Hans Journal of Data Mining 数据挖掘, 2023, 13(3), 222-229
Published Online July 2023 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2023.133022

Hans汉斯

基于最大决策熵的快速属性约简算法

袁　梅

烟台大学计算机与控制工程学院，山东　烟台

收稿日期：2023年5月27日；录用日期：2023年6月27日；发布日期：2023年7月5日

摘　要

在大数据时代背景下，各领域数据爆炸式增长，数据类型复杂多样。针对决策系统中基于最大决策熵的属性约简算法在大规模数据集下运行效率低的问题，提出了一种基于启发式的快速属性约简算法。本文提出的算法首先研究了属性和对象在属性约简过程中的变化对其产生影响，其次提出了属性重要度保序性的相关定理。最后通过UCI数据集对提出算法的有效性进行验证，结果表明提出的快速属性约简算法的运行效率更高。

关键词

快速属性约简算法，粗糙集，最大决策熵，决策系统

Fast Attribute Reduction Algorithm Based on Maximum Decision Entropy

Mei Yuan

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

Received: May 27th, 2023; accepted: Jun. 27th, 2023; published: Jul. 5th, 2023

Abstract

In the era of big data, data in various fields is growing explosively, and data types are complex and diverse. Aiming at the low efficiency of attribute reduction algorithm based on maximum decision entropy in decision system under large data sets, a fast attribute reduction algorithm based on heuristic is proposed. The algorithm proposed in this paper firstly studies the influence of the changes of attributes and objects in the process of attribute reduction, and then puts forward the related theorem about the rank preservation of attributes. Finally, the effectiveness of the proposed algorithm is verified by the UCI data set, and the results show that the proposed fast attribute reduction algorithm is more efficient.

文章引用: 袁梅. 基于最大决策熵的快速属性约简算法[J]. 数据挖掘, 2023, 13(3): 222-229.
DOI: 10.12677/hjdm.2023.133022

Keywords

Fast Attribute Reduction Algorithm, Rough Set, Maximum Decision Entropy, Decision System

1. 引言

粗糙集理论是用于处理不精确、不一致、不完备信息和知识的有效工具[1] [2]。如今，学者们对粗糙集理论已经进行了深入探索，相应的属性约简[3] [4] [5] [6]方法也较为完善。Kryszkiewicz [7]在不完备决策系统下引入广义决策保持约简，介绍了相关决策规则的提取，并提出了基于差别矩阵的广义决策保持约简方法。差别矩阵方法虽然可以求出所有约简结果，但其效率相对于启发式算法较低。2002 年王国胤等[8]从信息论观点出发，将条件信息熵作为启发式信息，设计了启发式属性约简算法；2018 年，Gao [9]提出了最大决策熵的启发式属性约简算法。2019 年 Zhang 等[10]等提出了启发式的广义决策属性约简。

现阶段，对于大规模数据集，有关属性约简的快速算法研究已取得许多成果。2006 年，徐章艳等[11]提出了基于基数排序的快速属性约简算法；2010 年，Qian 等[12]提出了正域加速属性约简算法，2018 年，Du 等[13]在序决策系统下提出了快速属性约简算法。另外，增量式属性约简算法[14] [15] [16] [17]利用已有的信息进行增量更新，不需要重新计算，从而实现算法效率的提高。本文从对象和属性的角度考虑研究，通过理论分析和实验结果均表明了该算法的有效性。

2. 基本概念

定义 1 [1]信息系统是由四元组 $IS=(U,AT,V,f)$ 组成，其中 U 表示论域，是非空有限对象组成的集合；AT 表示非空有限属性集合；V_p 表示属性 $p\in AT$ 的值域，有 $V=\bigcup_{P\in AT}V_p$；f 是一个映射函数，$f:U\times AT\to V$ 为论域 U 中的每一个对象在 $\forall p\in AT$ 上都有一个值。若 $AT=C\cup D$，其中 C 表示非空有限的条件属性集合，D 表示非空有限的决策属性集合，且 $C\cap D\neq\varnothing$，则四元组记为 $DS=(U,AT=C\cup D,V,f)$ 称为决策信息系统。

定义 2 [1]四元组 $DS=(U,AT=C\cup D,V,f)$ 为一个决策信息系统，对任意非空属性集合 $P\subseteq AT$，有 P 在 U 上的不可区分关系定义为：

$$IND(P)=\{(x,y)\in U\times U\mid p(x)=p(y),\forall p\in P\} \tag{1}$$

不可区分关系 $IND(P)$ 是一个满足自反性、对称性和传递性的等价关系。由不可区分关 $IND(P)$ 导出对论域 U 的划分为 $U/IND(P)=\{[x]_P\mid x\in U\}$，通常简写为 U/P，其中 $[x]_P$ 表示包含 x 的等价类，易得 $[x]_{IND(P)}=\bigcap_{p\in P}[x]_p$。

定义 3 [1]决策信息系统的四元组 $DS=(U,AT=C\cup D,V,f)$，由决策属性 D 导出 U 的划分为 $U/D=\{D_1,D_2,\cdots,D_m\}(1\le m\le|U|)$，对 $\forall P\subseteq C$，决策类 U/D 关于条件属性集 P 的下近似和上近似的定义为：

$$\underline{P}(U/D)=\{\underline{P}(D_1),\underline{P}(D_2),\cdots,\underline{P}(D_m)\} \tag{2}$$

$$\overline{P}(U/D)=\left\{\overline{P}(D_1),\overline{P}(D_2),\cdots,\overline{P}(D_m)\right\} \tag{3}$$

决策类 U/D 关于条件属性集 P 的正域和边界域的定义：

$$POS_P(U/D)=\bigcup_{D_i\in U/D}\underline{P}(D_i) \tag{4}$$

$$BND_P(U/D)=\bigcup_{D_i\in U/D}\overline{P}(D_i)-\bigcup_{D_i\in U/D}\underline{P}(D_i) \tag{5}$$

定义 4 [9]决策信息系统 $DS=(U,AT=C\cup D,V,f)$，U 在 C 以及 D 上的划分分别为 $U/C=\{U_1,U_2,\cdots,U_m\}$，$U/D=\{Y_1,Y_2,\cdots,Y_n\}$，其中 $m=|U/C|$，$n=|U/D|$。对于任意一个等价类 $U_i\in U/C$，该等价类的最大包含度以及最大决策分别定义为：

$$MP(D|U_i)=\max\left(P(Y_1|U_i),P(Y_2|U_i),\cdots,P(Y_n|U_i)\right) \tag{6}$$

$$MD(D|U_i)=\left\{f(y,D)\,|\,y\in Y_j\wedge P(Y_1|U_i)=MP(D|U_i)\right\} \tag{7}$$

定义 5 [9]决策信息系统 $DS=(U,AT=C\cup D,V,f)$，U 在 C 以及 D 上的划分分别为 $U/C=\{U_1,U_2,\cdots,U_m\}$，$U/D=\{Y_1,Y_2,\cdots,Y_n\}$，其中 $m=|U/C|$，$n=|U/D|$。C 相对于 D 的最大包含度的概率分布定义为：

$$\begin{aligned}MS(D|C)=&\left(\left(MP(D|U_1),1-MP(D|U_1)\right),\left(MP(D|U_2),1-MP(D|U_2)\right),\cdots,\right.\\&\left.\left(MP(D|U_m),1-MP(D|U_m)\right)\right)\end{aligned} \tag{8}$$

定义 6 [9]决策信息系统 $DS=(U,AT=C\cup D,V,f)$，若 $Q\subseteq C$，Q 相对于 D 的最大包含度的概率分布定义为 $MS(D|Q)=\left(\left(MP(D|U_1),1-MP(D|U_1)\right),\left(MP(D|U_2),1-MP(D|U_2)\right),\cdots,\left(MP(D|U_m),1-MP(D|U_m)\right)\right)$，那么对于任意一个等价类 $U_i\in U/Q$ 的最大决策熵以及 B 相对于 D 的最大决策熵分别定义为：

$$MH(D|U_i)=-P(U_i)\left(MP(D|U_i)\log MP(D|U_i)+(m-1)\left(\frac{1-MP(D|U_i)}{m-1}\right)\log\left(\frac{1-MP(D|U_i)}{m-1}\right)\right) \tag{9}$$

$$MH(D|B)=\sum_{i=1}^{m}MH(D|U_i) \tag{10}$$

3. 基于最大决策熵的启发式约简

定义 7 [9]决策信息系统 $DS=(U,C\cup D,V,f)$，若 $Q\subseteq C$，$\forall q\in Q$，q 的内部属性重要度定义为：

$$Sig_U^{inner}(q,Q,C,D)=MP_C^U(D|Q-\{q\})-MP_C^U(D|Q) \tag{11}$$

定义 8 [9]决策信息系统 $DS=(U,C\cup D,V,f)$，若 $Q\subseteq C$，$\forall q\in C-Q$，q 的外部属性重要度定义为：

$$Sig_U^{outer}(q,Q,C,D)=MP_C^U(D|Q)-MP_C^U(D|Q\cup\{q\}) \tag{12}$$

定义 9 [9]决策信息系统 $DS=(U,C\cup D,V,f)$，$Q\subseteq C$，$\forall q\in Q$，若 $Sig_U^{inner}(q,Q,C,D)>0$，则 q 为核属性；若 $Sig_U^{inner}(q,Q,C,D)=0$，则 q 为冗余属性。

定义 10 [9]决策信息系统 $DS=(U,C\cup D,V,f)$，若 $Q\subseteq C$ 是 C 的一个约简，当且仅当满足以下两个条件：

1) $MH(D|Q)=MH(D|C)$；

2) 对 $\forall Q'\subset Q$，有 $MH(D|Q')\neq MH(D|C)$。

Table 1. Fast reduction algorithm based on maximum decision entropy (ACC_HA_MDE)
表 1. 基于最大决策熵的快速约简算法

输入：决策系统。

输出：约简结果 Re。

1.初始化，$core=\varnothing$，$Re=\varnothing$；

2.计算 U 在 C 和 D 上的等价类；

3.计算每个属性的内部属性重要度，并求出核；

4.令 $Re=core$，$i=1$，$U_1=U$，$C_1=C$，$C_{del}=\varnothing$；

5.重复：选择属性重要度最大的属性加入 Re，并在该过程中删除冗余属性和对象；

6.去冗余；

7.输出 Re。

4. 基于最大决策熵的加速算法

定理 1 决策信息系统 $DS=(U,C\cup D,V,f)$，$P\subseteq C$，若 $\forall a,b\in C-P-P'$，其中，$C'=C-P'$，$P'=\left\{c\,|\,MH_C^U(D\,|\,P)=MH_C^U\left(D\,|\,(P\cup\{c\})\right),c\in C-P\right\}$，$U'=U-POS_P^C(U,D)$，并且 $Sig_U^{outer}(a,P,C,D)\geq Sig_U^{outer}(b,P,C,D)$，则 $Sig_{U'}^{outer}(a,P,C',D)\geq Sig_{U'}^{outer}(b,P,C',D)$。

证明：若 $U/P=\{U_1,U_2,\cdots,U_p,U_{p+1},\cdots,U_m\}$，$U/D=\{Y_1,Y_2,\cdots,Y_n\}$，其中 $U_p,U_{p+1},\cdots,U_m\subseteq POS_P^C(U,D)$。因此对每个等价类 $U_i\subseteq POS_P^C(U,D)$，存在决策类 Y，使 $U_i\cap Y=U_i$。用 $MH_C^U(D\,|\,P)$ 表示在 U 上的最大决策熵。

$$
\begin{aligned}
MH_C^U(D\,|\,P)&=-\sum_{i=1}^{m}\left(P(U_i)\left(MP(D\,|\,U_i)\log MP(D\,|\,U_i)+(m-1)\left(\frac{1-MP(D\,|\,U_i)}{m-1}\right)\log\left(\frac{1-MP(D\,|\,U_i)}{m-1}\right)\right)\right)\\
&=-\sum_{i=1}^{p}\left(P(U_i)\left(MP(D\,|\,U_i)\log MP(D\,|\,U_i)+(m-1)\left(\frac{1-MP(D\,|\,U_i)}{m-1}\right)\log\left(\frac{1-MP(D\,|\,U_i)}{m-1}\right)\right)\right)\\
&=-\frac{|U'|}{|U|}\sum_{i=1}^{p}\left(P(U_i)\left(MP(D\,|\,U_i)\log MP(D\,|\,U_i)+(m-1)\left(\frac{1-MP(D\,|\,U_i)}{m-1}\right)\log\left(\frac{1-MP(D\,|\,U_i)}{m-1}\right)\right)\right)\\
&=\frac{|U'|}{|U|}MH_C^{U'}(D\,|\,P)
\end{aligned}
$$

由于 $U'=U-POS_P^C(U,D)$，所以 $POS_P^C(U',D)=\varnothing$，又因为 $C'=C-P'$，其中 $P'=\left\{c\,|\,MH_C^U(D\,|\,P)=MH_C^U\left(D\,|\,(P\cup\{c\})\right),c\in C-P\right\}$，$C'\cap P'=\varnothing$，$P\subset P'$，故 $MH_C^U(D\,|\,P)=MH_{C'}^U(D\,|\,P)$。同理可得 $MH_C^{U'}(D\,|\,P)=MH_{C'}^{U'}(D\,|\,P)$。因此 $MH_C^U(D\,|\,P)=(|U'|/|U|)MH_C^{U'}(D\,|\,P)=(|U'|/|U|)MH_{C'}^{U'}(D\,|\,P)$，所以 $Sig_U^{outer}(a,P,C,D)/Sig_{U'}^{outer}(a,P,C',D)=|U'|/|U|$。因此，若 $Sig_U^{outer}(a,P,C,D)\geq Sig_U^{outer}(b,P,C,D)$，则 $Sig_{U'}^{outer}(a,P,C',D)\geq Sig_{U'}^{outer}(b,P,C',D)$。

通过表 1 所示的 ACC_HA_MDE 算法，可以快速计算基于最大决策熵的属性约简。其中 ACC_HA_MDE 在步骤 5 的时间复杂度为 $O\left(a_i\sum_{i=1}^{b}(b_i-i+1)^2\right)$；而在表 2 所示的 HA_MDE 算法在步骤 4 的时间复杂度为 $O(ab^2)$。因此，ACC_HA_MDE 的效率更高。

Table 2. Attribute reduction algorithm based on maximum decision entropy (HA_MDE) [9]
表 2. 基于最大决策熵的属性约简算法[9]

输入：决策系统。

输出：约简结果 Re。

1.初始化， $core=\varnothing$， $Re=\varnothing$；

2.计算 U 在 C 和 D 上的等价类；

3.计算每个属性的内部属性重要度，并求出核；

4.重复：选择属性重要度最大的属性加入 Re；

5.去冗余；

6.输出 Re。

5. 实验分析

实验环境采用 Intel Corei3-9100 (3.6 GHz)处理器、8 GB 内存和 Windows10 操作系统。算法使用 Python 语言进行编写，在开发工具 PyCharm2020 上编译运行。

为了验证提出算法的有效性，本实验选取了 8 组 UCI 数据集，为了更好验证所提出算法的有效性，需要对数据集进行预处理。首先将数据集使用 WEKA3.8.5 进行等频离散化，并将数据集中名词性数据转化为整数表示。对于缺失数据，利用对应属性下占最多比例的属性值进行替换。表 3 展示了每个数据集的相关信息。在实验过程中，将各数据集按对象数目分成 10 份(每份为$\lceil |U|/10 \rceil$)，或将各数据集的属性每份分成$\lceil |C|/10 \rceil$。

Table 3. Data set information
表 3. 数据集信息

	数据集名称	对象数	特征数	类别数
1	Audiology	200	69	24
2	Glass Identification	214	10	7
3	Chronic-Kidney-Disease	400	24	2
4	Connectionist Bench	528	10	11
5	Hill-valley-with-testing	606	100	2
6	Audit-risk	776	26	2
7	QSAR Biodegradation	1055	41	2
8	Kr-vs-kp	3196	36	2

约简效率对比

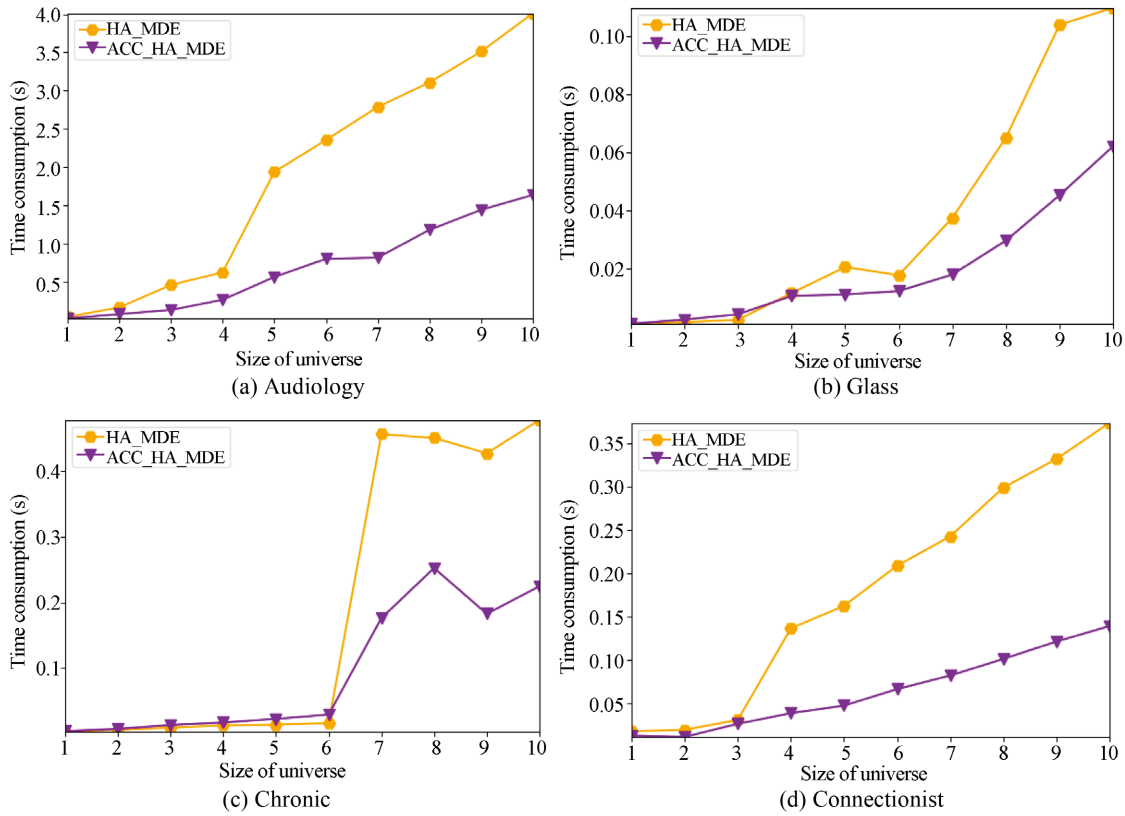
本节对 HA_MDE 与 ACC_HA_MDE 两种算法的约简效率进行比较分析，通过 8 组 UCI 数据集进行实验展示。表 4 展示了 HA_MDE 与 ACC_HA_MDE 两种算法的时间以及约简长度。可以看到 7 个数据集的 ACC_HA_MDE 算法的时间比 HA_MDE 算法的消耗的时间少。例如 Audit-risk 数据集，本文提出的算法 ACC_HA_MDE 耗时 0.32 s，而启发式算法 HA_MDE 运行时间为 2.01 s，HA_MDE 运行时间约为 ACC_HA_MDE 运行时间的 6.281 倍。而 Hill 数据集，本文提出的算法 ACC_HA_MDE 相对于 HA_MDE

耗时较多，由于该数据集在迭代中删除的正域与属性不够多，使消耗时间增多。由于删除的是冗余属性和对象，所以两种算法的约简长度相同。

Table 4. Comparison of time and reduction length of HA_MDE and ACC_HA_MDE algorithms
表 4. HA_MDE 和 ACC_HA_MDE 两种算法的时间与约简长度比较

	数据集名称	HA_MDE		ACC_HA_MDE	
		运行时间/s	约简长度	运行时间/s	约简长度
1	Audiology	4.27	14	**1.80**	14
2	Glass Identification	0.12	6	**0.06**	6
3	Chronic-Kidney-Disease	0.48	7	**0.22**	7
4	Connectionist Bench	0.41	7	**0.15**	7
5	Hill-valley-with-testing	11.02	14	14.27	14
6	Audit-risk	2.01	8	**0.32**	8
7	QSAR Biodegradation	14.52	12	**3.41**	12
8	Kr-vs-kp	9.18	29	**2.55**	29

图 1 中用实心六边形折线表示 HA_MDE、用实心倒三角形折线表示 ACC_HA_MDE，展示了两种算法在 8 组数据集上随论域大小变化的时间消耗曲线，横坐标表示论域大小，纵坐标表示算法运行时间。从图 1 中可以看到除了 Hill 数据集外，其余 7 个数据集在本文提出的算法 ACC_HA_MDE 运行时间相对于 HA_MDE 运行时间较短。因此，本文提出的 ACC_HA_MDE 算法相对于启发式 HA_MDE 算法提高了算法效率。



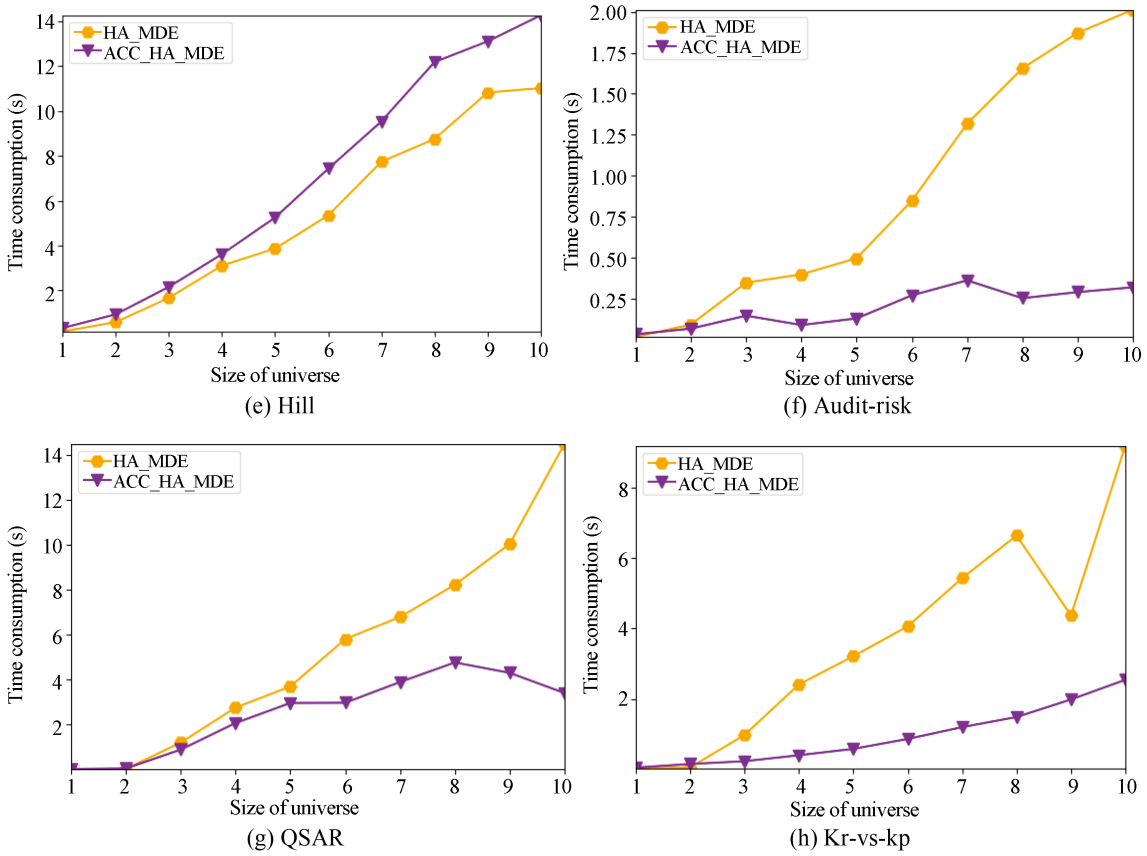


Figure 1. Comparison of algorithm reduction efficiency with object increase
图 1. 随着对象增加算法约简效率的比较

6. 总结

本文针对基于最大决策熵的约简目标提出了在完备决策信息系统下的快速属性约简算法。在每轮迭代中首先删除一部分正域，使数据集中对象数目减少，以提高算法的效率；其次删除冗余属性，可以进一步提高算法的效率。实验选取 8 组 UCI 数据集对提出的算法进行有效性验证，实验结果表明：本文提出算法的效率优于经典算法的效率，实现了对经典算法的优化。

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] 杨习贝, 颜旭, 徐苏平, 于化龙. 基于样本选择的启发式属性约简方法研究[J]. 计算机科学, 2016, 43(1): 40-43.

[3] Chen, H.M., Li, T.R., Cai, Y., Luo, C. and Fujita, H. (2016) Parallel Attribute Reduction in Dominance-Based Neighborhood Rough Set. *Information Sciences*, **373**, 351-368. https://doi.org/10.1016/j.ins.2016.09.012

[4] Wang, C.Z., Shao, M.W., Sun, B.Q. and Hu, Q.H. (2015) An Improved Attribute Reduction Scheme with Covering Based Rough Sets. *Applied Soft Computing*, **26**, 235-243. https://doi.org/10.1016/j.asoc.2014.10.006

[5] Min, F., Zhang, Z.H. and Dong, J. (2018) Ant Colony Optimization with Partial-Complete Searching for Attribute Re-

duction. *Journal of Computational Science*, **25**, 170-182. https://doi.org/10.1016/j.jocs.2017.05.007

[6] Miao, D.Q., Zhao, Y., Yao, Y.Y., Li, H.X. and Xu, F.F. (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

[7] 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

[8] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759-766.

[9] Gao, C., Lai, Z.H., Zhou, J., Zhao, C.R. and Miao, D.Q. (2108) Maximum Decision Entropy-Based Attribute Reduction in Decision-Theoretic Rough Set Model. *Knowledge-Based Systems*, **143**, 179-191. https://doi.org/10.1016/j.knosys.2017.12.014

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

[11] 徐章艳, 刘作鹏, 杨炳儒, 宋威. 一个复杂度为 $max(O(|C||U|),O(|C|^2|U/C|))$ 的快速属性约简算法[J]. 计算机学报, 2006, 29(3): 391-399.

[12] Qian, Y.H., Liang, J.Y., Pedrycz, W. and Dang, C.Y. (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] Du, W.S. and Hu, B.Q. (2018) A Fast Heuristic Attribute Reduction Approach to Ordered Decision Systems. *European Journal of Operational Research*, **264**, 440-452. https://doi.org/10.1016/j.ejor.2017.03.029

[14] Sang, B.B., Chen, H.M., Yang, L., Zhou, D.P., Li, T.R. and Xu, W.H. (2021) Incremental Attribute Reduction Approaches for Ordered Data with Time-Evolving Objects. *Knowledge-Based Systems*, **212**, Article ID: 106583. https://doi.org/10.1016/j.knosys.2020.106583

[15] Dong, L.J. and Chen, D.G. (2020) Incremental Attribute Reduction with Rough Set for Dynamic Datasets with Simultaneously Increasing Samples and Attributes. *International Journal of Machine Learning and Cybernetics*, **11**, 1339-1355. https://doi.org/10.1007/s13042-020-01065-y

[16] Shu, W.H., Qian, W.B. and Xie, Y.H. (2020) Incremental Feature Selection for Dynamic Hybrid Data Using Neighborhood Rough Set. *Knowledge-Based Systems*, **194**, Article ID: 105516. https://doi.org/10.1016/j.knosys.2020.105516

[17] 鲍迪, 张楠, 童向荣, 岳晓东. 区间值决策表的正域增量式属性约简算法[J]. 计算机应用, 2019, 39(8): 2288-2296.