

基于剥离商集的快速广义决策保持属性约简算法

刘中凯, 张楠

(烟台大学计算机与控制工程学院, 烟台 264005)

摘要: 属性约简作为粗糙集理论中的主要研究问题之一, 其目标是在不影响原始数据分类能力的前提下, 寻找一个更为简洁的属性子集。在约简策略中, 基于广义决策保持的属性约简因其具有良好的可解释性与较强的应用背景, 得到广泛应用。现有广义决策保持属性约简算法的计算效率在面对大规模数据集时常受到显著制约, 如何高效地实现广义决策保持下的属性约简, 成为当前研究亟需解决的关键问题。针对上述问题, 本文探讨了广义决策保持过程中正域对象的稳定性特征, 借助剥离商集方法, 构建了简化且更具代表性的等价类, 进而提出了一种基于广义决策保持的高效启发式属性约简算法。该算法在确保约简正确性的基础上, 可以显著提升约简过程的计算效率。实验结果表明, 相较于现有算法, 本文提出的算法在8个UCI数据集上具有更高的计算效率。

关键词: 粗糙集; 属性约简; 启发式算法; 剥离商集

中图分类号: TP18 **文献标志码:** A

引用格式: 刘中凯, 张楠. 基于剥离商集的快速广义决策保持属性约简算法[J]. 数据采集与处理, 2026, 41(4): 1118–1132. LIU Zhongkai, ZHANG Nan. Efficient attribute reduction algorithm for generalized decision preservation based on stripped quotient set[J]. Journal of Data Acquisition and Processing, 2026, 41(4): 1118–1132.

引言

1982年粗糙集理论在文献[1]中被首次提出, 它主要处理现实生活中具有不确定性、残缺性等特点的数据。粗糙集作为一种数据分析的数学理论与模型, 可以深入分析数据, 挖掘其中的潜在信息, 并揭示属性与决策之间的内在联系。由此, 粗糙集在诸多热门领域得到应用, 如知识表示、特征选择和机器学习等^[2-5]。

属性约简^[6-11]是粗糙集理论中最重要的研究和应用之一, 旨在不降低决策能力的前提下, 从信息中去除冗余属性, 同时保持某种分类能力不变, 保留最具代表性的信息。随着数据规模的不断扩大, 经典的属性约简算法在效率上面临严峻挑战, 而在实际应用中, 算法的高效性非常重要, 诸多学者在如何设计高效的属性约简算法中做出了重大贡献^[12-19], 这些改进为约简算法在其他领域中的应用提供重要可拓性。广义决策保持约简^[20]由于能保持系统原有的决策能力, 成为研究热点之一。文献[20]提出了广义决策保持的差别矩阵, 能计算出广义决策保持的所有约简; 文献[21]提出了一种基于广义决策保持的启发式算法, 提高了约简效率; 文献[22]采用哈希表存储等价类和广义决策, 解决了计算广义决策相似度需多次遍历全部决策类的问题, 提升了算法效率。

剥离商集也称剥离集, 由Nguyen等^[12]学者提出。该方法目标是从属性划分中分离出对正域产生决定性影响的局部结构, 减少数据计算量, 以此提升属性约简过程中算法实行效率。文献[13]将剥离

商集方法应用到经典粗糙集中,给出了一种新的显著性度量方法,利用度量设计了有效的约简算法,提高了约简效率。文献[14]将剥离商集方法应用到信息熵保持约简中,定义了基于剥离商集的属性约简概念,并提出了一种新的算法。该算法分类精度优于香农条件熵和梁氏条件熵的约简精度,并且适用于不一致决策表。文献[15]将剥离商集方法应用到邻域粗糙集中,并引入了剥离邻域覆盖的新概念,减少原覆盖不必要的容差类,在混合和不完全决策表中定义了一种新的高效启发式约简算法,并验证了该算法在分类精度上优于同类算法。本文将剥离商集^[12]方法应用到广义决策保持的属性约简算法中,结合哈希表存储^[22],给出一种新的度量方式,实现基于剥离商集的快速广义决策保持属性约简。

1 基础知识

为进一步研究基于剥离商集的广义决策保持快速属性约简算法,本节给出粗糙集理论和剥离商集的相关概念。

1.1 粗糙集理论和剥离商集

本节主要介绍了粗糙集理论和剥离商集的相关概念,包括决策系统、不可分辨关系、下近似、正域和剥离商集等相关概念。

定义 1^[1] 四元组 $IS=(U, AT, V, f)$ 为一个信息系统,其中 U 包含信息系统中的所有对象,这个非空有限集合称为论域, $|U|$ 为论域中对象的个数; AT 为属性的集合; $V=\bigcup\{V_s|s\in AT\}$, 其中 V_s 是属性 $s\in AT$ 的值域; $f:U\times AT\rightarrow V$ 为映射函数, $f(x_i, s)$ 为对象 $x_i\in U (1\leq i\leq |U|)$ 在属性 $s\in AT$ 上的取值,记为 $s(x_i)$ 。

如果 IS 中的 $AT=C\cup D$, 其中 $C=\{a_1, a_2, \dots, a_{|C|}\}$ 和 $D=\{d\}$ 分别表示条件属性集合和决策属性的集合,则记 $DS=(U, AT=C\cup D, V, f)$ 为一个决策系统, $V=V_C\cup V_D, f(x_i, d)=d(x_i)$ 表示对象 $x_i\in U (1\leq i\leq |U|)$ 的决策值。

定义 2^[1] 四元组 $DS=(U, AT=C\cup D, V, f)$ 为一个决策系统,对于 $\forall B\subseteq C, x, y\in U\times U$, 定义在属性 B 下的不可分辨关系 $IND(B)$ 为

$$IND(B)=\{(x, y)|\forall b\in B, b(x)=b(y)\} \quad (1)$$

式中不可分辨关系 $IND(B)$ 为一个满足自反性、对称性和传递性的等价关系,由等价关系可计算属性 B 对论域 U 的划分为 $U/IND(B)=\{[x_i]_B|x_i\in U\}=\{E_B(x_i)|x_i\in U\}$, 简记 U/B , 其中 $[x_i]_B=E_B(x_i)$ 表示包含对象 x_i 的等价类,通过等价关系对论域 U 划分所得的等价类,可以得出下近似的相关定义。

定义 3^[1] 四元组 $DS=(U, AT=C\cup D, V, f)$ 为一个决策系统,对于 $\forall B\subseteq C, X\subseteq U$, 定义 X 关于 B 的下近似为

$$\underline{B}(X)=\{x_i\in U|[x_i]_B\subseteq X\} \quad (2)$$

在下近似相关的定义基础上,可进一步引出正域的定义。具体而言,正域由所有能够在下近似中被完全正确分类的对象组成,它反映了通过条件属性集合对决策属性的确定能力。

定义 4^[1] 四元组 $DS=(U, AT=C\cup D, V, f)$ 为一个决策系统,对于 $\forall B\subseteq C, X\subseteq U$, 定义 X 关于 B 的正域为

$$POS_B(X)=\underline{B}(X) \quad (3)$$

为更好地理解其在约简过程中的作用,下面对剥离商集的基本概念及其构造方法进行系统阐述。

定义 5^[12] 四元组 $DS=(U, AT=C\cup D, V, f)$ 为一个决策系统,对于 $\forall B\subseteq C, x_i\in U$, 定义集合 U/B 的剥离商集为

$$\widehat{U/B}=\{[x_i]_B\in U/B \parallel [x_i]_B|>1\} \quad (4)$$

基于剥离商集的概念, Nguyen等^[12]给出了 D -剥离商集的概念, 用于优化属性约简中的等价类表示, 下面给出 D -剥离商集的相关定义。

定义6^[12] 给出两个剥离商集 $\widehat{U/B}$ 和 $\widehat{U/D}$, 对于 $\forall [x_i]_B \in \widehat{U/B}, [x_i]_D \in \widehat{U/D}$ 定义 U/B 下的 D -剥离商集 $\widehat{U/B^D}$ 可表示为

$$\widehat{U/B^D} = \{[x_i]_B | [x_i]_B \not\subseteq [x_i]_D\} \quad (5)$$

如果两个 D -剥离商集为已知条件, 则两个 D -剥离商集的点乘可由以下定义进行计算。

定义7^[12] 给出两个剥离商集 $\widehat{U/B_1^D}$ 和 $\widehat{U/B_2^D}$, $\forall [x_i]_D \in \widehat{U/D}$, 则两个剥离商集点乘可表示为

$$(\widehat{U/B_1^D}) \cdot (\widehat{U/B_2^D}) = \bigcup_{[x_i]_{B_1} \in \widehat{U/B_1^D}} Q([x_i]_{B_1}) \quad (6)$$

式中: $Q([x_i]_{B_1}) = \bigcup_{[x_i]_{B_2} \in \widehat{U/B_2^D}} \{[x_i]_{B_1} \cap [x_i]_{B_2} | [x_i]_{B_1} \cap [x_i]_{B_2} \not\subseteq [x_i]_D\}$, 且 $(\widehat{U/B_1^D}) \cdot (\widehat{U/B_2^D})$ 可记作 $\widehat{U/(B_1 \cup B_2)^D}$ 。

由于 D -剥离商集不包含任何属于正域的对象, 因此该集合与正域之间构成了一个互补的关系, 将该集合记为 $\langle \widehat{U/B^D} \rangle$, 进一步给出相关概念形式化定义。

定义8^[12] 四元组 $DS = (U, AT = C \cup D, V, f)$ 为一个决策系统, 对于 $\forall B \subseteq C$, 定义 $\langle \widehat{U/B^D} \rangle$ 为

$$\langle \widehat{U/B^D} \rangle = U \setminus POS_B(D) \quad (7)$$

式中 $U \setminus POS_B(D)$ 表示从论域 U 中剔除正域对象后的集合。

2 广义决策保持属性约简的度量与算法

本节共分为3个小节: 第1小节介绍广义决策保持约简的基本概念, 包括广义决策、相似度及属性重要度; 第2小节提出一种新的约简判断指标, 给出其数学性质及相应的算法描述; 第3小节在此基础上详细描述了基于该指标的启发式属性约简算法流程, 为后续算法验证提供理论与方法支持。

2.1 广义决策保持约简

结合第1节中的基础知识, 本节主要介绍了广义决策保持的相关概念, 包括广义决策、相似度和属性重要度计算方式等相关概念。通过对这些概念的系统阐述, 为后续算法的设计与实现提供了理论和数学工具支持。

定义9^[21] 四元组 $DS = (U, AT = C \cup D, V, f)$ 为一个决策系统, 对于 $\forall B \subseteq C, x_i, x_j \in U, [x_i]_B \in \widehat{U/B}$, 定义对象 x_i 在属性子集 B 下的广义决策为

$$\delta_B(x_i) = \{f(x_j, d) | x_j \in [x_i]_B\} \quad (8)$$

通过广义决策的定义, 可以计算部分条件属性与所有条件属性之间的相似度, 具体计算方式由定义10给出。

定义10^[21] 四元组 $DS = (U, AT = C \cup D, V, f)$ 为一个决策系统, 对于 $\forall P \subseteq C, x_i \in U$, 定义广义决策保持约简的相似度 $\text{Sim}(P, C)$ 为

$$\text{Sim}(P, C) = \begin{cases} \frac{|\bigcup_{i=1}^{|U|} ([x_i]_C)|}{|U|} & \delta_P(x_i) = \delta_C(x_i) \\ 0 & \text{其他} \end{cases} \quad (9)$$

根据上述相似度计算公式, 可以对各个相关的条件属性进行重要度评估。通过计算每个属性在整体数据集中的区分能力, 能够有效地衡量其在决策过程中的贡献程度。

定义 11^[21] 四元组 $DS=(U, AT=C \cup D, V, f)$ 为一个决策系统, 设 $\forall B \subseteq C, b \in B, P = B - \{b\}$, 定义广义决策保持约简的内部属性重要度 $\text{Sig}^{\text{inner}}(b, B, C)$ 为

$$\text{Sig}^{\text{inner}}(b, B, C) = \text{Sim}(B, C) - \text{Sim}(P, C) \quad (10)$$

通过计算内部属性的重要度, 可以确定决策系统中的核属性, 这些核属性对于维持决策系统的分类能力具有不可或缺的作用。在此基础上, 进一步计算外部属性的重要度, 可以获得该决策系统的一个广义决策保持属性约简。

定义 12^[21] 四元组 $DS=(U, AT=C \cup D, V, f)$ 为一个决策系统, 设 $\forall B \subseteq C, q \in C - B, Q = B \cup \{q\}$, 定义广义决策保持约简的外部属性重要度 $\text{Sig}^{\text{outer}}(q, B, C)$ 为

$$\text{Sig}^{\text{outer}}(q, B, C) = \text{Sim}(Q, C) - \text{Sim}(B, C) \quad (11)$$

2.2 约简度量指标差别度

本节提出了一种新的广义决策保持属性约简度量指标, 并初步讨论了其相关性质。通过对该指标基本特性的分析, 尝试为实际约简过程提供可能的理论参考。这些工作或可为后续约简算法的设计与改进提供相关的基础, 有望对提升属性约简方法的理论完整性与应用价值有所帮助。

定义 13 四元组 $DS=(U, AT=C \cup D, V, f)$ 为一个决策系统, 设 $\forall P \subseteq C, x_i \in U$, 其中 $|U|$ 表示论域中所有对象的个数, 定义广义决策保持的差别度 $\text{Dis}(P, C, U)$ 为

$$\text{Dis}(P, C, U) = \frac{|\text{Dif}_P(C)|}{|U|} \quad (12)$$

式中 $\text{Dif}_P(C) = \{x_i | \delta_P(x_i) \neq \delta_C(x_i)\}$ 是由条件属性 P 与条件属性 C 下广义决策不相等的对象构成的集合。这一指标在属性约简和决策规则提取中, 能够帮助识别属性之间的依赖关系和冗余, 并为属性约简算法提供启发式搜索度量指标。

定理 1 四元组 $DS=(U, AT=C \cup D, V, f)$ 为一个决策系统, 对于任意属性子集 $P \subseteq Q \subseteq C$, 广义决策保持差别度满足单调性, 有

$$\text{Dis}(P, C, U) \geq \text{Dis}(Q, C, U) \quad (13)$$

证明: 对于任意属性子集 $P \subseteq Q \subseteq C$, 由划分的单调性可知 U/Q 的划分比 U/P 更细, 即对于任意对象 $x \in U$ 都有其等价类满足包含关系 $[x]_Q \subseteq [x]_P$, 根据广义决策的定义 $\delta_B(x_i) = \{f(x_j, d) | x_j \in [x_i]_B\}$, 可以推导出 3 层包含关系: $\delta_C(x_i) \subseteq \delta_Q(x_i) \subseteq \delta_P(x_i)$, 当存在对象 $x \in \text{Dif}_Q(C)$ 即满足 $\delta_Q(x_i) \neq \delta_C(x_i)$ 时, 结合上述关系可推出 $\delta_P(x_i) \neq \delta_C(x_i)$, 从而建立集合包含关系 $\text{Dif}_Q(C) \subseteq \text{Dif}_P(C)$ 。通过基数运算转化为 $|\text{Dif}_Q(C)| \leq |\text{Dif}_P(C)|$, 两边同时除以论域大小 $|U|$, 即 $|\text{Dif}_Q(C)|/|U| \leq |\text{Dif}_P(C)|/|U|$, 可得 $\text{Dis}(Q, C, U) \leq \text{Dis}(P, C, U)$; 特别地, 当 Q 比 P 多包含至少一个对某些对象的广义决策产生影响的属性时, 即存在 $x \in U$ 使得关系 $\delta_P(x_i) \neq \delta_C(x_i)$ 但 $\delta_Q(x_i) = \delta_C(x_i)$ 成立, 此时不等式有关系 $\text{Dis}(Q, C, U) < \text{Dis}(P, C, U)$ 严格成立, 证毕。

例 1 在表 1 所示的决策系统中, 记 $P = \{C_1\}, Q = \{C_1, C_2\}, \forall x_i \in U$, 决策属性 D 对 U 划分的等价类为 $U/D = \{\{x_1, x_2, x_3\}, \{x_4\}, \{x_5\}\}$, 在所有条件属性 C 下划分的包含对象 x_i 的等价类为 $[x_1]_C = [x_2]_C = \{x_1, x_2\}, [x_3]_C = \{x_3\}, [x_4]_C = \{x_4\}, [x_5]_C = \{x_5\}$ 。条件属性 C 下对象的广义决策为 $\delta_C(x_1) = \delta_C(x_2) = \delta_C(x_3) = \{0\}, \delta_C(x_4) = \{1\}, \delta_C(x_5) = \{2\}$, 在条件属性 P 下划分的包含对象 x_i 的等价类为 $[x_1]_P = [x_2]_P = [x_3]_P = \{x_1, x_2, x_3\}, [x_4]_P = [x_5]_P = \{x_4, x_5\}$, 条件属性 P 下对象

表 1 决策系统

Table 1 A decision system

U	C_1	C_2	C_3	D
x_1	1	0	3	0
x_2	1	0	3	0
x_3	1	1	3	0
x_4	0	2	3	1
x_5	0	3	3	2

的广义决策为 $\delta_P(x_1)=\delta_P(x_2)=\delta_P(x_3)=\{0\}$, $\delta_P(x_4)=\delta_P(x_5)=\{1,2\}$, 由于 $\delta_P(x_4)\neq\delta_C(x_4)$, $\delta_P(x_5)\neq\delta_C(x_5)$, 将对象 x_4 、 x_5 加入到差别度计算公式中, 由此可得条件属性 P 与条件属性 C 之间的差别度 $\text{Dis}(P, C, U)=0.4$; 同理, 条件属性 Q 与条件属性 C 之间的差别度为 $\text{Dis}(Q, C, U)=0$ 。

由上述讨论可以得出, 对于 $\forall P\subseteq Q\subseteq C$, $1\geq\text{Dis}(P, C, U)\geq\text{Dis}(Q, C, U)\geq 0$ 成立。如果 $P\subseteq C$ 是广义决策保持的约简, 则满足定义 14 给出的关系。

定义 14 四元组 $DS=(U, AT=C\cup D, V, f)$ 为一个决策系统, $P\subseteq C$ 是广义决策保持的约简, 则满足: (1) $\text{Dis}(P, C, U)=0$; (2) $\forall R\subset P, \text{Dis}(R, C, U)>0$ 。

条件(1)确保了广义决策保持约简前后的差别度判断指标是一致的; 条件(2)确保了 P 中属性都是必要的, 它保证了约简结果中没有冗余属性。

2.3 启发式约简算法设计及复杂度分析

本节在剥离商集和差别度相关定义的基础上, 设计出一种基于剥离商集的快速广义决策保持属性约简算法。该算法充分利用剥离商集划分等价类的优势, 结合差别度对属性辨识能力的度量机制, 实现对冗余属性的剔除, 从而在保证决策分类能力的前提下, 有效提高属性约简的计算效率。

从例 1 中可以发现, 计算 $\text{Dis}(Q, C, U)$ 时, 论域中的对象被全部判别了一次, 而在计算 $\text{Dis}(P, C, U)$ 时, $E_P(x_1)=E_P(x_2)=E_P(x_3)=\{x_1, x_2, x_3\}\in\{\{x_1, x_2, x_3\}, \{x_4\}, \{x_5\}\}$, 表明对象 $\{x_1, x_2, x_3\}$ 已经被划分到正域中, 因此计算 $\text{Dis}(Q, C, U)$ 时, $\{x_1, x_2, x_3\}$ 不必重复考虑, 借助剥离商集方法, 可将在条件属性 P 下已经划分到正域内的部分对象提前去除掉, 随着条件属性增加计算时, 数据量将大大减少。

定理 2 四元组 $DS=(U, AT=C\cup D, V, f)$ 为一个决策系统, 设 $\forall P\subseteq C$, 差别度满足关系: $\text{Dis}(P, C, U)=\text{Dis}(P, C, \langle U/P^D \rangle)$ 。

证明: 对于 $\forall x_i\in U$, $\text{Dis}(P, C, U)=\text{Dis}(P, C, \langle U/P^D \rangle)+\text{Dis}(P, C, \text{POS}_P(D))$, 如果 $[x_i]_P\in U/D$, 则 x_i 为正域中的对象, 且 $\delta_P(x_i)=\delta_C(x_i)$, 因此 x_i 不会被 $\text{Dif}_P(C)$ 考虑, 即 $\text{Dis}(P, C, \text{POS}_P(D))=0$, 故 $\text{Dis}(P, C, U)=\text{Dis}(P, C, \langle U/P^D \rangle)$, 得证。

定理 2 表明计算差别度时, 不必考虑论域 U 中的所有对象。基于定理 2 给出一种新的内部属性重要度以及外部属性重要度的计算公式。

定义 15 四元组 $DS=(U, AT=C\cup D, V, f)$ 为一个决策系统, $P\subseteq C, a\in P, B=P-\{a\}$, 定义广义决策保持的内部属性重要度为

$$\text{Sig}^{\text{inner}}(a, P, C, D)=\text{Dis}(B, C, \langle U/B^D \rangle)-\text{Dis}(P, C, \langle U/P^D \rangle) \quad (14)$$

定义 16 四元组 $DS=(U, AT=C\cup D, V, f)$ 为一个决策系统, $P\subseteq C, a\in C-P, Q=P\cup\{a\}$, 定义广义决策保持的外部属性重要度为

$$\text{Sig}^{\text{outer}}(a, P, C, D)=\text{Dis}(P, C, \langle U/P^D \rangle)-\text{Dis}(Q, C, \langle U/Q^D \rangle) \quad (15)$$

基于条件属性的重要度度量公式, 给出决策系统中核属性的判定方法。

定理 3 四元组 $DS=(U, AT=C\cup D, V, f)$ 为一个决策系统, $P\subseteq C$, 若 $\exists p\in P$ 使得 $\text{Sig}^{\text{inner}}(p, P, C, D)>0$, 则 $p\in\text{core}$ 。

证明:

如果 $\text{Sig}^{\text{inner}}(p, P, C, D)>0$, 即 $\text{Dis}(P-\{p\}, C, \langle U/(P-\{p\})^D \rangle)-\text{Dis}(P, C, \langle U/P^D \rangle)>0$, 则属性 p 被称为广义决策保持不可或缺的属性, 称为条件属性 P 的核。

如果 $P\subseteq C$ 是基于剥离商集的广义决策保持的约简, 则满足下面定义 17 给出的关系。

定义 17 四元组 $DS = (U, AT = C \cup D, V, f)$ 为一个决策系统, $P \subseteq C$ 是广义决策保持的约简, 则满足: (1) $\text{Dis}(P, C, \widehat{U/P^D}) = 0$; (2) $\forall R \subset P, \text{Dis}(R, C, \widehat{U/R^D}) > 0$ 。

条件(1)确保了基于剥离商集的快速广义决策保持约简前后的差别度判断指标是一致的; 条件(2)确保了 P 中属性都是必要的, 保证了约简结果中没有冗余属性。

结合以上描述, 下面给出相关算法的设计与时间复杂度分析。

算法 1 快速计算 D -剥离商集的算法 (Efficient stripped quotient computation algorithm, ESQCA)

输入: 决策系统 $DS = \langle U, C \cup D, V, f \rangle$, 条件属性 $B_1, B_2 \subseteq C$

输出: $\widehat{U/B_1^D}$ 与 $\widehat{U/B_2^D}$ 的积

- (1) 初始化等价类集合 $\widehat{U/B_1^D} = \emptyset, \widehat{U/B_2^D} = \emptyset$
- (2) 创建空字典 content, 用于临时存储等价类
- (3) 对于 $\forall u \in U$ 执行:
- (4) 初始化 key 为 1 个空字符串
- (5) 将 U 中第 1 个对象在 B_1 下对应的条件属性值连接到 key 上, 形成联合属性值 key
- (6) 如果 key 不在 content 中:
- (7) content[key] = {u}
- (8) 否则, 将 u 添加到 content[key]
- (9) 将每个 content[key] 添加到 $E_{B_1}(u)$ 中
- (10) 通过定义 5 得到 $\widehat{U/B_1}$
- (11) 通过定义 6 得到 $\widehat{U/B_1^D}$
- (12) 同理, 对 B_2 重复步骤(2~11)得到 $\widehat{U/B_2^D}$
- (13) 通过定义 7 得到 $\widehat{U/(B_1 \cup B_2)^D}$
- (14) 返回 $\widehat{U/(B_1 \cup B_2)^D}$

算法 1 为基于 D -剥离商集算法^[13]的改进算法, 在计算等价类时采用哈希存储^[22], 随后删除等价类中对象数目小于等于 1 的集合, 最后将不符合 D -剥离商集判断的集合删除。算法 1 中, 步骤(2~9)为采用哈希存储计算等价类, 时间复杂度为 $O(|C||U|)$, 步骤(10)为判断等价类长度是否大于 1, 时间复杂度为 $O(|U|)$, 步骤(11)为保留非决策类子集的等价类, 时间复杂度为 $O(|U|)$, 步骤(12)时间复杂度为 $O(|C||U|)$, 步骤(13)为计算两个 D -剥离商集点乘, 采用文献[13]方法, 时间复杂度为 $O(|U|)$ 。综上所述, 算法 1 时间复杂度为 $O(|C||U|)$ 。

通过引入算法 1, 本文提出了一种基于剥离商集的快速广义决策保持属性约简算法, 具体过程如算法 2 所示。

算法 2 基于剥离商集的快速广义决策保持属性约简算法 (Fast generalized decision-preserving attribute reduction algorithm based on stripped quotient sets, BGDSP)

输入: 决策系统 $DS = \langle U, C \cup D, V, f \rangle$

输出: 属性约简 red

- (1) 初始化 $\text{red} = \emptyset, R = C$
- (2) 对于 $\forall x_i \in U$, 计算 $\delta_c(x_i)$

- (3) 对于 C 中每个属性 a_i :
- (4) 如果 $\text{Sig}^{\text{inner}}(a_i, R, C, D) > 0$:
- (5) 将属性 a_i 加入 red ;
- (6) 如果属性约简 red 非空:
- (7) 使用算法 1 步骤(2~11), 计算当前属性约简 red 的 D -剥离商集 $\widehat{U/\text{red}^D}$
- (8) 计算当前约简差别度 $\text{Dis}(\text{red}, C, \langle \widehat{U/\text{red}^D} \rangle)$
- (9) 如果 $\text{Dis}(\text{red}, C, \langle \widehat{U/\text{red}^D} \rangle) \neq 0$:
- (10) 对于 $\forall a_i \in C \setminus \text{red}$:
- (11) 使用算法 1, 计算 $\widehat{U/(R \cup \{a_i\})^D}$
- (12) 计算 $\text{Sig}^{\text{outer}}(a_i, R, C, \langle \widehat{U/\text{red}^D} \rangle)$, 记录最大判别指标对应的条件属性 a_0
- (13) 将属性 a_0 加入 red ;
- (14) 更新 red , 重新计算 $\text{Dis}(\text{red}, C, \langle \widehat{U/\text{red}^D} \rangle)$
- (15) 返回属性约简 red

算法 2 中, 步骤(2~5)为寻找核属性, 其中步骤(2)借助哈希存储, 可快速计算出指定对象的广义决策, 时间复杂度为 $O(|U|)^{[22]}$, 因此步骤(2~5)时间复杂度为 $O(|C||U|)$, 步骤(7)为计算当前约简的 D -剥离商集, 时间复杂度为 $O(|C||U|)$, 步骤(9~14)为计算外部属性重要度, 并将最大度量指标对应的条件属性加到约简中, 因此步骤(9~14)时间复杂度为 $O(|C|^2|U|)$ 。综上所述, 算法 2 时间复杂度为 $O(|C|^2|U|)$, 优于经典广义决策保持前向约简算法^[21]的 $O(|C|^3|U| + |C||U|^2)$ 以及哈希广义决策保持前向约简算法^[22]的 $O(|C|^3|U|)$, 且算法 2 中参与计算的对象数量会随着条件属性的不断增加而逐渐减少, 因此算法 2 的计算效率会更高。

3 实 验

本节实验将算法 BGDSP 与经典广义决策保持前向约简算法 FGRAG^[21]、哈希广义决策保持前向约简算法 HSFGR^[22]、哈希广义决策保持后向约简算法 HSBGR, 以及结合蚁群算法^[23-24]思想的广义决策保持约简算法 ACOFG, 针对约简效率进行比较, 与经典广义决策保持前向约简算法 FGRAG 针对约简结果进行正确性及稳定性分析实验。

本节选取了 8 组 UCI 数据集, 均为离散型数据。所选数据集详细信息如表 2 所示。所有实验均在 Intel Core i5-12400F 处理器、16 GB 内存的 Windows 11 操作系统下进行, 实验平台为 2024.1.4 版本的 PyCharm, 采用版本为 3.10 的 Python 实现算法的设计与测试, 确保实验的可执行性。

3.1 差别度单调性与迭代对象数量统计

为检验算法 BGDSP 在约简过程中的差别度单调性, 图 1 展示了不同数据集上, 随约简属性数量逐步增加时差别度的变化趋势。图 1 以所有核属性对应的差别度值作为初始迭代结果, 随后依次引入外部属性中度量指标变化最大的属性进行计算。通过差别度的变化曲线可直观反映其单调性保持情况。

表 2 数据集信息

Table 2 Dataset information

数据集名称	对象数	特征数	类别数
QSAR biodegradation(Qsar)	1 054	41	2
chess2(Ches)	2 300	6	2
kr-vs-kp(Krvs)	3 194	38	2
Hypothyroid(Hypo)	3 770	29	4
Ticdata2000(Ticd)	5 821	85	2
Statlog(Stat)	6 435	36	6
Phishing(Phis)	11 055	30	2
Letter(Lete)	20 000	16	16

图1横坐标表示属性约简过程中外部属性选择的迭代轮次,纵坐标表示每轮加入属性后对应的差别度值。曲线整体呈现单调递减趋势,说明每轮选入的属性均提升了样本的区分能力,保证了约简过程的稳定性与有效性。此外,该趋势也进一步验证了BGDSP算法引入剥离商集方法不会影响最终的广义决策保持约简结果,确保了约简结果的正确性与一致性。

为了更直观地展示所提算法BGDSP在迭代过程中的高效性与优化机制,图2通过3组关键指标的对比曲线,动态呈现了算法在每一轮属性选择中对数据结构的影响。图2横坐标表示选取条件属性的循环总数,纵坐标表示不同指标的数量。图2中分别记录了以蓝色曲线为代表的原始数据集中的对象总数、以黑色曲线为代表的条件属性下划分所得的等价类数量,以及以红色曲线为代表的经过剥离商集处理后,实际参与后续计算的剩余对象数量。

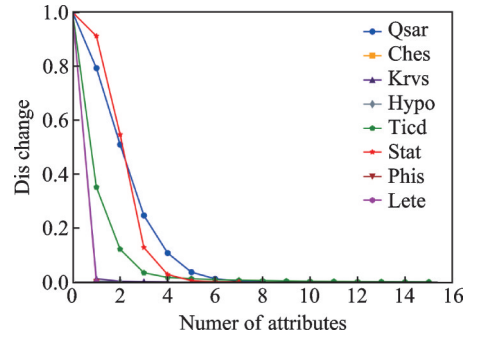
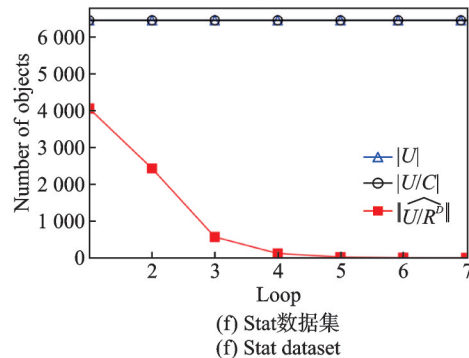
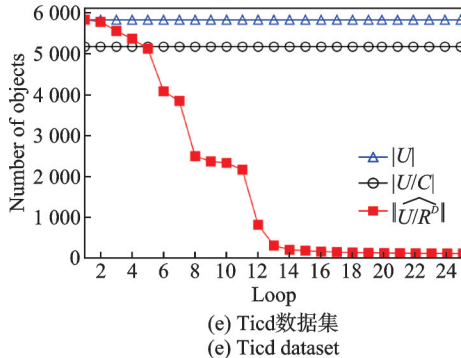
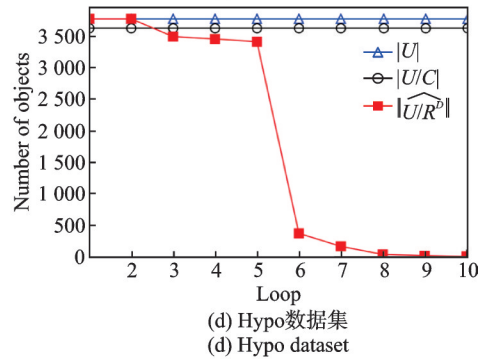
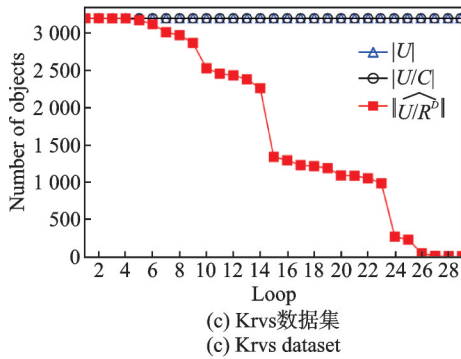
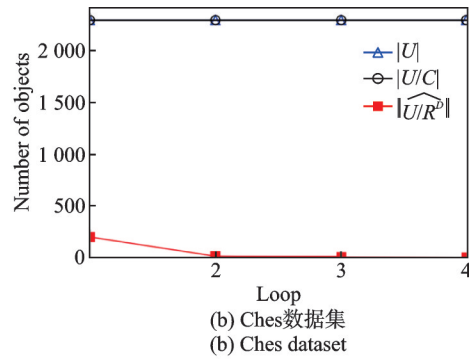
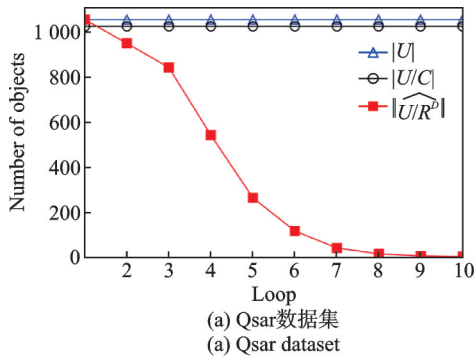


图1 差别度单调性统计

Fig.1 Statistical analysis of the monotonicity property of discernibility degree



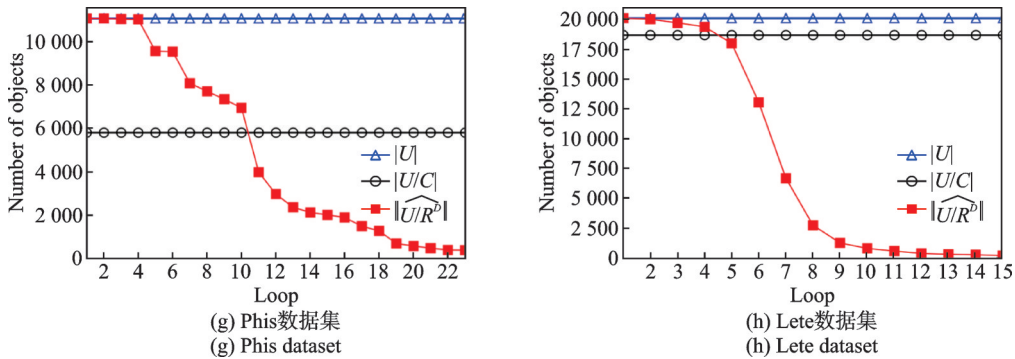


图2 每次循环过程中的对象数量

Fig.2 Number of objects in each iteration

随着迭代轮次的增加,算法逐步选入条件属性并更新当前约简属性集。在此过程中,红色曲线所对应的实际计算对象数量持续下降,表明算法在每轮迭代中均能识别并剔除一定比例的冗余对象。这一趋势不仅显著降低了参与计算的数据规模,也体现了剥离商集在有效过滤信息方面的积极作用。

从曲线变化整体趋势可以看出:蓝色曲线代表的对象总数保持不变,黑色曲线代表的等价类数量一直保持不变。表明经典算法在约简过程中,参与计算的始终为所有数据集下的等价类。而红色曲线呈明显下降趋势表明:随着迭代次数的不断增加,被剥离的冗余对象逐渐增多,实际参与计算的对象数量显著减少。这一现象充分说明BGDSP算法在去除无关信息方面的有效性,同时在保证约简结果正确性的前提下,降低了计算复杂度。实验结果验证了该算法在面对大规模数据时所展现出的适应性与计算效率,为实际应用中的快速属性约简提供了一种可能有效的思路。

3.2 约简效率对比

为了验证BGDSP算法在属性约简任务中的高效性,本节统计了FGRAG、HSFGR、HSBGR、ACOFG以及BGDSP等多算法在多个不同数据集上的约简耗时情况。进一步地,为评估各算法在数据规模变化下的性能表现,分别从数据对象数量增加和条件属性数量增加两个维度出发,统计并分析了各算法约简时间的变化趋势,通过对比统计结果,直观反映了BGDSP算法在约简效率与稳定性方面的优势。

图3展示了5种属性约简算法在8个不同数据集上的属性约简时间统计结果。其中横坐标表示数据集名称,纵坐标采用对数坐标形式展示对应的约简耗时,以方便呈现不同数量级下的时间差异。从图3中可以看出,ACOFG算法和FGRAG算法约简耗时相对较高,且ACOFG算法在不同数据集上呈现的性能存在一定的不一致性,表现出较弱的稳定性。BGDSP算法在所有数据集上的运行时间均低于其他4种对比算法,特别是在处理大规模数据集时,表现出较为明显的性能优势。这一结果充分验证了BGDSP算法在面对高维、海量数据时的有效性。

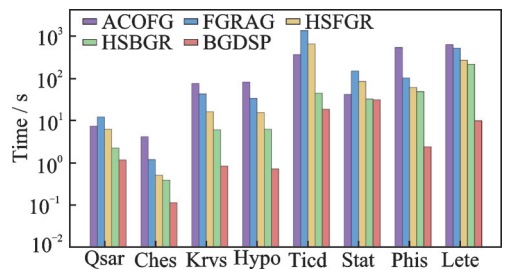


图3 约简时间统计

Fig.3 Reduction time statistics

为进一步分析各算法在不同数据规模条件下的性能适应能力,后续设计了两组实验,分别从数据对象数量增加和条件属性数量增加两个维度,对各算法在数据集上的运行表现进行系统评估,结果如图4,5所示,以展现BGDSP算法在复杂数据环境中的稳定性与处理能力。图4中横坐标表示对象数

量,图5中横坐标表示条件属性数量,纵坐标均表示算法约简所需时间。从图中曲线可以看出,BGDSP算法的约简效率更为高效,且受数据集规模影响较小,运行更加稳定。

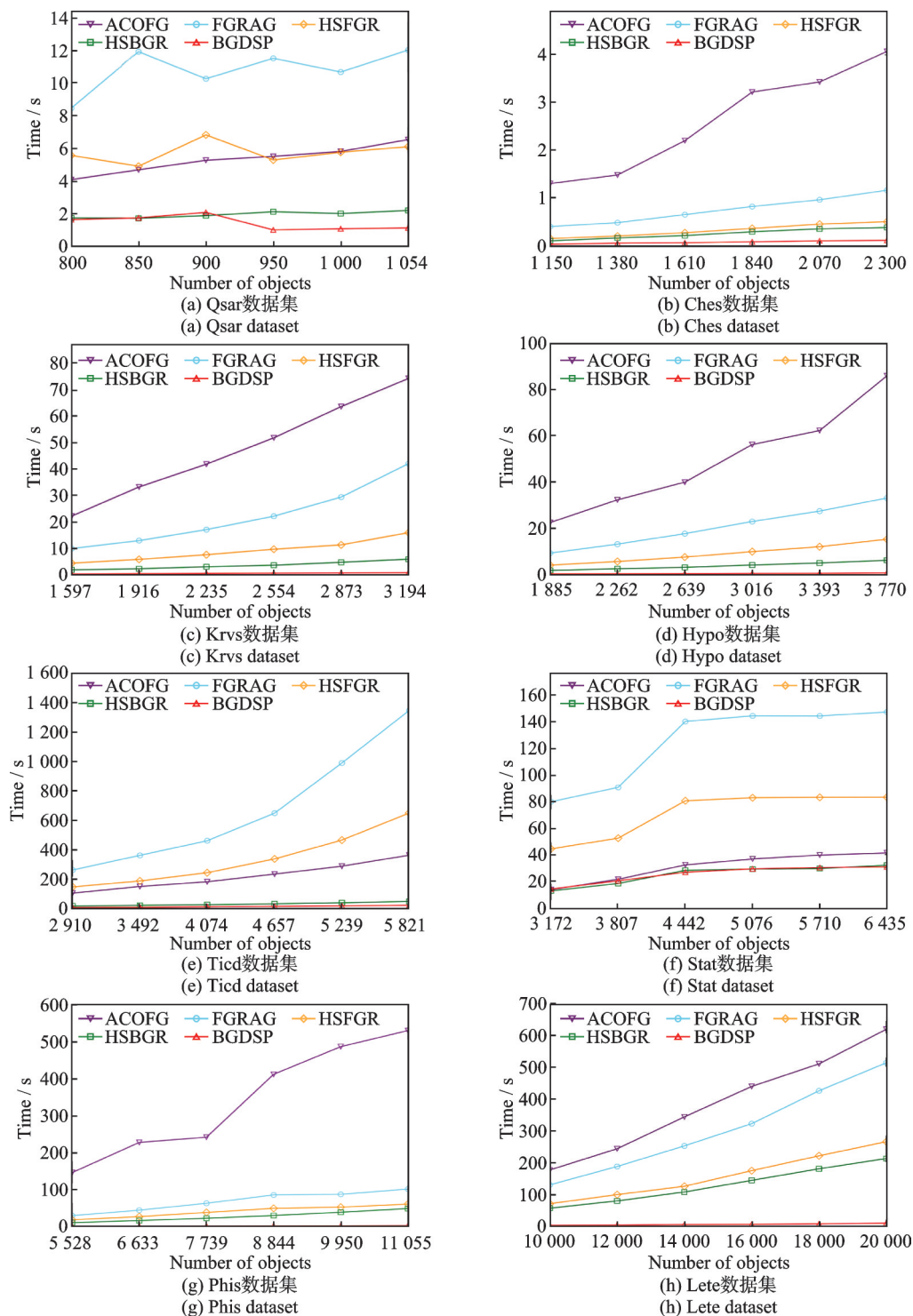


图4 对象增加时间变化比较
Fig.4 Comparative analysis of time variation with increasing objects

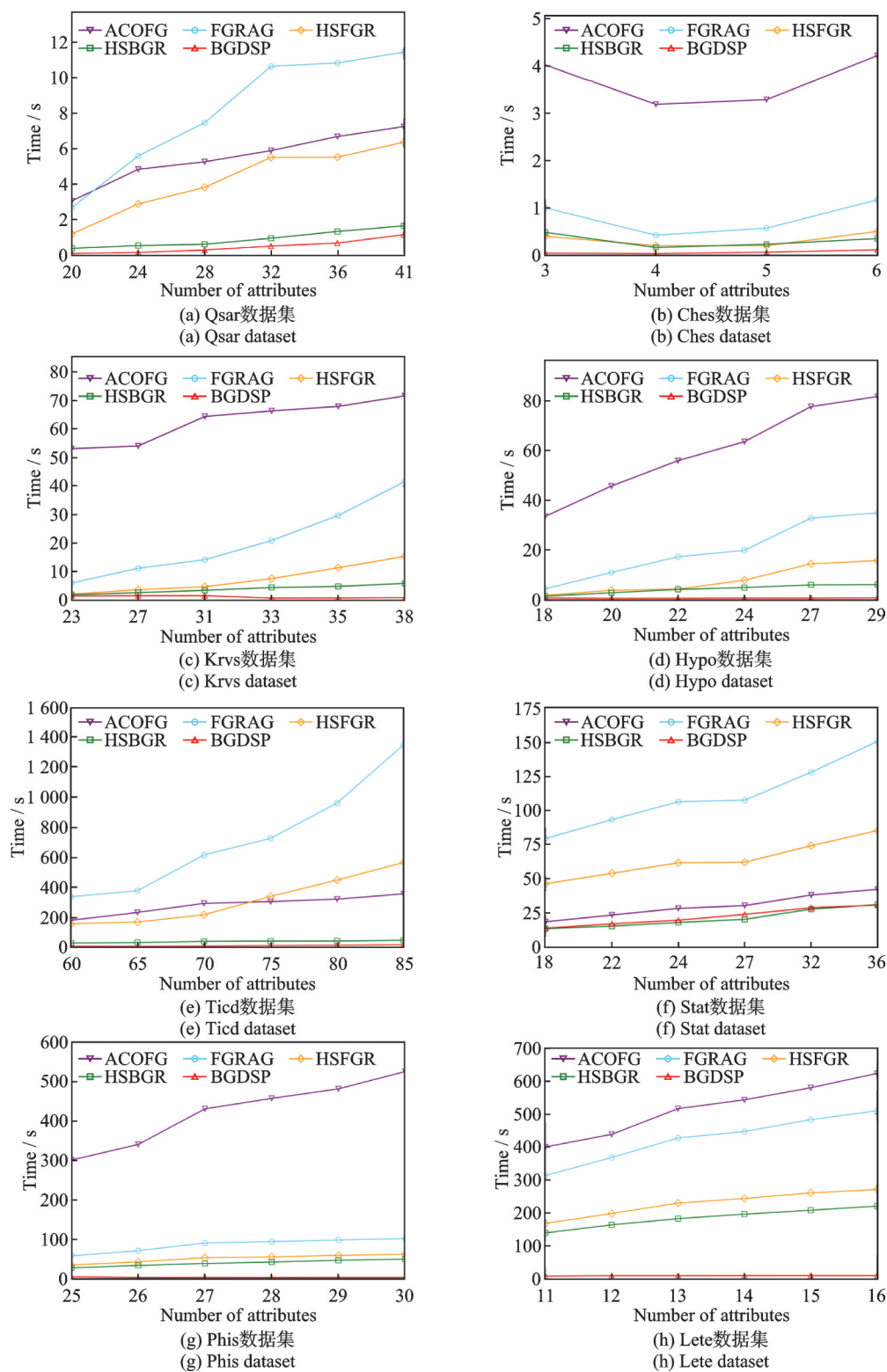


图5 条件属性增加时间变化比较

Fig.5 Comparative analysis of time variation with increasing conditional attributes

3.3 约简有效性与分类精度实验

为验证BGDSP算法约简结果的有效性,表3统计了FGRAG算法与BGDSP算法的约简结果,通过对比证明了BGDSP算法约简结果的准确性。

表3 约简结果统计
Table 3 Reduction result statistics

数据集	FGRAG		BGDSP		约简一致性
	约简结果	长度	约简结果	长度	
Qsar	{0,1,4,8,14,15,16,21,30,34}	10	{0,1,4,8,14,15,16,21,30,34}	10	一致
Ches	{0,2,3,5}	4	{0,2,3,5}	4	一致
Krvs	{0,2,3,4,5,6,⋯,32,33,34,35}	29	{0,2,3,4,5,6,⋯,32,33,34,35}	29	一致
Hypo	{0,1,2,7,9,17,19,21,23,28}	10	{0,1,2,7,9,17,19,21,23,28}	10	一致
Ticd	{0,1,4,6,9,13,⋯,63,67,79,82}	25	{0,1,4,6,9,13,⋯,63,67,79,82}	25	一致
Stat	{0,3,10,13,17,24,35}	7	{0,3,10,13,17,24,35}	7	一致
Phis	{0,1,2,3,5,6,⋯,26,27,28,29}	23	{0,1,2,3,5,6,⋯,26,27,28,29}	23	一致
Lete	{1,2,3,4,5,6,⋯,12,13,14,15}	15	{1,2,3,4,5,6,⋯,12,13,14,15}	15	一致

通过表3可以得知,在实验所选数据集中,FGRAG算法与BGDSP算法在约简结果上完全一致,表明BGDSP算法依然遵循广义决策保持的约简准则。

为进一步验证BGDSP算法的稳定性,表4和表5展示了所选数据在支持向量机(Support vector machine, SVM)与K最近邻(K-nearest neighbor, KNN)分类器下的分类精度,其中第2列为原始数据的

表4 使用SVM分类器的分类精度
Table 4 Classification accuracy of the selected algorithms using SVM classifier

数据集	Raw data	FGRAG	BGDSP
Qsar	84.70±5.57	81.60±6.67	81.60±6.67
Ches	99.60	99.60	99.60
Krvs	93.60±4.26	94.00±3.49	94.00±3.49
Hypo	92.30±0.08	92.30±0.08	92.30±0.08
Ticd	94.00±0.07	94.00±0.07	94.00±0.07
Stat	86.70±3.87	85.00±4.02	85.00±4.02
Phis	94.70±0.70	95.00±0.86	95.00±0.86
Lete	57.80±1.21	58.20±1.23	58.20±1.23

表5 使用KNN分类器的分类精度
Table 5 Classification accuracy of the selected algorithms using KNN classifier

数据集	Raw data	FGRAG	BGDSP
Qsar	83.10±5.11	80.80±6.67	80.80±6.67
Ches	99.50±1.29	99.80±0.22	99.80±0.22
Krvs	77.20±5.35	81.10±5.23	81.10±5.23
Hypo	92.30±1.03	92.30±0.92	92.30±0.92
Ticd	93.50±0.40	93.40±0.40	93.40±0.40
Stat	85.90±4.10	84.40±3.75	84.40±3.75
Phis	94.80±1.31	94.50±1.61	94.50±1.61
Lete	67.80±1.34	66.70±0.98	66.70±0.98

分类精度,第3、4列分别为FGRAG算法和BGDSP算法约简后的分类精度。实验结果显示,BGDSP算法与FGRAG算法在分类精度上保持一致,证明了BGDSP算法约简的稳定性。

4 结束语

本文针对广义决策保持属性约简的效率问题,提出了一种基于剥离商集的快速启发式算法BGDSP。该算法充分利用正域对象的稳定性,构建了更为简化且具有代表性的等价类,在保证约简正确性的基础上提高了约简效率,为相关算法设计提供了一种可能有效的新思路。

参考文献:

- [1] PAWLAK Z. Rough sets[J]. *International Journal of Computer & Information Sciences*, 1982, 11(5): 341-356.
- [2] GAETA A, LOIA V, LOMASTO L, et al. A novel approach based on rough set theory for analyzing information disorder[J]. *Applied Intelligence*, 2023, 53(12): 15993-16014.
- [3] ZHAO Jie, WU Daiyang, ZHOU Yongxin, et al. Rough set theory-based group incremental approach to feature selection[J]. *Information Sciences*, 2024, 675: 120733.
- [4] SEWWANDI M A N D, LI Y F, ZHANG J L. A class-specific feature selection and classification approach using neighborhood rough set and K-nearest neighbor theories[J]. *Applied Soft Computing*, 2023, 143: 110366.
- [5] QIAN Wenbin, WAN Lijuan, SHU Wenhao. Semi-supervised feature selection based on discernibility matrix and mutual information[J]. *Applied Intelligence*, 2024, 54(13): 7278-7295.
- [6] 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.
MIAO Duoqian, HU Guirong. A heuristic algorithm for knowledge reduction[J]. *Journal of Computer Research and Development*, 1999, 36(6): 42-45.
- [7] 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.
ZHANG Nan, MIAO Duoqian, YUE Xiaodong. Knowledge reduction in interval-valued information systems[J]. *Journal of Computer Research and Development*, 2010, 47(8): 1362-1371.
- [8] REN Chungel, ZHU Ping. Attribute reduction based on interval-set rough sets[J]. *Soft Computing*, 2024, 28(3): 1893-1908.
- [9] ZHANG Xianrong, YAO Yiyu. Tri-level attribute reduction in rough set theory[J]. *Expert Systems with Applications*, 2022, 190: 116187.
- [10] GAO Can, LAI Zhihui, ZHOU Jie, et al. Maximum decision entropy-based attribute reduction in decision-theoretic rough set model[J]. *Knowledge-Based Systems*, 2018, 143: 179-191.
- [11] XIE Jingjing, HU Baoqing, JIANG Haibo. A novel method to attribute reduction based on weighted neighborhood probabilistic rough sets[J]. *International Journal of Approximate Reasoning*, 2022, 114: 1-17.
- [12] NGUYEN H S, NGUYEN N T, NGUYEN H S, et al. Some observations on representation of dependency degree $k[C]$ // *Proceedings of 2017 9th International Conference on Knowledge and Systems Engineering*. Warsaw: IEEE, 2017: 13-17.
- [13] NGUYEN N T, SARTRA W. A new approach for reduction of attributes based on stripped quotient sets[J]. *Pattern Recognition*, 2020, 97: 106999.
- [14] NGUYEN N T, SARTRA W. On reduction of attributes in inconsistent decision tables based on information entropies and stripped quotient sets[J]. *Expert Systems with Applications*, 2019, 137: 308-323.
- [15] NGUYEN N T, SARTRA W. A novel feature selection method for high-dimensional mixed decision tables[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, 33(7): 3024-3037.
- [16] QIAN Y H, LIANG J Y, PEDRYCZ W, et al. Positive approximation: An accelerator for attribute reduction in rough set theory[J]. *Artificial Intelligence*, 2010, 174(9/10): 597-618.
- [17] NI Peng, ZHAO Suyun, WANG Xizhao, et al. PARA: A positive-region based attribute reduction accelerator[J]. *Information Sciences*, 2019, 503: 533-550.
- [18] XIA Shuyin, BAI Xinyu, WANG Guoyin, et al. An efficient and accurate rough set for feature selection, classification, and knowledge representation[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2022, 35(8): 7724-7735.

- [19] SANG Binbin, CHEN Hongmei, YANG Lei, et al. Incremental feature selection using a conditional entropy based on fuzzy dominance neighborhood rough sets[J]. IEEE Transactions on Fuzzy Systems, 2021, 30(6): 1683-1697.
- [20] MIAO Duoqian, ZHAO Yan, YAO Yiyu, et al. Relative reducts in consistent and inconsistent decision tables of the Pawlak rough set model[J]. Information Sciences, 2009, 179(24): 4140-4150.
- [21] ZHANG Nan, GAO Xueyi, YU Tianyou. Heuristic approaches to attribute reduction for generalized decision preservation[J]. Applied Sciences, 2019, 9(14): 2841-2860.
- [22] 赵昱德. 一种广义决策保持的快速启发式属性约简算法[J]. 计算机科学与应用, 2024, 14(2): 260-267.
ZHAO Yude. A fast heuristic attribute reduction algorithm for generalized decision preservation[J]. Computer Science and Application, 2024, 14(2): 260-267.
- [23] BLUM C. Ant colony optimization: A bibliometric review[J]. Physics of Life Reviews, 2024, 51: 87-94.
- [24] KORZENÍ M, GISTEREK I. Applying ant colony optimization to reduce tram journey times[J]. Sensors, 2024, 24(19): 6226-6238.

作者简介:



刘中凯(2002-),男,硕士研究生,研究方向:粗糙集、粒计算与人工智能,E-mail: liuzhongkai812@163.com。



张楠(1979-),通信作者,男,副教授,研究方向:粒计算、认知信息学、数据挖掘与人工智能,E-mail: zhangnan0851@163.com。

(编辑:刘彦东)

Efficient Attribute Reduction Algorithm for Generalized Decision Preservation Based on Stripped Quotient Set

LIU Zhongkai, ZHANG Nan*

(School of Computer Science and Control Engineering, Yantai University, Yantai 264005, China)

Abstract: Attribute reduction is one of the core research problems in rough set theory, aiming to identify a more concise subset of attributes without compromising the classification capability of the original dataset. Among various reduction strategies, attribute reduction based on generalized decision preservation has been widely applied due to its strong interpretability and practical applicability. However, existing algorithms for generalized decision-preserving attribute reduction often suffer from significant computational inefficiency when dealing with large-scale datasets. How to efficiently achieve attribute reduction under the generalized decision-preserving framework has become a critical issue that demands urgent attention. To address this problem, this paper explores the stability characteristics of objects in the positive region during the generalized decision-preserving process. By employing the stripped quotient set method, a simplified and more representative equivalence class structure is constructed. On this basis, we propose an efficient heuristic attribute reduction algorithm under the generalized decision preservation framework. The proposed algorithm significantly improves computational efficiency while ensuring the correctness of the reduction results. Experimental results on eight benchmark datasets from University of California, Irvine (UCI) repository demonstrate that the proposed algorithm achieves higher reduction efficiency compared to existing methods.

Highlights:

1. A fast generalized decision-preserving attribute reduction algorithm based on stripped quotient sets (BGDSP) is proposed to improve the computational efficiency of large-scale attribute reduction.
2. A stripped quotient set strategy excludes positive-region objects from subsequent calculations, and a discriminability measure with a proved monotonicity property is developed to guide attribute selection.
3. Experiments on eight data sets from University of California, Irvine (UCI) Machine Learning Repository show that BGDSP achieves the same reduction results as the forward greedy reduction algorithm for generalized decision preservation (FGRAG).

Key words: rough set; attribute reduction; heuristic algorithms; stripped quotient set

Foundation item: Yantai City Science and Technology Development Program Project (No.2022XDRH016).

Received: 2025-06-15; **Revised:** 2025-07-30

***Corresponding author, E-mail:** zhangnan0851@163.com.