(b)粒球进行迭代,根据定义,最终结果如(d),可以得到最终多粒度粒球粗糙集的正域。

图 1 例 1 的多粒度粒球粗糙集的生成过程

Fig.1 The generation process of granular-ball rough set model based on multi-granulation in example 1

4 实验结果与分析

对于本文提出的多粒度粒球粗糙集模型生成正域算法,本节通过实验分析验证其可行性和有

效性。所有实验硬件环境配置为 AMD Ryzen 5 2500U CPU@ 2.00 GHz 和 8 GB 内存的个人计算机,算法运行的软件环境为 Matlab R2021b。

为了验证所提算法的有效性,在本节中,从 UCI 中选取了 10 组基准数据集进行相关的实验对比分析,其中前 4 组为连续型数据,后 6 组为离散型数据。数据的具体描述见表 5。

表 5 数据集描述

Tab.5 Data sets description

ID	数据集名称	样本数	属性数	决策类数
1	Wine	178	13	3
2	Iris	150	4	3
3	Segmentation	2 310	19	7
4	Pima	768	8	2
5	Mushroom	7 535	22	2
6	Zoo	101	16	7
7	Lymphography	148	18	4
8	Mushroom1	8 124	22	2
9	German	1 000	21	2
10	Anneal	798	39	5

4.1 参数分析

在数学实验中,选择最优的参数进行对比实验是一个重要的步骤。多粒度粒球粗糙集模型(MGBRST)生成正域算法中需要考虑的参数为最小粒径 $size(GB)$,即颗粒球 GB 内的最小样本数,它决定了颗粒球 GB 是否继续迭代;由于正域生成算法设置纯度阈值为“1”,因此实验不考虑纯度阈值的影响。选取 German、Zoo、Lymphography、Anneal 进行参数分析, $size(GB)$ 的选择范围为 2~7。考虑参数对算法的时间消耗/s 和包含度/% 的影响,可视化结果如图 2、3 所示。

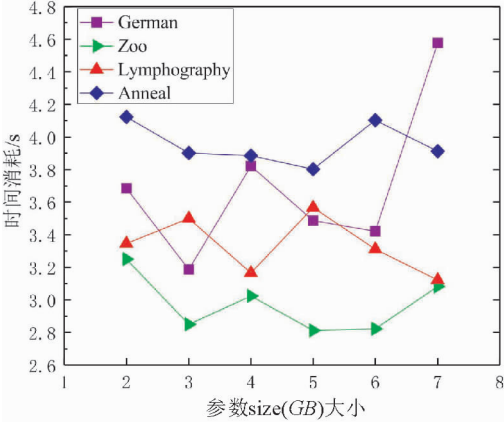


图 2 参数 $size(GB)$ 的大小对时间消耗的影响

Fig.2 The effect of parameter $size(GB)$ on time consumption

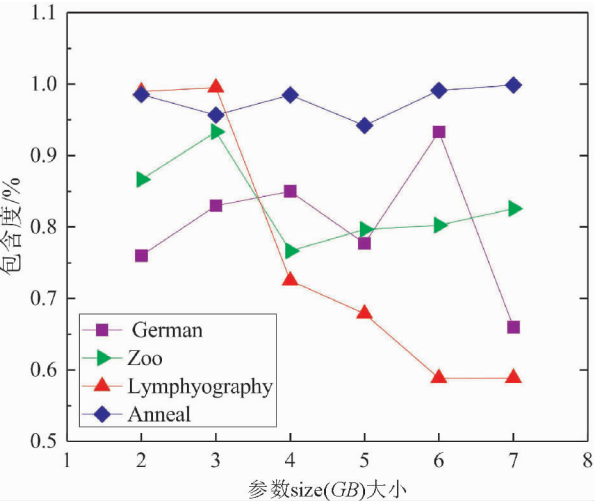


图3 参数 size(GB)的大小对包含度的影响

Fig.3 The effect of parameter size (GB) on inclusion

包含度的定义如下:在多粒度粒球粗糙集模型(MGBRST)中,由于粒球刻画了数据之间的内在联系和规律,因此生成的正域的样本数从理论上来说将少于多粒度粗糙集(MGRST),并且可能包含在多粒度粗糙集(MGRST)中。故本文根据梁等提出的思想,即包含度可作为建立粗糙集数据分析中的度量的主要依据^[26],定义在MGBRST模型上正域生成的样本关于MGRST模型的包含度 α 为

$$\alpha(X_{MG}/X_{MGB}) = \frac{|X_{MG} \cap X_{MGB}|}{|X_{MGB}|} \quad (17)$$

其中: X_{MG} 表示在MGRST模型下生成正域的样本集; X_{MGB} 表示在MGBRST模型下生成正域的样本集; $|X_{MGB}|$ 表示集合 X_{MGB} 的基数。

参数越小,迭代的次数越多,时间消耗理论上就越大;除此外,参数的设定还与数据类别有关,且k-means算法具有随机性,因此找到合适的参数是必要的。如图2所示,参数size(GB)的大小对4个数据集的影响不同。数据集Anneal受参数的影响较小,而数据集German受参数的影响较大。参数的大小还影响了包含度的大小,最小粒径越小,迭代的次数越多,理论上得到的粒球就越多,包含度可能就越大;此外,参数过小也可能出现生成单点集的情况,这样也一定程度上减少了正域样本的生成,导致包含度下降。如图3所示数据集Anneal受参数的影响较小,而数据集Lymphography受参数的影响较大,曲线呈大幅度下降趋势。综上所述,结合图2、3,选择最优的参数。如German数据集中选择size(GB)的值为4,能更好地结合时间消耗小和包含度高这两个优

点。各数据集的参数大小选择如表6所示。

表6 各数据集的参数选择

Tab.6 Parameter selection for each dataset

ID	Size(GB)
1	3
2	2
3	6
4	3
5	5
6	3
7	3
8	5
9	4
10	7

4.2 正域生成的时间消耗对比

在本节实验中,将针对钱等提出的多粒度粗糙集模型(MGRST)^[12]及本文提出的多粒度粒球粗糙集模型(MGBRST),进行时间消耗的对比,最终采集的时间消耗为两种方法分别求解正域所需时间的均值,具体结果如表7和图4所示。

表7 两种模型计算正域的时间消耗

Tab.7 Two models calculate the time consumption of positive region

ID	MGRST/s	MGBRST/s
1	0.541 091	0.484 320
2	0.271 866	0.179 734
3	3.460 916	2.414 930
4	1.596 670	1.130 863
5	12.135 100	12.522 220
6	3.499 518	2.913 096
7	1.882 662	3.564 317
8	23.741 600	22.754 823
9	3.070 070	3.461 867
10	4.122 445	3.978 553
均值	5.429 494	5.340 472

在数据的预处理阶段,由于MGRST模型处理的是离散型数据,因此,对前4组数据使用等宽离散化方法进行离散化,再利用MGRST模型计算离散数据。MGBRST模型处理前4组连续数据以及后6组离散数据。

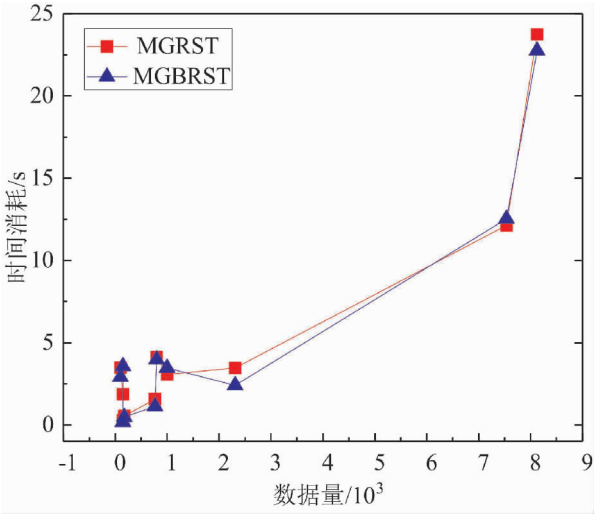


图 4 两种算法计算正域的时间消耗

Fig. 4 Two algorithms calculate the time consumption of positive region

观察表 7 和图 4, 不难得出如下结论: 利用多粒度粒球粗糙集模型 (MGBRST) 计算正域的时间消耗大多低于传统的多粒度粗糙集模型 (MGRST)。这说明粒球与多粒度粗糙集的结合, 可以有效地提升计算正域的效率; 其次, 多粒度粒球粗糙集模型 (MGBRST) 不仅可以处理连续型数据, 还可以处理离散型数据。

4.3 正域生成的样本数和包含率对比

为了更好地对比 MGBRST 模型和 MGRST 模型正域生成的关系, 本节主要讨论两种模型在 6 个数据的正域样本数和包含度, 包含度的计算如式 (17) 所示。

实验选取后 6 个离散型数据, 分别利用两种模型对 4 个数据集求解, 可得 4 个数据集在两种模型下生成的正域的样本个数和包含度 α 的结果如表 8 和图 5、6 所示。

表 8 两种模型生成的正域样本个数和包含度 α 的值
Tab. 8 Number and inclusion degree α of positive region samples generated by the two models

ID	MGRST	MGBRST	α
5	3 151	1 264	0.999 2
6	61	38	1.000 0
7	20	7	0.914 4
8	4 896	3 884	0.918 6
9	653	110	0.922 4
10	538	96	1.000 0
均值	-	-	0.959 1

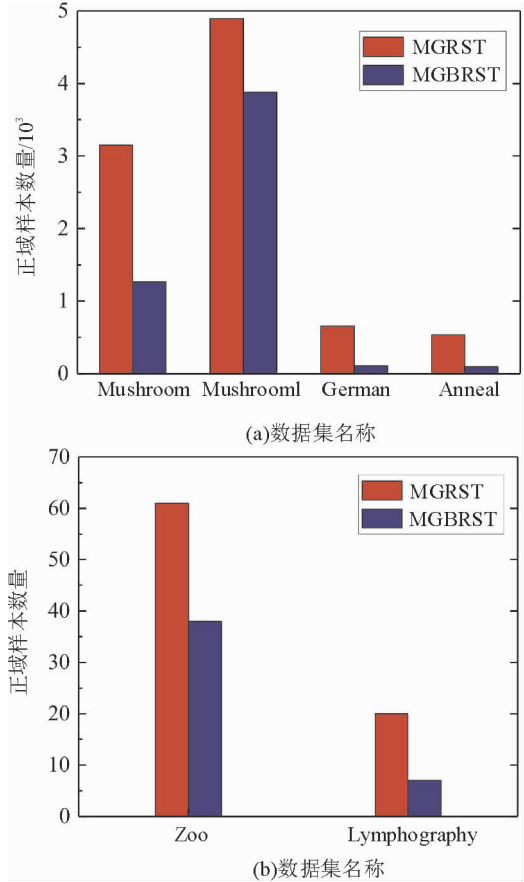


图 5 两种模型下生成的正域样本个数

Fig. 5 Number of positive region samples generated under two models

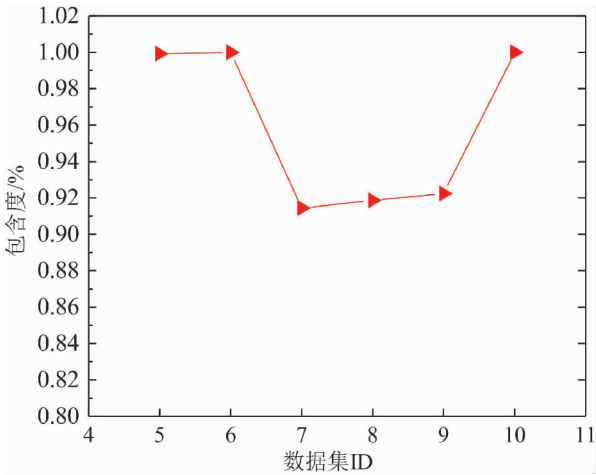


图 6 6 种数据集下的包含度

Fig. 6 Inclusion degree under six types of data set

观察表 8 和图 5、6 可知, 6 个数据集在 MGBRST 模型下生成正域的样本个数明显小于 MGRST 模型下生成正域的样本个数; 且包含度 α 基本上稳定在 0.9 以上, Zoo 数据集和 Anneal 数据集的包含度 α 为 1。这说明 MGBRST 模型是 MGRST 模型的一个拓展模型, 它结合了粒球粗糙

集更细致的描述能力,更好地刻画了数据之间的内在联系和规律,可以进一步划分正域中不被需要的那一部分。

5 结语

考虑到粒球粗糙集不能处理多粒度数据问题。为了解决该问题,本文将粒球思想与多粒度粗糙集模型结合,提出了多粒度粒球粗糙集模型,并讨论了多粒度粒球粗糙集的相关性质。此外,给出了该模型正域的计算算法。研究发现,多粒度粒球粗糙集能更好地刻画数据之间的内在联系,能进一步划分正域中的样本,减少样本数,便于决策者决策。最后,本文通过实验表明了多粒度粒球粗糙集模型的可行性和有效性。今后将进一步探索多粒度粒球粗糙集的相关研究,着重考虑基于该模型的粒度约简等问题。

参考文献

- [1] PAWLAK Z. Rough sets[J]. International Journal of Computer & Information Sciences, 1982, 11: 341-356.
- [2] WEI J M, WANG S Q, YUAN X J. Ensemble rough hypercuboid approach for classifying cancers [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(3): 381-391.
- [3] 马文萍,黄媛媛,李豪,等. 基于粗糙集与差分免疫模糊聚类算法的图像分割[J]. 软件学报, 2014, 25(11): 2675-2689.
MA W P, HUANG Y Y, LI H, et al. Image segmentation based on rough set and differential immune fuzzy clustering algorithm[J]. Journal of Software, 2014, 25(11): 2675-2689.
- [4] QIAN Y H, XU H, LIANG J Y, et al. Fusing monotonic decision trees[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(10): 2717-2728.
- [5] QIAN Y H, LIANG X Y, WANG Q, et al. Local rough set: A solution to rough data analysis in big data [J]. International Journal of Approximate Reasoning, 2018, 97: 38-63.
- [6] VIDHYA K A, GEETHA T V. Entity resolution framework using rough set blocking for heterogeneous web of data[J]. Journal of Intelligent & Fuzzy Systems, 2018, 34(1): 659-675.
- [7] HU M J, YAO Y Y. Structured approximations as a basis for three-way decisions in rough set theory[J]. Knowledge-Based Systems, 2019, 165: 92-109.
- [8] 梁吉业,钱宇华,李德玉,等. 面向大数据的粒计算理论与方法研究进展[J]. 大数据, 2016, 2(4): 13-23.
LIANG J Y, QIAN Y H, LI D Y, et al. Research development on granular computing theory and method for big data[J]. Big Data research, 2016, 2(4): 13-23.
- [9] KRYSZKIEWICZ M. Rough set approach to incomplete information systems [J]. Information Sciences, 1998, 112(1/2/3/4): 39-49.
- [10] 王国胤. Rough 集理论在不完备信息系统中的扩充 [J]. 计算机研究与发展, 2002, 39(10): 1238-1243.
WANG G Y. Extension of rough set under incomplete information systems[J]. Journal of Computer Research and Development, 2002, 39(10): 1238-1243.
- [11] STEFANOWSKI J, TSOUKIAS A. Incomplete information tables and rough classification [J]. Computational Intelligence, 2001, 17(3): 545-566.
- [12] 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.
- [13] QIAN Y H, LIANG J Y, DANG C Y. Incomplete multigranulation rough set [J]. IEEE Transactions on Systems Man and Cybernetics-Part A: Systems and Humans, 2010, 40(2): 420-431.
- [14] QIAN Y H, LI S Y, LIANG J Y, et al. Pessimistic rough set based decisions: A multigranulation fusion strategy [J]. Information Sciences, 2014, 264: 196-210.
- [15] HU Q H, YU D R, XIE Z X. Neighborhood classifiers [J]. Expert Systems with Applications, 2008, 34(2): 866-876.
- [16] 李楠,谢娟英. 基于邻域粗糙集的增量特征选择 [J]. 计算机技术与发展, 2011, 21(11): 149-152.
LI N, XIE J Y. A feature subset selection algorithm based on neighborhood rough set for incremental updating datasets [J]. Computer Technology and Development, 2011, 21(11): 149-152.
- [17] 胡清华,赵辉,于达仁. 基于邻域粗糙集的符号与数值属性快速约简算法 [J]. 模式识别与人工智能, 2008, 21(6): 732-738.
HU Q H, ZHAO H, YU D R. Efficient symbolic and numerical attribute reduction with neighborhood rough sets [J]. Pattern Recognition and Artificial Intelligence, 2008, 21(6): 732-738.
- [18] 彭潇然,刘遵仁,纪俊. 自适应的邻域粗糙集邻域大小取值方法[J]. 计算机应用研究, 2019, 36(1):

- 144-147.
- PENG X R, LIU Z R, JI J. Adaptable method for determining neighborhood size of neighborhood rough set [J]. *Application Research of Computers*, 2019, 36 (1): 144-147.
- [19] XIA S Y, LIU Y S, DING X, et al. Granular ball computing classifiers for efficient scalable and robust learning [J]. *Information Sciences*, 2019, 483 (1): 136-152.
- [20] XIA S Y, WANG C, WANG G Y, et al. GBRS: An unified model of Pawlak rough set and neighborhood rough set[EB/OL]. (2022-01-10) [2023-06-01]. <http://arxiv.org/abs/2201.03349>.
- [21] XIA S Y, PENG D W, MENG D Y, et al. Ball k-means: Fast adaptive clustering with no bounds[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(1): 87-99.
- [22] XIA S Y, ZHANG H, LI W H, et al. GBNRS: A novel rough set algorithm for fast adaptive attribute reduction in classification[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2022, 34(3): 1231-1242.
- [23] XIA S Y, ZHENG S Y, WANG G Y, et al. Granular ball sampling for noisy label classification or imbalanced classification[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, 34(4): 2144-2155.
- [24] 陈中华, 巴婧, 徐泰华, 等. 一种面向粒球粗糙集的快速约简求解方法[J]. *小型微型计算机系统*, 2023, 44(1): 24-29.
- CHEN Z H, BA J, XU T H, et al. Quick strategy for searching granular ball rough set based reduct[J]. *Journal of Chinese Computer Systems*, 2023, 44(1): 24-29.
- [25] XIE J, XIA S Y, WANG G Y, et al. GBMST: An efficient minimum spanning tree clustering based on granular-ball computing [EB/OL]. (2023-03-02) [2023-06-01]. <http://arxiv.org/abs/2303.01082>.
- [26] 梁吉业, 徐宗本, 李月香. 包含度与粗糙集数据分析中的度量[J]. *计算机学报*, 2001, 24(5): 544-547.
- LIANG J Y, XU Z B, LI Y X. Inclusion degree and measures of rough set data analysis[J]. *Chinese Journal of Computers*, 2001, 24(5): 544-547.

(编辑 张 欢)