Computer Science and Application 计算机科学与应用, 2024, 14(2), 260-267
Published Online February 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.142026

Hans汉斯

一种广义决策保持的快速启发式属性约简算法

赵昱德

烟台大学计算机与控制工程学院，山东 烟台

收稿日期：2024年1月19日；录用日期：2024年2月19日；发布日期：2024年2月27日

摘　要

属性约简是粗糙集理论的重要概念之一，旨在获得一个可以保持原始信息系统分类能力的最小属性子集。广义决策保持约简是粗糙集中的属性约简方法之一，其目标为维护决策系统中的决策结果，确保在约简过程中不丢失原始决策。这意味着约简后的系统仍可正确地进行决策，而决策规则的有效性和决策能力得以保持。传统的广义决策保持启发式属性约简算法注重算法的有效性，而算法的效率有待优化。传统算法在计算广义决策保持相似度时需多次遍历每个对象的等价类与决策类，存在大量的重复计算。为了克服这个问题，我们通过引入哈希表来存储每个对象的等价类与其广义决策，使得计算广义决策保持相似度时可针对计算对象直接得出结果而不是依次遍历，由此提出了广义决策保持的快速启发式属性约简算法。最后，通过6组UCI数据集验证了本文提出算法的有效性与高效性。

关键词

属性约简，粗糙集，启发式算法，广义决策保持

A Fast Heuristic Attribute Reduction Algorithm for Generalized Decision Preservation

Yude Zhao

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

Received: Jan. 19th, 2024; accepted: Feb. 19th, 2024; published: Feb. 27th, 2024

Abstract

Attribute reduction is one of the key concepts in rough set theory, aimed at obtaining the minimal subset of attributes that retains the classification capability of the original information system. Generalized Decision-Preserving Reduction is a method of attribute reduction within rough set

文章引用: 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.
DOI: 10.12677/csa.2024.142026

theory, which focuses on maintaining the outcomes of decision-making systems, ensuring that the original decisions are not lost during the reduction process. This implies that the reduced system can still make correct decisions, with the effectiveness and capability of decision rules maintained. Traditional algorithms prioritize effectiveness, but their efficiency needs improvement. These conventional algorithms require multiple traversals through each object's equivalence class and decision class to compute the similarity degree for generalized decision preservation, leading to extensive redundant computations. To overcome this issue, we introduce the use of hash tables to store each object's equivalence class and its generalized decision, allowing direct computation of similarity degree for generalized decision preservation for a computational object, rather than sequential traversals, thereby accelerating the algorithm. Finally, the efficacy and efficiency of the proposed algorithm are validated through six sets of UCI datasets.

Keywords

Attribute Reduction, Rough Set, Heuristic Algorithms, Generalized Decision Preservation

1. 引言

1982 年波兰数学家 Pawlak [1] [2]为了处理不准确、不确定的数据，提出了粗糙集理论。在粗糙集理论中，属性约简[3]-[8]是一个关键概念，其主旨在于通过识别数据集中最关键的属性集，在不影响原始分类能力的情况下简化信息系统的表示，减少冗余属性，提高对大规模数据和高维数据的处理效率。广义决策保持约简[9] [10] [11]是粗糙集理论属性约简中的一个重要的算法，旨在选择的最小属性子集可以保持决策能力，提高决策系统的效率和可解释性。与传统属性约简侧重于等价关系和近似空间不同，广义决策保持约简强调在简化属性集的同时，确保维持决策系统的原始决策，这有助于保证约简后属性子集的可解释性，使得约简结果更容易为人理解和接受。

随着数据集规模的不断增大和特征维度的提高，传统的属性约简算法可能面临巨大的计算复杂性和资源消耗。而在实际应用中，如大规模数据挖掘和智能决策系统，对计算时间的迅速响应是至关重要的。在这一背景下，研究者致力于设计高效的属性约简算法，以加速属性约简的过程[12] [13] [14] [15] [16]。通过降低计算复杂性，这些改进不仅有助于在有限时间内处理大规模数据，同时也提升了属性约简方法的实用性和可扩展性。

由于广义决策保持约简算法的计算过程需计算每个对象所在等价类的广义决策相似度，而计算广义决策相似度需多次遍历全部决策类并依次计算结果，若使等价类与决策类一一对应，则可通过单次计算直接得到广义决策相似度，从而避免了重复计算。本文在广义决策保持启发式属性约简的基础上，提出了可使每个等价类与其相关决策类对应的广义决策哈希表，由此实现了广义决策保持的快速启发式属性约简算法。最后，经过实验证明了本文所提算法的高效性。

2. 基本概念

2.1. Pawlak 经典粗糙集理论

定义 1. [1]给定信息系统 IS 是一个四元组：$IVDS=(U,A,V,f)$，其中 U 是一个非空有限集合称为论

域，A 表示一个非空有限的属性集合，V 表示的是属性集合 A 的值域，f 是一个映射函数，其中 $f:U\times AT\to V$，$f(x,a)$ 表示对象 $x\in U$ 在属性 $a\in A$ 上的取值，简记为 $a(x)$。若属性集 A 可分为条件属性 C 与决策属性 D，即 $A=C\cup D$，则四元组 $(U,A=C\cup D,V,f)$ 为一个决策系统，$f(x,d)$ 表示对象 x 在属性 $d\in D$ 上的取值，简记为 $d(x)$。

定义 2. [1]在决策系统 $DS=(U,A=C\cup D,V,f)$ 中，设 $\forall P\subseteq C$，若 $x_a,x_b\in U$，则属性集 Q 下的不可分辨关系定义为：

$$Ind(P)=\{(x_a,x_b)\in U\times U\mid \forall p\in P,p(x_a)=p(x_b)\} \tag{1}$$

显然，$Ind(Q)$ 是对称的、传递的、自反的等价关系。对于论域 U，$Ind(P)$ 可将其划分为不同等价类，即 $U/Ind(P)=U/P=\{[x]_P\mid x\in U\}$，其中 $[x]_P=\{y\in U\mid (x,y)\in Ind(P)\}$。对于决策属性 D，不可分辨关系划分所得的 U/D 中的对象被称为决策类。

定义 3. [1]在决策系统 $DS=(U,A=C\cup D,V,f)$ 中，设 $\forall P\subseteq C$，$D_m\subseteq U/D$，属性集 P 对于 D_m 的上近似与下近似的定义为：

$$\underline{P}(D_m)=\{x\mid [x]_P\subseteq D_m\}=\cup\{[x]_P\mid [x]_P\subseteq D_m\} \tag{2}$$

$$\overline{P}(D_m)=\{x\mid [x]_P\cap D_m\neq\varnothing\}=\cup\{[x]_P\mid [x]_P\cap D_m\neq\varnothing\} \tag{3}$$

对于 $\forall P\subseteq C$，其下近似中的对象为确定属于 D_m，其上近似中的对象代表其可能属于 D_m。由下近似与上近似可将论域 U 划分为三个区域，即正域、边界域、负域，定义如下：

$$Pos_P=\underline{P}(D_m) \tag{4}$$

$$Bnd_P=\overline{P}(D_m)-\underline{P}(D_m) \tag{5}$$

$$Neg_P=U-\overline{P}(D_m) \tag{6}$$

其中正域与负域分别代表了确定属于 D_m 的对象与确定不属于 D_m 的对象，为确定性信息，而边界域代表了可能属于 D_m 的对象，为不确定性信息。此外，论域 U 与正域、边界域、负域的关系为：

$$U=Pos_P(D_m)+Bnd_P(D_m)+Neg_P(D_m) \tag{7}$$

2.2. 基于广义决策保持的启发式属性约简算法

定义 4. [11]在决策系统 $DS=(U,A=C\cup D,V,f)$ 中，设 $\forall P\subseteq C$，对于 $x_a\in U$，$[x_a]_P\in U/P$，对象 x_a 在属性集 P 下的广义决策为：

$$\delta_P(x_a)=\{d(x_b)\mid x_b\in[x_a]_P\} \tag{8}$$

定义 5. [11]在决策系统 $DS=(U,A=C\cup D,V,f)$ 中，设 $\forall P\subseteq C$，$x_a\in U$，若 $[x_a]_P\in U/P$，$[x_a]_C\in U/C$，广义决策保持的相似度定义为：

$$Sim(P,C)=\begin{cases}\dfrac{\left|\cup_{i=1}^{|U|}\left([x_a]_P\cap[x_a]_C\right)\right|}{|U|}, & \delta_P(x_a)=\delta_C(x_a)\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

定义 6. [11]在决策系统 $DS=(U,A=C\cup D,V,f)$ 中，当且仅当以下两个条件成立时，$P\subseteq C$ 为此决策系统的一个广义决策保持约简：

1) 联合充分性：$Sim(P,C)=1$；

2) 个体必要性： $\forall Q \subset P$，满足 $Sim(P,Q) \neq 1$。

定义 7. [11]在决策系统 $DS=(U,A=C\cup D,V,f)$ 中，设 $\forall P \subseteq C$， $p \in P$，广义决策保持的内部属性重要度定义为：

$$Sig^{inner}(p,P,C)=Sim(P,C)-Sim(P-\{p\},C) \tag{10}$$

定义 8. [11]在决策系统 $DS=(U,A=C\cup D,V,f)$ 中，设 $\forall P \subseteq C$，若 $\exists p \in P$ 存在 $Sig^{inner}(p,P,C)>0$，则 $p \in Core$。

定义 9. [11]在决策系统 $DS=(U,A=C\cup D,V,f)$ 中，设 $\forall P \subseteq C$， $p \in P$，广义决策保持的内部属性重要度定义为：

$$Sig^{outer}(p,P,C)=Sim(P\cup\{p\},C)-Sim(P,C) \tag{11}$$

基于上述定义，可知广义决策保持的前向启发式属性约简算法(A Forward Greedy Reduction Algorithm for Generalized Decision Preservation, FGRAG) [11]如表 1 所示。

Table 1. A forward greedy reduction algorithm for generalized decision preservation (FGRAG)
表 1. 广义决策保持的前向启发式属性约简算法

输入： 决策系统 $DS=(U,C\cup D,V,f)$。

输出： 约简 P。

1. 设 $Core=\varnothing$ 并对于 $x_i \in U$ 计算 $\delta_C(x_i)$；
2. 根由定义 8 求出 $Core$，并使 $P=Core$；
3. 循环选择 $C-P$ 中外部属性重要度最高的属性加入属性集 P，直到 $Sim(P,C)=1$；
4. 对于属性集 P 中的属性，若其内部属性重要度为 0，则将其自属性集 P 中删去；
5. 返回约简集合 P。

由于计算广义决策保持的相似度时，对于 $\delta_P(x)=\delta_C(x)$ 的对象需要遍历所有等价类才可得出 $Sim(P,C)$，即时间复杂度为 $O(|U|)$，对于每个符合条件的对象，总体时间复杂度为 $O(|U|^2)$。因此，在上述 FGRAG 算法中，步骤 1 与步骤 2 的时间复杂度为 $O(|C||U|+|C|(|C||U|+|U|^2))$，步骤 3 的时间复杂度为 $O(|C|(|C|^2|U|+|U|^2))$，步骤 4 的时间复杂度为 $O(|C|(|C||U|+|U|^2))$，步骤 5 的时间复杂度为 $O(1)$。因此，FGRAG 算法的时间复杂度为 $O(|C|^3|U|+|C||U|^2)$。

3. 广义决策保持的快速启发式属性约简算法

由于计算广义决策保持的相似度时需多次遍历等价类，所以本节主要通过实现广义决策保持的相似度的快速计算方法从而加速传统广义决策保持的启发式属性约简算法。

在计算广义决策保持的相似度前，需计算属性子集 $P \subseteq C$ 下的广义决策哈希表，如表 2 所示：

Table 2. Algorithm for computing generalized decision hash tables (GDHT)
表 2. 广义决策哈希表的计算算法

输入： 决策系统 $DS=(U,C\cup D,V,f)$， $P \subseteq C$。

输出： $P \subseteq C$ 下的广义决策哈希表 $Hash_P$。

1. 设 $Hash_P=\varnothing$；

续表

2. 对于每个 $x_i \in U$，将其条件属性值组为字符串 $str_{x_i}^P$：

2.1. 若 $str_{x_i}^P$ 在 $Hash_P$ 的键中不存在，则将 $str_{x_i}^P$ 作为键，$equ_P = [x_i]$ 与 $\delta_P = d(x_i)$ 作为值加入 $Hash_P$；

2.2. 若 $str_{x_i}^P$ 在 $Hash_P$ 的键中存在，则将 x_i 加入到 equ 中并更新 δ_P；

3. 返回广义决策哈希表 $Hash_P$。

在算法 GDHT 中，由于仅遍历一次论域，所以时间复杂度为 $O(|C||U|)$。在算法结束后，广义决策哈希表 $Hash_P$ 中所存储的内容为键与其对应的值，其中键为由不同等价类的条件属性值组成的字符串，值为该条件属性值下的等价类与其广义决策。得益于哈希表查找的高效性，若已知某一对象的条件属性值，则可在 $O(1)$ 的时间复杂度下确定其所属的等价类与其广义决策。

此外，若 $P = C$，则 $Hash_C$ 为一个特殊的广义决策哈希表，由于在后续计算过程中需频繁查找 $Hash_C$ 且 $Hash_C$ 仅需计算一次，所以后续算法均默认 $Hash_C$ 的存在。对于 $P \subseteq C$，通过 $Hash_C$ 与 $Hash_P$ 可得广义决策保持的相似度的快速计算算法，如表 3 所示：

Table 3. A fast algorithm for computingsimilarity degree for generalized decision preservation(FCSG)
表 3. 广义决策保持的相似度的快速计算算法

输入： 决策系统 $DS = (U, C \cup D, V, f)$，$Hash_C$，$Hash_P$。

输出： $P \subseteq C$ 下的广义决策哈希表 $Hash_P$。

1. 设 $temp = 0$；
2. 对于每个 $x_i \in U$，将其在属性集合 C 与属性集合 P 下的属性值组为字符串 $str_{x_i}^C$ 与字符串 $str_{x_i}^P$；
3. 于 $Hash_C$ 中查找 $str_{x_i}^C$，于 $Hash_P$ 中查找 $str_{x_i}^P$：
2.1. 若 $\delta_C = \delta_P$，则 $temp = temp + |equ_C \cap equ_P|$；
2.2. 若 $\delta_C \neq \delta_P$，则跳过此 x_i；
4. $Sim(P,C) = temp/|U|$；
5. 返回广义决策保持相似度 $Sim(P,C)$

在算法 FCSG 中，通过哈希表查找的高效性可快速定位指定对象的等价类与广义决策，而不用重复遍历。算法时间复杂度为 $O(|U|)$。由此可得出广义决策保持的快速启发式属性约简算法(A Fast Heuristic Attribute Reduction Algorithm for Generalized Decision Preservation, FAGD)，如表 4 所示：

Table 4. A fast heuristic attribute reduction algorithm for generalized decision preservation (FAGD)
表 4. 广义决策保持的快速启发式属性约简算法

输入： 决策系统 $DS = (U, C \cup D, V, f)$。

输出： 约简 P。

1. 由算法 GDHT 计算 $Hash_C$；
2. 结合算法 FCSG，通过定义 8 求出 $Core$，并使 $P = Core$；
3. 结合算法 FCSG，循环选择 $C - P$ 中外部属性重要度最高的属性加入属性集 P，直到 $Sim(P,C) = 1$；
4. 对于属性集 P 中的属性，若其内部属性重要度为 0，则将其自属性集 P 中删去；
5. 返回约简集合 P。

在算法 FAGD 中，步骤 1 与步骤 2 的时间复杂度为 $O(|C||U| + |C|((|C||U| + |U|))$，步骤 3 的时间复杂度

为$O\left(|C|\left(|C|^2|U|+|C||U|\right)\right)$，步骤 4 的时间复杂度为$O\left(|C|\left(|C||U|+|C||U|\right)\right)$，步骤 5 的时间复杂度为$O(1)$。因此，FAGD 算法的时间复杂度为$O\left(|C|^3|U|\right)$，优于算法 FGRAG 的$O\left(|C|^3|U|+|C||U|^2\right)$。

4. 实验分析

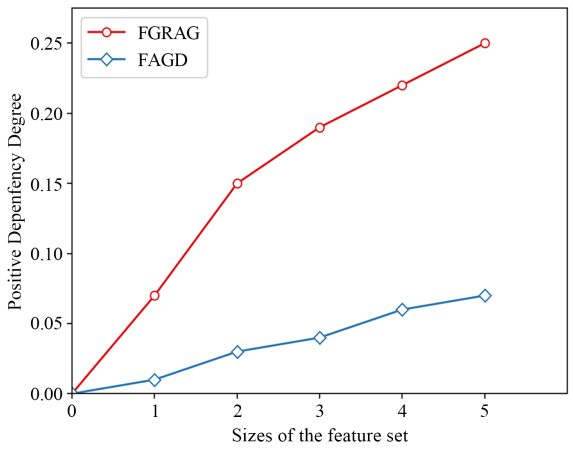
本实验选取 6 个 UCI 数据集进行约简效率的对比，数据集的详细信息如表 5 所示。本节主要对传统的广义决策保持的前向启发式属性约简算法(FGRAG)与本文提出的广义决策保持的快速启发式属性约简算法(FAGD)约简效率进行对比。本实验在 i7-8750h 处理器、16.00 GB 内存和 MacOS 12.6 操作系统下进行。

Table 5. DatasetInformation
表 5. 实验使用的数据集

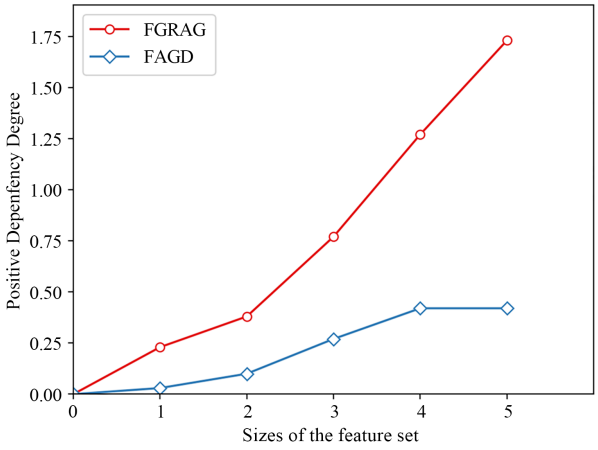
序号	名称	对象数	特征数	分类数
1	Breast-w	699	9	2
2	Vehicle	846	18	4
3	German	1000	20	2
4	House	506	13	4
5	SolarFlare	1066	12	6
6	Primary-tumor	339	17	21

约简效率

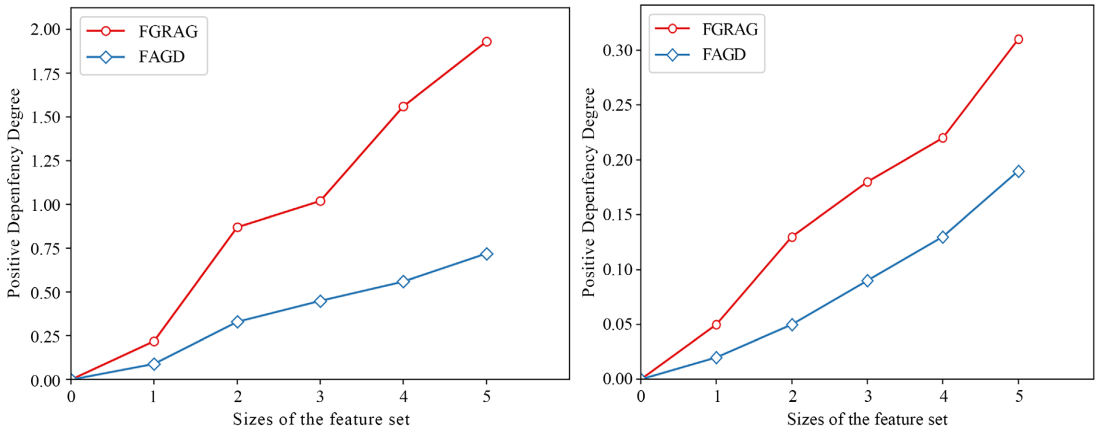
传统的广义决策保持的前向启发式属性约简算法(FGRAG)与本文提出的广义决策保持的快速启发式属性约简算法(FAGD)按照论域规模变化的时间效率对比图。对象规模变化规则如下：将论域平均划分为 5 份。横坐标表示参与实验的论域规模，即原始数据的 20%、40%、60%、80%、100%。纵坐标是约简所消耗的时间单位是秒，红色是广义决策保持的前向启发式属性约简算法(FGRAG)，蓝色是广义决策保持的快速启发式属性约简算法(FAGD)。



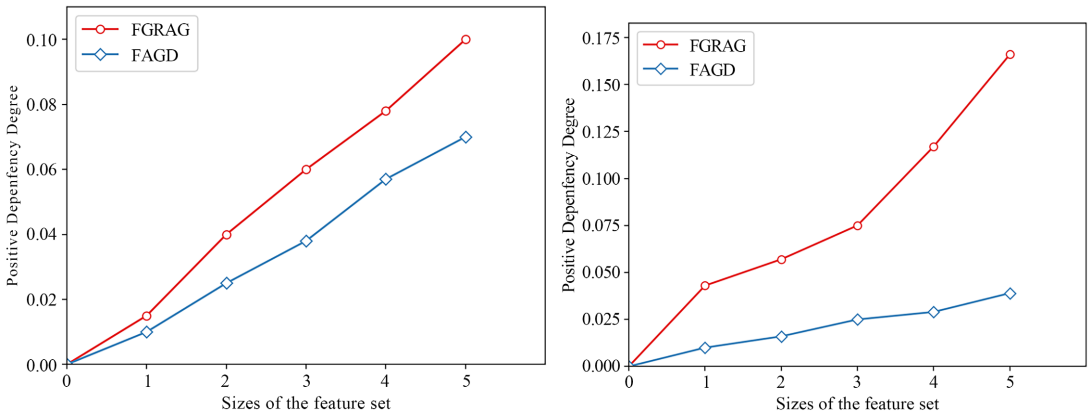
(a) Breast-w



(b) Vehicle



(c) German (d)House



(e) Solar Flare (f) Primary-tumor

Figure 1. Comparison of reduction efficiency
图 1. 约简效率对比

由图 1 可知，当论域规模较小时，FAGD 算法与 FGRAG 算法的效率差别不大。当随着论域规模逐渐增加时，两种算法的执行时间均有上升，但是本文提出的算法随着论域规模的增加变化比较均匀，稳定性更好。本文提出的 FAGD 算法的折线始终保持在 FGRAG 算法折线的下方，表明 FAGD 算法一直保持较好的效率。当论域规模较大时，两种算法直接的差距更为明显，以数据集 Vehicle 为例，FGRAG 算法运行时间为 1.73 秒，而 FAGD 算法仅为 0.42 秒，仅为 FGRAG 算法的四分之一左右，相对来说可以证明本文提出的算法更加适用于大规模的数据集。因此，本文提出的 FAGD 算法与 FGRAG 算法相比更具稳定性与高效性性。

5. 结论

在本文中，首先提出了广义决策哈希表概念，通过此概念实现了广义决策保持的相似度的快速计算算法，并通过加速广义决策保持相似度的计算效率实现了广义决策保持的快速启发式属性约简算法(FAGD)。最后通过选取的六组 UCI 数据集验证了算法的高效性。从实验结果可以看出，本文所提的算法相比于传统算法效率更高。

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 and Information Sciences*, **11**, 341-356. https://doi.org/10.1007/BF01001956

[2] Pawlak, Z. (1992) Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Berlin. https://doi.org/10.1007/978-94-011-3534-4_7

[3] 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(6): 42-45.

[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] Sang, B., Chen, H., Yang, L., *et al*. (2021) Feature Selection for Dynamic Interval-Valued Ordered Data Based on Fuzzy Dominance Neighborhood Rough Set. *Knowledge-Based Systems*, **227**, Article ID: 107223. ttps://doi.org/10.1016/j.knosys.2021.107223

[6] Yang, X., Li, M., Fujita, H., *et al*. (2022) Incremental Rough Reduction with Stable Attribute Group. *Information Sciences*, **589**, 283-299. https://doi.org/10.1016/j.ins.2021.12.119

[7] Xie, J., Hu, B.Q. and Jiang, H. (2022) A Novel Method to Attribute Reduction Based on Weighted Neighborhood Probabilistic Rough Sets. *International Journal of Approximate Reasoning*, **144**, 1-17. https://doi.org/10.1016/j.ijar.2022.01.010

[8] Zhang, X. and Yao, Y. (2022) Tri-Level Attribute Reduction in Rough Set Theory. *Expert Systems with Applications*, **190**, Article ID: 116187. https://doi.org/10.1016/j.eswa.2021.116187

[9] Kryszkiewicz, M. (1998) Rough Set Approach to Incomplete Information Systems. *Information Sciences*, **112**, 39-49. https://doi.org/10.1016/S0020-0255(98)10019-1

[10] Miao, D.Q., Zhao, Y., Yao, Y.Y., *et al*. (2009) Relative Reducts in Consistent and Inconsistent Decision Tables of the Pawlak Rough Set Model. *Information Sciences*, **179**, 4140-4150. https://doi.org/10.1016/j.ins.2009.08.020

[11] Zhang, N., Gao, X. and Yu, T. (2019) Heuristic Approaches to Attribute Reduction for Generalized Decision Preservation. *Applied Sciences*, **9**, Article 2841. https://doi.org/10.3390/app9142841

[12] Qian, Y.H., Liang, J.Y., Pedrycz, W., *et al*. (2010) Positive Approximation: An Accelerator for Attribute Reduction in Rough Set Theory. *Artificial Intelligence*, **174**, 597-618. https://doi.org/10.1016/j.artint.2010.04.018

[13] Xia, H., Chen, Z., Wu, Y.M., *et al*. (2022) Attribute Reduction Method Based on Improved Granular Ball Neighborhood Rough Set. 2022 7*th International Conference on Cloud Computing and Big Data Analytics* (*ICCCBDA*), Chengdu, 22-24 April 2022, 13-16. https://doi.org/10.1109/ICCCBDA55098.2022.9778889

[14] Sang, B.B., Chen, H.M., Yang, L., *et al*. (2022) Incremental Feature Selection Using a Conditional Entropy Based on Fuzzy Domi-nance Neighborhood Rough Sets. *IEEE Transactions on Fuzzy Systems*, **30**, 1683-1697. https://doi.org/10.1109/TFUZZ.2021.3064686

[15] Xia, S.Y., Wang, G.Y. and Gao, X. (2023) An Efficient and Accurate Rough Set for Feature Selection, Classification, and Knowledge Representation. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 7724-7735. https://doi.org/10.1109/TKDE.2022.3220200

[16] Mao, H., Wang, S.Y., Liu, C., *et al*. (2023) Hypergraph-Based Attribute Reduction of Formal Contexts in Rough Sets. *Expert Systems with Applications*, **234**, Article ID: 121062. https://doi.org/10.1016/j.eswa.2023.121062