Computer Science and Application 计算机科学与应用, 2025, 15(4), 301-308
Published Online April 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.154102

Hans汉斯

基于区间值决策系统的正域快速求解算法

张国辉

烟台大学计算机与控制工程学院，山东 烟台

收稿日期：2025年3月18日；录用日期：2025年4月17日；发布日期：2025年4月25日

摘　要

区间值决策系统在不确定性数据处理领域具有重要应用价值，而正域的高效计算是提升系统整体性能的关键。由于相容类的计算是正域求解的必要步骤，但是传统相容类计算方法在处理大规模数据时面临较高的计算复杂度，导致属性约简的效率显著下降。针对这一问题，本文提出了一种利用哈希思想的快速计算方法。该方法通过哈希函数对每个对象进行快速分区，缩小相容类的查找范围，减少冗余计算，从而显著提升了相容类的计算效率，以至于能够优化属性约简的整体性能。通过在8个UCI数据集上的实验验证，本文所提方法在计算速度上较传统方法具有明显优势，为区间值决策系统的高效属性约简算法提供了新思路。

关键词

粗糙集，区间值决策系统，相容类，正域

Fast Algorithms for Computing Positive Regions Based on Interval-Valued Decision Systems

Guohui Zhang

School of Computer and Control Engineering, Yantai University, Yantai Shandong

Received: Mar. 18th, 2025; accepted: Apr. 17th, 2025; published: Apr. 25th, 2025

Abstract

Interval-valued decision systems play a crucial role in uncertain data processing, where the efficient computation of the positive domain is essential for enhancing overall system performance. Since the computation of compatible classes is a necessary step in solving the positive domain, traditional methods for computing compatible classes face high computational complexity when handling

文章引用: 张国辉. 基于区间值决策系统的正域快速求解算法[J]. 计算机科学与应用, 2025, 15(4): 301-308.
DOI: 10.12677/csa.2025.154102

large-scale data, leading to a significant decline in attribute reduction efficiency. To address this issue, this paper proposes a fast computation method based on hashing techniques. By utilizing hash functions to rapidly partition each object, the proposed method narrows the search scope for compatible classes, reduces redundant computations, and significantly improves the efficiency of compatible class computation, thereby optimizing the overall performance of attribute reduction. Experimental validation on eight UCI datasets demonstrates that the proposed method achieves a significant speed advantage over traditional approaches, providing new insights for developing efficient attribute reduction algorithms in interval-valued decision systems.

Keywords

Rough Set, Interval-Valued Decision System, Compatible Classes, Positive Region

1. 引言

随着信息技术的飞速发展，数据规模和复杂性呈指数级增长，如何从海量、高维、不确定的数据中高效提取有价值的信息已成为数据挖掘与知识发现领域的关键研究问题。Pawlak 于 1982 年提出粗糙集理论[1]，该理论是一种处理不确定性和不完备信息的数学工具，随着技术的发展，这一工具在数据分析、模式识别和机器学习等领域展现了其重要价值，得到了广泛应用。区间值决策系统[2]-[4]作为粗糙集理论的重要扩展，能够有效处理区间特征的数据，在不确定数据分析和决策支持中发挥着重要作用。

在区间值决策系统中，属性约简[5]-[8]是核心任务之一，而正域中的相容类的求解是属性约简的关键步骤。然而随着数据规模的增大和维度的提升，传统的相容类计算方法逐渐暴露出计算效率较慢的问题，成为系统性能的主要瓶颈。因此提升相容类的计算效率对于优化区间值决策系统的整体性能具有重要意义。近年来，研究者们提出了多种属性约简的优化方法，如差别矩阵[9]-[12]、启发式算法等，但这些方法在处理大规模数据时仍存在性能不足的问题，因此提高属性约简的效率成为了研究热题。

针对现有方法的局限性，本文提出了一种基于区间值决策系统的正域快速求解算法。该算法通过哈希函数对对象进行快速分区，并利用对象在属性集合下相容类的单调性优化计算过程，从而显著减少冗余计算，提升相容类的计算效率，进而优化属性约简的整体性能。通过在 8 个 UCI 数据集上的实验验证，本文所提方法在计算速度上较传统方法具有明显优势，能够更好地满足大规模、高维度数据的处理需求，为区间值决策系统的高效计算提供了新的解决方案。

2. 基本概念

定义 1：给定一个四元组 $IVDS=(U, A=C\cup D, V, f)$ 为区间值决策系统，$U=\{x_1, x_2\cdots x_n\}$ 为论域，$C=\{a_1, a_2\cdots a_m\}$ 为条件属性集合，$D=\{d\}$ 为决策属性的集合，V_a 是条件属性 a 的值域，$f: U\times A\to V$ 是一个对象与属性值间的映射函数，$f(x,a)=a(x)$ 表示对象 $x\in U$ 在条件属性 $a\in C$ 上的取值，而 $a(x)$ 的值为一个区间，$a(x)=\left[a^-, a^+\right]$，其中 a^- 代表区间的左边界，a^+ 代表区间的右边界。

定义 2：假设区间值决策系统 $IVDS=(U, A=C\cup D, V, f)$ 中的任意两个区间 $I_1=\left[l^-, l^+\right]$ 和 $I_2=\left[e^-, e^+\right]$，则区间 I_1 和区间 I_2 之间的交运算和并运算定义如下：

$$I_1 \cap I_2 = \begin{cases} \left[\max(l^-, e^-),\ \min(l^+, e^+)\right], & \max\left(l^-, e^-\right) \le \min\left(l^+, e^+\right) \\ \varnothing, & otherwise \end{cases} \tag{1}$$

$$I_1 \cup I_2 = \begin{cases} \left[\min\left(l^-, e^-\right),\ \max\left(l^+, e^+\right)\right], & \max\left(l^-, e^-\right) \le \min\left(l^+, e^+\right) \\ \left[l^-, l^+\right] \cup \left[e^-, e^+\right], & otherwise \end{cases} \tag{2}$$

定义 3 [13]：假设区间值决策系统 $IVDS = \left(U,\ A = C \cup D,\ V, f\right)$ 中的两个对象 $\forall x, y \in U$，条件属性 $a_m \in C$，则对象 x 和 y 在条件属性 a_m 下的取值分别为 $x_m = \left[l^-, l^+\right]$，$y_m = \left[e^-, e^+\right]$，则对象 x 相较于对象 y 在条件属性 a_m 下的优势度定义为：

$$G_{x_m \ge y_m} = \min\left(1,\ \max\left(\frac{l^+ - e^-}{\left(l^+ - l^-\right) - \left(e^+ - e^-\right)}, 0\right)\right) \tag{3}$$

优势度具有以下性质：

1) $G_{x_m \ge y_m} \ne G_{y_m \ge x_m}$；

2) $0 \le G_{x_m \ge y_m} \le 1$；

3) $G_{x_m \ge y_m} + G_{y_m \ge x_m} = 1$；

4) $G_{x_m \ge x_m} = G_{y_m \ge y_m} = 0.5$。

定义 4 [13]：假设区间值决策系统 $IVDS = \left(U,\ A = C \cup D,\ V, f\right)$ 中的两个对象 $\forall x, y \in U$，条件属性 $a_m \in C$，则对象 x 和 y 在条件属性 a_m 下的优势度分别为 $G_{x_m \ge y_m}$ 和 $G_{y_m \ge x_m}$，$\forall P \subseteq C\left(\left|P\right| = s\right)$，则对象 x 和对象 y 在条件属性子集 P 下的相似距离定义为：

$$\vartheta\left(x, y\right) = \sqrt{\sum_{m=1}^{s}\left(G_{x_m \ge y_m} - G_{y_m \ge x_m}\right)^2} \tag{4}$$

则相似距离 ϑ 具有以下性质：

1) $\vartheta\left(x, x\right) = 0$；

2) $\vartheta\left(x, y\right) = \vartheta\left(y, x\right)$；

3) $\vartheta\left(x, z\right) \le \vartheta\left(x, y\right) + \vartheta\left(y, z\right)$；

4) $\vartheta\left(x, z\right) \ge \vartheta\left(x, y\right) - \vartheta\left(y, z\right)$。

定义 5：给定一个区间值决策系统 $IVDS = \left(U,\ A = C \cup D,\ V, f\right)$，相似距离阈值 $0 \le \varepsilon \le 1$，对于非空条件属性子集 $\forall P \subseteq C$，则定义在条件属性子集 P 下 ε-相容关系 ζ_P^ε 为：

$$\zeta_P^\varepsilon = \left\{\left(x_i, x_j\right) \middle| \left(x_i, x_j\right) \in U^2,\ \forall a_k \in P \wedge \vartheta\left(x_i, x_j\right) \le \varepsilon\right\} \tag{5}$$

定义 6：给定一个区间值决策系统 $IVDS = \left(U,\ A = C \cup D,\ V, f\right)$，对于非空条件属性子集 $\forall P \subseteq C$，对于 $\forall x_i \in U$，则对象 x_i 在条件属性子集 P 下 ε-相容类 $\mathrm{T}_P^\varepsilon\left(x_i\right)$ 定义为：

$$\mathrm{T}_P^\varepsilon\left(x_i\right) = \left\{x_j \middle| \forall x_j \in U,\ \left(x_i, x_j\right) \in \zeta_P^\varepsilon\right\} \tag{6}$$

而在条件属性子集 P 下的相容类集合为：

$$\mathrm{T}_P^\varepsilon\left(U\right) = \left\{\mathrm{T}_P^\varepsilon\left(x_1\right), \mathrm{T}_P^\varepsilon\left(x_2\right), \cdots, \mathrm{T}_P^\varepsilon\left(x_{|U|}\right)\right\} \tag{7}$$

定义 7：给定一个区间值决策系统 $IVDS = \left(U,\ A = C \cup D,\ V, f\right)$，$\forall P \subseteq C$，对于 $\forall X \subseteq U$，则目标集合 X 关于条件属性子集 P 的下、上近似定义为：

$$\underline{RS_P}(X)=\left\{x_i \middle| x_i \in U,\ \mathrm{T}_P^{\varepsilon}(x_i)\subseteq X\right\} \tag{8}$$

$$\overline{RS_P}(X)=\left\{x_i \middle| x_i \in U,\ \mathrm{T}_P^{\varepsilon}(x_i)\cap X\neq\varnothing\right\} \tag{9}$$

由此可推出正域为 $POS_P(X)=\underline{RS_P}(X)$，边界域为 $BND_P(X)=\overline{RS_P}(X)-\underline{RS_P}(X)$，负域为 $NEG_P(X)=U-\overline{RS_P}(X)$。

根据上述定义，给出了基于区间值决策系统的下近似经典计算算法，算法详情如表 1 所示。

Table 1. Classical algorithm for computing positive region in interval-valued decision systems (CCMLA)
表 1. 区间值决策系统的正域经典求解算法

输入：一个区间值决策系统 $IVDS=(U,\ AT=C\cup D,\ V,\ f)$，相似距离阈值 ε

输出：$IVDS$ 的正域集合 $POS_C(D)$

1) 初始化：$POS_C(D)=\varnothing$

2) 对于 $\forall x_i\in U$，根据定义 5 计算对象 x_i 的相容类 $\mathrm{T}_P^{\varepsilon}(x_i)$

3) 若 $\mathrm{T}_P^{\varepsilon}(x_i)$ 完全包含在决策类中，$POS_C(D)=POS_C(D)\cup x_i$

4) 输出正域集合 $POS_C(D)$

假设区间值决策系统 $IVDS$ 中包含 n 个对象，m 个条件属性，在计算相容类时，需要遍历数据集中的每个对象，并依据定义计算相容关系，在全部属性下计算正域的时间复杂度为 $O(m\cdot n^2)$。

3. 区间值决策系统的正域快速求解算法

定义 8 [14]：给定一个区间值决策系统 $IVDS=(U,\ A=C\cup D,\ V,\ f)$，相似距离阈值 $0\le\varepsilon\le1$，对于 $\forall b\subseteq C$，通过哈希函数可以将论域 U 中的所有对象在条件属性 b 下映射到有限个区域 $Q_1,Q_2,\cdots Q_t$ 中，其中区域 Q_t 中的对象集合定义为：

$$Q_t=\left\{x_i \middle| \forall x_i\in U,\ \left\lceil \vartheta_b(x_0,x_i)/\varepsilon\right\rceil=t\right\} \tag{10}$$

其中 x_0 表示在每个条件属性下的最小区间值，选取方法为：先确定下边界最小的区间，若下边界相同，则选择上边界最小的区间。$\vartheta_b(x_0,x_i)$ 表示为在条件属性 b 下两者之间的相似距离，Q_t 集合中的所有对象表示都和当前条件属性下最小区间值的相似距离在 $((t-1)\varepsilon,\ t\varepsilon)$ 之间。

定理 1：给定一个区间值决策系统 $IVDS=(U,\ A=C\cup D,\ V,\ f)$，相似距离阈值 $0\le\varepsilon\le1$，对于 $\forall b\subseteq C$，根据上述定义所示，通过哈希函数可以将论域 U 中的所有对象在条件属性 b 下映射到有限个区域 $Q_1,Q_2,\cdots Q_t$ 中，则对于 $\forall x_i\in Q_g\ (g=2,3\cdots t-1)$，则对象 x_i 的相容类对象仅包含在区域 $Q_{g-1},Q_g,\cdots Q_{g+1}$ 中；如果 $\forall x_i\in Q_1$，则对象 x_i 的相容类对象仅包含在区域 Q_1，Q_2 中；如果 $\forall x_i\in Q_t$，则对象 x_i 的相容类对象仅包含在区域 Q_t，Q_{t-1} 中。

证明：图 1 为有 6 个区域的哈希映射图 $Q_1,Q_2,\cdots Q_6$，假设存在两个对象 $x_1,\ x_2\in U$，根据上述定义在条件属性 b 下，对象 x_1 映射到了区域 Q_2，对象 x_2 映射到了区域 Q_4，即 $x_1\in Q_2$，$x_2\in Q_4$，则有 $\varepsilon<\vartheta_b(x_0,x_1)\le2\varepsilon$，$3\varepsilon<\vartheta_b(x_0,x_2)\le4\varepsilon$，此时 $\vartheta_b(x_0,x_2)-\vartheta_b(x_0,x_1)>\varepsilon$，并且由相似距离函数的性质可知 $\vartheta_b(x_1,x_2)<\vartheta_b(x_0,x_1)+\vartheta_b(x_0,x_2)$，并且 $\vartheta_b(x_1,x_2)>\vartheta_b(x_0,x_2)-\vartheta_b(x_0,x_1)$，所以有 $\vartheta_b(x_1,x_2)>\varepsilon$，所以对象 x_1 和对象 x_2 在条件属性 b 下不构成相容关系，即不在同一个相容类中。对于 $\forall x_i,\ x_j\in U$，$\forall Q_t\in Q$，上述均可被同理证明，证明完毕。

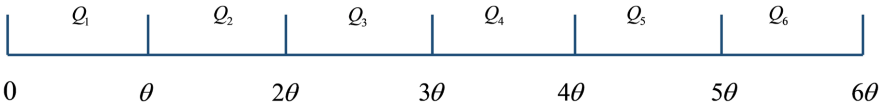


Figure 1. Hash mapping graph
图 1. 哈希映射图

根据上述定义可以得出基于区间值决策系统的下近似快速计算算法，算法详情如表 2 所示。

Table 2. Fast algorithm for computing positive region in interval-valued decision systems (FCMLA)
表 2. 区间值决策系统的正域快速求解算法(FCMLA)

输入： 一个区间值决策系统 $IVDS=\left(U,\ AT=C\cup D,\ V,\ f\right)$，相似距离阈值 ε

输出： $IVDS$ 的正域集合 $POS_C\left(D\right)$

1) 初始化：$POS_C\left(D\right)=\varnothing$

2) 对于 $\forall a_m\in C$，求出在条件属性 a_m 下的最小值

3) 对于 $\forall x_i\in U$，根据定义 9 计算对象 x_i 的映射区域 Q_t

4) 对于 $\forall x_j\in U$，遍历对象 x_j 所在区域以及相邻区域求得 $\mathrm{T}_P^{\varepsilon}\left(x_j\right)$

5) 若 $\mathrm{T}_P^{\varepsilon}\left(x_j\right)$ 完全包含在决策类中，$POS_C\left(D\right)=POS_C\left(D\right)\cup x_j$

6) 输出正域集合 $POS_C\left(D\right)$

假设区间值决策系统 $IVDS$ 中包含 n 个对象，m 个条件属性，传统的相容类计算方法需要遍历两次论域 U，因此在寻找所有对象相容类的时间复杂度为 $O\left(m\cdot n^2\right)$，而用哈希函数映射的方法，只需要遍历当前对象的区域以及邻近区域中的对象即可，不用再去遍历所有的对象，具体的时间复杂度由对象所在区域和邻近区域内的数量决定，时间复杂度明显小于 $O\left(m\cdot n^2\right)$。当所有的对象经过哈希函数的映射都在一个区域内，此时的时间复杂度最大，时间复杂度应为 $O\left(m\cdot n^2\right)$；而当每个对象都被映射到单独一个区域内，那么也就是经过哈希函数的映射，每个区域内只有一个对象时，此时的时间复杂度最小，时间复杂度达到了 $O\left(m\cdot n\right)$。

4. 实验分析

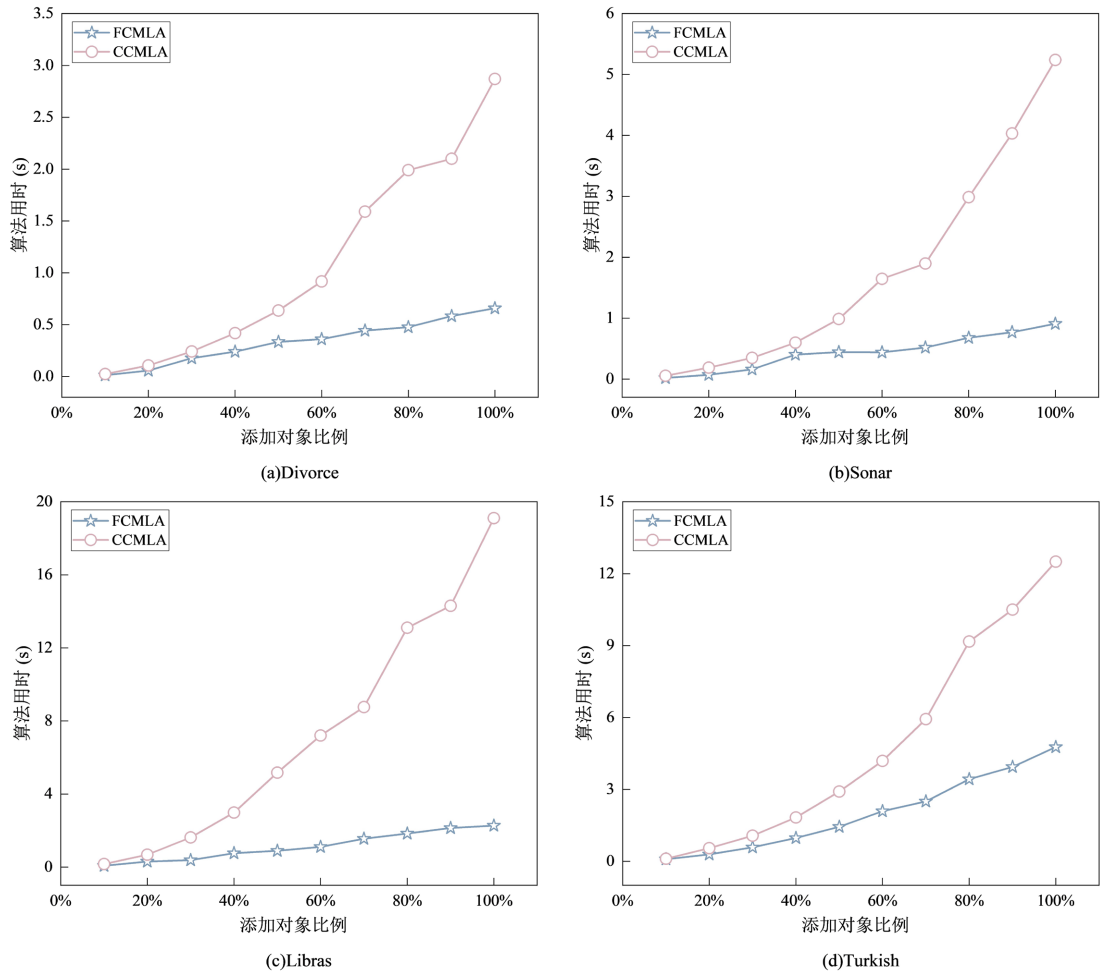
实验选取了 8 组 UCI 数据集，数据集的详细信息如表 3 所示，对于数据集中的符号型数据、缺失数据、名词型数据、连续型数据等不同的数据类型，分别采用数值化处理、插值法替换、{0、1}替换、等频分割等方法进行处理。等数据处理完之后，使用编写的区间值数据生成算法对上述数据集进行处理，使之生成新的区间值数据集。实验主要验证区间值决策系统的正域经典求解算法(CCMLA)与本文章提出的区间值决策系统的正域快速求解算法(FCMLA)的计算效率对比，本实验的实验环境为：Windows10 64 位操作系统；8GB 的内存；Intel(R) Core(TM) i7-8550U CPU；软件环境为：PyCharm；编程环境为：Python，相似距离阈值 $\varepsilon=0.2$。

实验对比了区间值决策系统的正域经典求解算法(CCMLA)同区间值决策系统的正域快速求解算法(FCMLA)计算下近似的时间消耗情况。将数据集的对象平均分成 10 份，每份占原数据集大小的 10%，数据集对象的初始数量从 10%开始，每次添加 10%的对象数量，直至添加至原数据集大小。图 2 表示随着对象数量的变化，两个算法下近似计算时间的变化情况，蓝色五角星折线是本文提出的算法 FCMLA，红色球形折线是经典算法 CCMLA，横坐标为添加对象比例，纵坐标是算法运行时间，单位为秒。

Table 3. UCI datasets
表 3. UCI 数据集

序号	数据集	样本数	属性数	类别数
1	Divorce	170	54	2
2	Sonar	208	60	2
3	Libras	360	90	15
4	Turkish	400	50	4
5	Musk	476	168	2
6	Breast Cancer	568	30	2
7	Sports	1000	59	2
8	Statlog	4435	36	6

从图 2 中可以看出，本文所提的算法 FCMLA 在 8 个数据集下都优于经典算法 CCMLA，而且随着对象数量的增加，两种算法所消耗的时间都有所上升，但 FCMLA 算法时间消耗起伏较小，而 CCMLA 算法的时间消耗起伏较大。在大数据集上更明显，比如 Statlog 数据集，随着对象数量的增加，CCMLA 算法时间消耗明显上升，而 FCMLA 算法时间消耗起伏很小，几乎为一条直线。



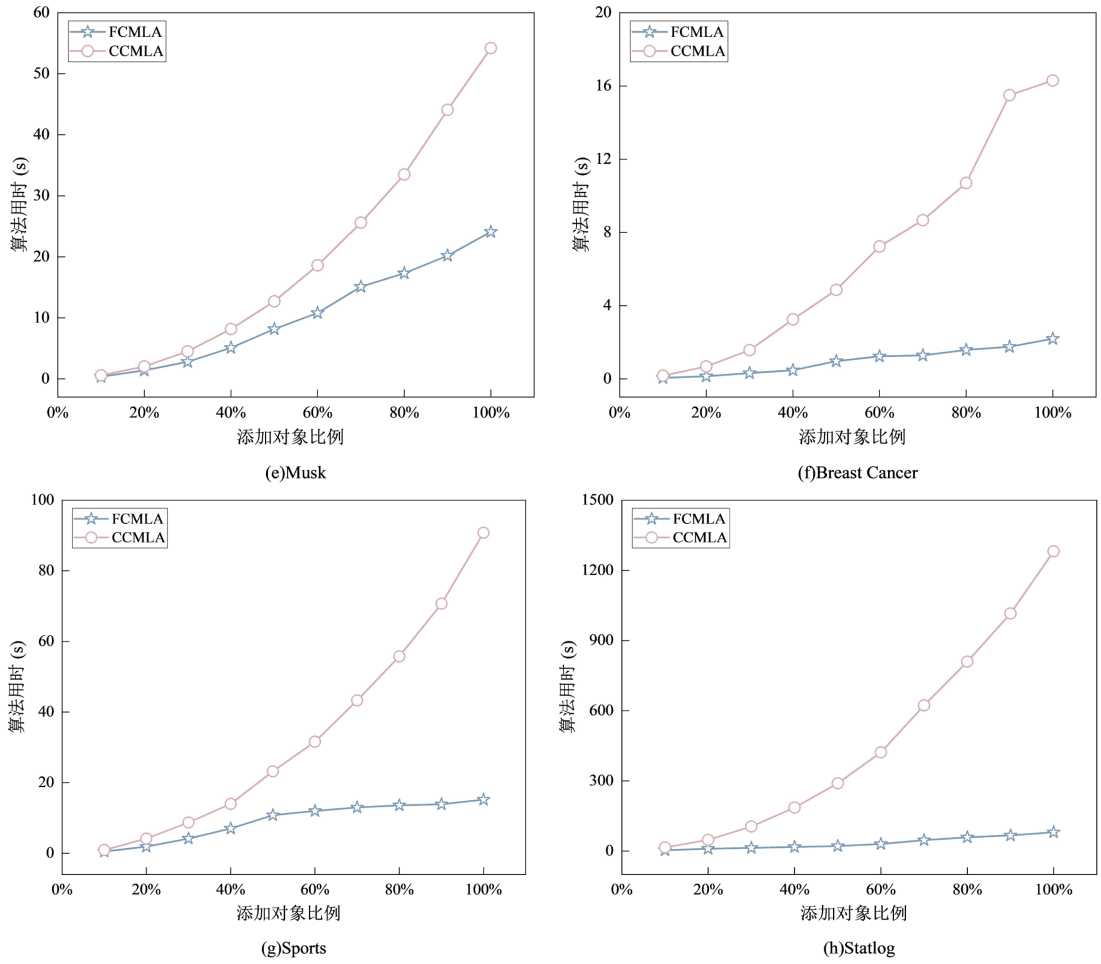


Figure 2. Time efficiency graph
图 2. 时间效率图

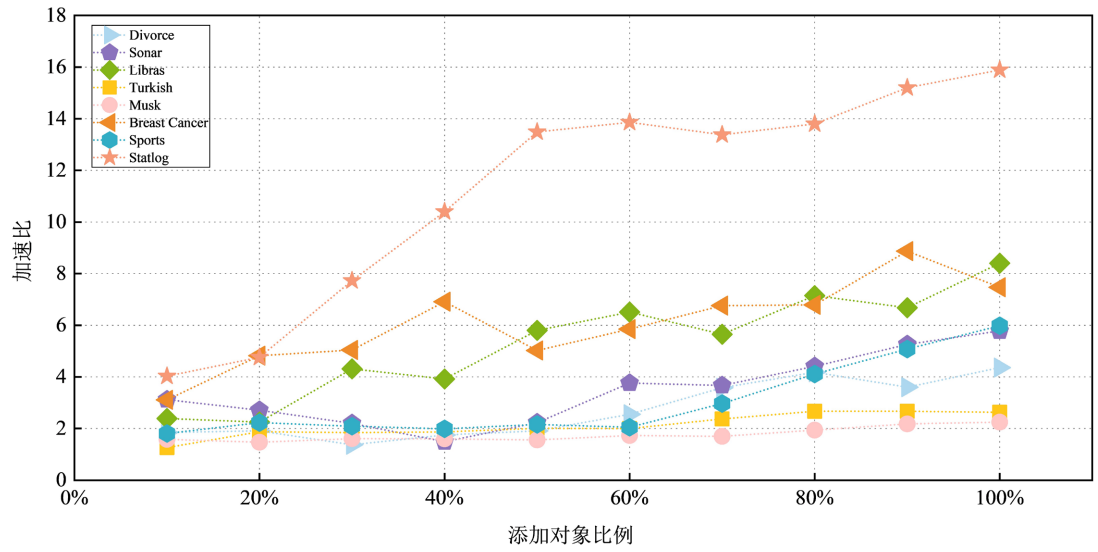


Figure 3. Time speedup ratio
图 3. 时间加速比

图 3 为 CCMLA 算法与 FCMLA 算法所用时间的比值，来进一步证实了 FCMLA 算法的优越性，可以观察到 FCMLA 算法在 Sonar 数据集上的加速比是 CCMLA 算法的 3.12 倍到 5.78 倍；在 Statlog 数据集上的加速比是 CCMLA 算法的 4.03 倍到 15.89 倍。

5. 结论

在本文中，针对传统正域计算算法在处理大规模区间值决策系统时计算复杂度较高的问题，提出了一种采用哈希思想的快速求解算法。该算法利用哈希函数对每个对象进行快速分区，并结合相容类的单调性优化计算过程，有效减少冗余计算，提高计算效率。实验结果表明，在 8 个 UCI 数据集上，本文所提算法在计算速度上优于传统算法，且在大规模数据集上的性能更稳定，具有较高的应用价值。

基金项目

本文受烟台市科技计划项目(编号：2022XDRH016)的资助。

参考文献

[1] Pawlak, Z. (1982) Rough Sets. *International Journal of Computer & Information Sciences*, **11**, 341-356. https://doi.org/10.1007/bf01001956

[2] Dai, J., Wang, W., Xu, Q. and Tian, H. (2012) Uncertainty Measurement for Interval-Valued Decision Systems Based on Extended Conditional Entropy. *Knowledge-Based Systems*, **27**, 443-450. https://doi.org/10.1016/j.knosys.2011.10.013

[3] Chen, B., Zhang, X. and Yuan, Z. (2024) Two-Dimensional Improved Attribute Reductions Based on Distance Granulation and Condition Entropy in Incomplete Interval-Valued Decision Systems. *Information Sciences*, **657**, Article 119910. https://doi.org/10.1016/j.ins.2023.119910

[4] 张楠, 苗夺谦, 岳晓冬. 区间值信息系统的知识约简[J]. 计算机研究与发展, 2010, 47(8): 1362-1371.

[5] Yao, Y. and Zhao, Y. (2008) Attribute Reduction in Decision-Theoretic Rough Set Models. *Information Sciences*, **178**, 3356-3373. https://doi.org/10.1016/j.ins.2008.05.010

[6] Wang, C., Huang, Y., Ding, W. and Cao, Z. (2021) Attribute Reduction with Fuzzy Rough Self-Information Measures. *Information Sciences*, **549**, 68-86. https://doi.org/10.1016/j.ins.2020.11.021

[7] Li, Z., Zhang, Q., Liu, S., Peng, Y. and Li, L. (2024) Information Fusion and Attribute Reduction for Multi-Source Incomplete Mixed Data via Conditional Information Entropy and D-S Evidence Theory. *Applied Soft Computing*, **151**, Article 111149. https://doi.org/10.1016/j.asoc.2023.111149

[8] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(6): 42-45.

[9] Yao, Y. and Zhao, Y. (2009) Discernibility Matrix Simplification for Constructing Attribute Reducts. *Information Sciences*, **179**, 867-882. https://doi.org/10.1016/j.ins.2008.11.020

[10] Qian, W., Wan, L. and Shu, W. (2024) Semi-Supervised Feature Selection Based on Discernibility Matrix and Mutual Information. *Applied Intelligence*, **54**, 7278-7295. https://doi.org/10.1007/s10489-024-05481-3

[11] 杨明. 一种基于改进差别矩阵的核增量式更新算法[J]. 计算机学报, 2006, 29(3): 407-413.

[12] Sowkuntla, P. and Sai Prasad, P.S.V.S. (2025) Parallel Attribute Reduction in High-Dimensional Data: An Efficient Mapreduce Strategy with Fuzzy Discernibility Matrix. *Applied Soft Computing*, **172**, Article 112870. https://doi.org/10.1016/j.asoc.2025.112870

[13] 徐伟华. 序信息系统与粗糙集[M]. 北京: 科学出版社, 2013.

[14] Yong, L., Wenliang, H., Yunliang, J. and Zhiyong, Z. (2014) Quick Attribute Reduct Algorithm for Neighborhood Rough Set Model. *Information Sciences*, **271**, 65-81. https://doi.org/10.1016/j.ins.2014.02.093