Computer Science and Application 计算机科学与应用, 2021, 11(1), 113-120
Published Online January 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.111013

Hans汉斯

序决策系统下近似约简的启发式算法

孙祖文，唐玉凯

烟台大学计算机与控制工程学院，山东 烟台
Email: sunzuwen444@163.com, iamtyk@126.com

收稿日期：2020年12月25日；录用日期：2021年1月19日；发布日期：2021年1月26日

摘　要

随着网络的进步，社会中产生了大量的高维数据，但很多统计方法难以直接应用到高维数据上。如何获得去噪简化且保存关键信息的低维度数据是一项急需解决的问题。粗糙集理论提供了一种数据降维的方法，被称为属性约简。属性约简的目标是保证原数据集的一种分类特征不变，获得其最小的属性子集。目前，基于不同的分类特征，已提出了许多不同的属性约简算法，如下近似保持、分布保持等。在序决策系统中，传统的下近似约简算法是基于差别矩阵的，计算复杂度高。为了解决这一问题，本文使用依赖度，设计了一个后向贪婪的启发式算法来计算下近似约简。实验使用6组UCI数据集。实验结果表明本文设计的算法可以得到正确的下近似约简，并在时间效率上优于传统的差别矩阵算法。

关键词

属性约简，启发式算法，粗糙集，序决策系统

Heuristic Algorithm to Attribute Reduction for Lower Approximation Preservation in Ordered Decision Systems

Zuwen Sun, Yukai Tang

School of Computer and Control Engineering, Yantai University, Yantai Shandong
Email: sunzuwen444@163.com, iamtyk@126.com

Received: Dec. 25th, 2020; accepted: Jan. 19th, 2021; published: Jan. 26th, 2021

Abstract

With the development of network, a large number of high-dimensional data are generated in the so-

文章引用: 孙祖文, 唐玉凯. 序决策系统下近似约简的启发式算法[J]. 计算机科学与应用, 2021, 11(1): 113-120.
DOI: 10.12677/csa.2021.111013

ciety, but many statistical methods are difficult to be directly applied to high-dimensional data. How to obtain the denoising simplified data with low dimensions and key information is an urgent problem. Rough set theory provides a method for reducing data dimensions, called attribute reduction. The purpose of attribute reduction is to obtain a minimal subset of attributes without changing a classification property of the original data. Until now, different attribute reduction methods are proposed for different classification properties, such as lower approximation preservation and distribution preservation. In an ordered decision system, the traditional attribute reduction algorithm is based on discernibility matrices, and the computation complexity is high. To solve this problem, we design a backward greedy heuristic algorithm to compute a reduct for lower approximation preservation. The experiments are conducted in six UCI data sets. The experimental results show that the algorithm gets a correct reduct and is better than the traditional discernibility matrix algorithm in time efficiency.

Keywords

Attribute Reduction, Heuristic Algorithm, Rough Set Theory, Ordered Decision Systems

1. 引言

粗糙集理论[1]是 Pawlak 提出的处理模糊数据分类的有效方法，已被使用在数据挖掘[2] [3]、数据分析[4]和机器学习[5]等领域。属性约简[6] [7] [8]是粗糙集理论中的一种数据降维方法。属性约简可以保证原数据集的一种分类特征不变，然后获得最小的特征子集。Pawlak 粗糙集使用等价关系来划分样本，但在现实生活中，样本往往需要使用序关系来进行描述，例如学生的成绩、产品的市场占有率等。序决策系统是经典决策系统的一种拓展。在序决策系统中，属性的值域是有序的，在样本间能建立偏序(优势)关系。为了从序决策系统中提取有效信息，Greco 等[9]提出了优势粗糙集方法(Dominance-based rough set approach, DRSA)。

如今，对于 DRSA 的研究已经较为深入。对于序决策系统中不同的分类特征，专家学者们设计了许多不同的属性约简算法。Yuan 等[10]提出了上近似保持约简和广义决策保持约简的算法，并证明了两种算法结果的等价性。Xu 等[11]提出了下近似保持约简算法。Qian 等[12] [13]将 DRSA 应用于集值数据和区间值数据，提出了优势关系保持的约简算法。

以上提及的算法是基于差别矩阵[14]、利用布尔运算法则来进行计算的。尽管这种方法能够得到所有的约简结果，但计算的时间复杂度高。为了提升属性约简的效率，专家学者们在序决策系统中研究设计了启发式算法[15] [16] [17]。启发式算法能够根据不同的分类特征得到一个约简结果，计算效率高。本文基于后向贪婪策略，在序决策系统中提出了下近似约简的启发式算法。实验表明，本文所提的算法能计算出一个正确的下近似约简，且在时间效率上优于 Xu 等[11]提出的差别矩阵算法。

2. 基本概念

给定一个序决策系统 $S=(O, A=M\cup N)$，O 是样本集合，也被称为论域。A 是属性集合，其中 M 是条件属性集合，N 是决策属性集合。对于 $\forall a\in A$，$\forall x\in O$，$a(x)$ 表示样本 x 在某个属性 a 下的取值。表 1 为一个序信息系统，论域 $O=\{x_1,x_2,x_3,x_4\}$，条件属性集 $M=\{a,b,c\}$，决策属性集 $N=\{d\}$。

Table 1. An ordered decision system
表 1. 序决策系统

O	a	b	c	d
x_1	1	1	2	1
x_2	2	2	1	2
x_3	1	3	2	2
x_4	1	4	5	3

定义 1 [9]给定一个序信息系统 $S=(O, A=M\cup N)$，对于 $B\subseteq A$，记

$$R_B^{\geq}=\left\{(x,y)\in O\times O \mid \forall a\in B, a(x)\geq a(y)\right\}.$$

则 $R_B^{\geq}$ 为属性集 B 下的优势关系。$(x,y)\in R_B^{\geq}$ 表示 x 在属性集 B 下优于 y，且 $R_B^{\geq}=\bigcap_{a\in B}R_a^{\geq}$。

优势关系是一个自反、传递、反对称的偏序关系，由优势关系可以导出论域上的一个覆盖。对于 $B\subseteq M$，$[x]_B^{\geq}=\left\{y\in O \mid (y,x)\in R_B^{\geq}\right\}$ 和 $[x]_B^{\leq}=\left\{y\in O \mid (x,y)\in R_B^{\geq}\right\}$ 为样本 x 关于属性集 B 的条件优势类和条件劣势类。对于决策属性集 N，$[x]_N^{\geq}=\left\{y\in O \mid (y,x)\in R_N^{\geq}\right\}$，称 $[x]_N^{\geq}$ 为样本 x 关于优势关系 $R_N^{\geq}$ 的决策优势类，关于 $R_N^{\geq}$ 的决策优势类的全体记为 $O/R_N^{\geq}=\{N_1,N_2,\cdots,N_t\}=\left\{[x]_N^{\geq} \mid x\in O\right\}$。

定义 2 [9]给定一个序决策系统 $S=(O, A=M\cup N)$，对于 $B\subseteq M$，$N_i\in O/R_N^{\geq}$，记

$$\underline{R_B^{\geq}}(N_i)=\left\{x\in O \mid [x]_B^{\geq}\subseteq N_i\right\};$$

$$\overline{R_B^{\geq}}(N_i)=\left\{x\in O \mid [x]_B^{\leq}\cap N_i\neq\varnothing\right\},$$

$\underline{R_B^{\geq}}(N_i)$ 和 $\overline{R_B^{\geq}}(N_i)$ 分别为 N_i 关于属性集 B 的下近似和上近似。

3. 下近似保持约简以及差别矩阵算法

在下近似集合中，每一个样本对应着数据集中的一条确定性规则，保证约简前后下近似不变，即保证了确定性规则不变。Xu 等[11]在序决策系统 S 中定义了下近似约简。

定义 3 [11]给定一个序决策系统 $S=(O, A=M\cup N)$，对于 $B\subseteq M$，$O/R_N^{\geq}=\{N_1,N_2,\cdots,N_t\}$，$B$ 是序决策系统 S 的一个下近似约简需满足以下两点：

1) $\forall N_i\in O/R_N^{\geq}$, $\underline{R_B^{\geq}}(N_i)=\underline{R_M^{\geq}}(N_i)$,

2) $\forall P\subset B$, $\exists N_i\in O/R_N^{\geq}$, $\underline{R_P^{\geq}}(N_i)\neq\underline{R_B^{\geq}}(N_i)$。

条件 1)保证了约简前后 S 的下近似保持不变，条件 2)则保证获得的约简 B 是能够满足 1)的最小属性子集。为了计算下近似约简，Xu 等[11]设计了一个差别矩阵算法，如表 2 所示。

Table 2. A discernibility matrix algorithm for lower approximation preservation (DMALA) [11]
表 2. 基于下近似保持的差别矩阵算法[11]

输入：序决策系统 $S=(O, A=M\cup N)$。
输出：S 的所有下近似约简。
1.在条件属性集 M 下，计算每个样本的条件优势类 $[x]_M^{\geq}$。
2.对 $O/R_N^{\geq}$ 的每一个 N_i 计算下近似 $\underline{R_M^{\geq}}(N_i)$。
3.构造差别矩阵。
4.根据差别矩阵得到差别函数。
5.使用分配律和吸收律对差别函数进行范式转换，将其变为极小析取式。
6.输出极小析取式中的合取项，每一个合取项就是一个下近似约简。

4. 基于依赖度的启发式算法

上一节介绍的差别矩阵算法是基于范式转换来计算约简的。研究已表明，范式的转换是一个 NP-Hard 问题，因此 Xu 等[11]设计的差别矩阵算法计算复杂度高，难以应用在多样本的高维数据集上。为了提升计算效率，本节基于依赖度，提出了一个下近似约简的启发式算法。

定义 4 [18]给定一个序决策系统 $S=(O, A=M\cup N)$，对于 $B\subseteq M$，$O/R_N^{\geq}=\{N_1, N_2, \cdots, N_t\}$，记

$$\gamma_B^{\geq}=\frac{1}{|O|}\sum_{i=1}^{t}\left|\underline{R_B^{\geq}}(N_i)\right|,$$

则 $\gamma_B^{\geq}$ 为属性集 B 关于决策属性 N 的依赖度，其中 $|\cdot|$ 表示集合中元素的个数。

基于依赖度，给出下近似约简的等价定义如下。

定义 5 [18]给定一个序决策系统 $S=(O, A=M\cup N)$，对于 $B\subseteq M$，B 是序决策系统 S 的一个下近似约简需满足以下两点：

1) $\gamma_B^{\geq}=\gamma_M^{\geq}$，

2) $\forall P\subset B, \gamma_P^{\geq}\neq\gamma_B^{\geq}$。

基于定义 5，本文将依赖度作为搜索条件，设计了一个后向贪婪的算法，在每次迭代过程中逐步搜索删除冗余属性。算法具体步骤如表 3 所示。

Table 3. A backward greedy algorithm for lower approximation preservation (BGALA)
表 3. 基于下近似保持的后向贪婪算法

输入：序决策系统 $S=(O, A=M\cup N)$。
输出：S 的下近似约简 B。

1.在条件属性集 M 下，对 $O/R_N^{\geq}$ 的每一个 N_i 计算下近似 $\underline{R_M^{\geq}}(N_i)$。
2.计算 $\gamma_M^{\geq}$。
3.令 $B\leftarrow M$。
4.对 M 中的每个条件属性 a，重复：
　计算 $\gamma_{B-\{a\}}^{\geq}$；
　若 $\gamma_{B-\{a\}}^{\geq}=\gamma_B^{\geq}$，那么 $B=B-\{a\}$。
5.输出 B。

令 m 表示样本个数，n 表示条件属性数目。算法 BGALA 第 1 步和第 2 步是为了计算原始属性集的依赖度，时间复杂度为 $O(m^2n^2)$。第 4 步是一个迭代的过程，在每次迭代中删除一个对依赖度无影响(冗余)的属性，其时间复杂度为 $O(m^3n^2)$。当所有冗余属性被删除，输出约简结果 B。算法的整体时间复杂度是 $O(m^3n^2)$。

5. 实验分析

本节通过实验对比 Xu 等[11]的算法 DMALA 和本文所提算法 BGALA。实验对比主要有两个方面。首先对比两种算法(DMALA, BGALA)的约简结果，验证 BGALA 可以获得和 DMALA 相同的约简。然后对 DMALA 和 BGALA 的运行效率进行对比。本节实验所用的硬件配置为：Windows 10 64 位操作系统，Inter(R) Core(TM) i5-8400 处理器，16G 运行内存。实验使用的编程语言为 Python3.6.5，编译器为 Pycharm Community Edition 2018.3.6。实验使用的 6 组 UCI 数据集如表 4 所示。

Table 4. Data sets description
表 4. 数据集描述

序号	数据集	样本数	属性数
1	Absenteeism at work	740	19

Continued

2	Chronic kidney disease	400	24
3	Dermatology	366	34
4	Flags	194	28
5	Lymphography	148	18
6	Stalog (vehicle)	846	18

5.1. 约简结果对比

表 5 为 DMALA 和 BGALA 两种算法的约简结果对比。属性按照从左到右的顺序使用从 1 开始的正整数来命名。从表中可以看出，DMALA 可以得到多个下近似约简，BGALA 只能得到一个下近似约简。而且，在 DMALA 计算出的所有下近似约简中，有一个约简和 BGALA 计算出的约简是相同的。这个实验结果说明 BGALA 的计算结果是 DMALA 的计算结果之一，是一个正确的下近似约简。

Table 5. The comparison of the reduction results
表 5. 约简结果对比

序号	数据集	DMALA	BGALA
1	Absenteeism at work	{1,2,3,4,5,6,7,11,12,13,14,15,16,17,19} **{1,2,3,5,6,7,11,13,14,15,16,17,18,19}**	**{1,2,3,5,6,7,11,13,14,15,16,17,18,19}**
2	Chronic kidney disease	{1,2,3,7,9,10,13,14,15,17} **{1,3,5,7,8,10,11,14,15,16,18,24}**	**{1,3,5,7,8,10,11,14,15,16,18,24}**
3	Dermatology	**{3,5,6,9,10,12,16,18,19,20,21,23, 24,26,31,33,34}**	**{3,5,6,9,10,12,16,18,19,20,21,23, 24,26,31,33,34}**
4	Flags	**{1,2,3,4,5,6,7,8,12,13,15,16,17,19,21, 22,23,24,25,27,28}** {1,2,3,4,5,6,7,8,11,12,13,15,16,17,18, 19,21,22,24,25,27} {1,2,3,4,5,6,7,8,11,12,13,15,16,17,19, 21,22,23,24,25,27} {1,2,3,4,5,6,7,8,12,13,15,16,17,18,19, 21,22,24,25,27,28}	**{1,2,3,4,5,6,7,8,12,13,15,16,17,19,21, 22,23,24,25,27,28}**
5	Lymphography	{1,2,3,4,5,6,7,8,11,12,13,14,16,17,18} **{2,3,4,5,6,7,8,10,11,12,13,14,16,17,18}**	**{2,3,4,5,6,7,8,10,11,12,13,14,16,17,18}**
6	Stalog(vehicle)	**{1,4,5,6,7,8,9,10,12,14,16,17}** {1,4,5,6,7,8,9,10,11,13,14,16,17}	**{1,4,5,6,7,8,9,10,12,14,16,17}**

5.2. 时间效率对比

图 1~6 展示了 DMALA、BGALA 在 6 组数据集上的效率对比。其中横轴代表属性的个数，纵轴代表运行时间。从图中可以看出，随着属性数的增加，BGALA 的运行时间曲线较为平稳，而 DMALA 的运行时间曲线则波动较大。属性数目较少时，BGALA 的运行时间略长于 DMALA 的运行时间，但当属性数目较大时，BGALA 的运行时间要远远短于 DMALA 的运行时间。这个实验结果表明，在处理高维数据时，BGALA 的时间性能要优于 DMALA。

6. 总结

在序决策系统中，下近似约简的差别矩阵算法时间复杂高，不利于应用在高维数据当中。本文通过引入

依赖度的概念，设计了一个后向贪婪的启发式算法来获得下近似约简。实验证明本文所提的算法能计算出一个正确的约简，并在时间性能上优于传统的差别矩阵算法。下一步将继续研究其它约简目标的启发式算法。

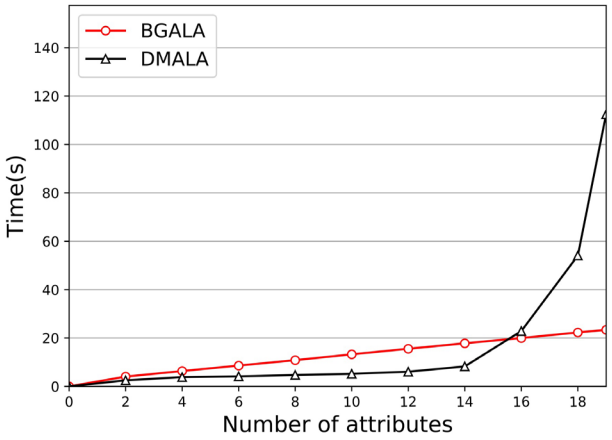

Figure 1. Absenteeism at work
图 1. Absenteeism at work

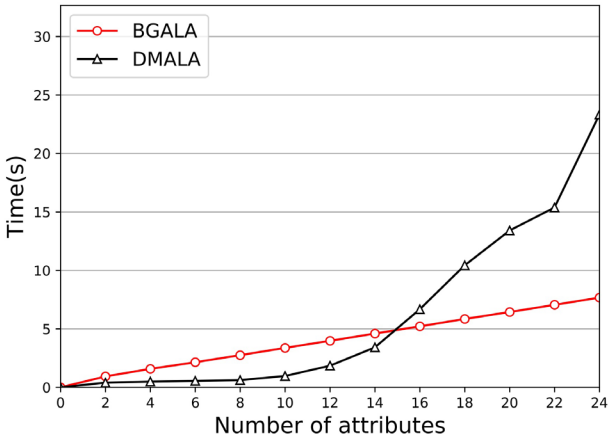

Figure 2. Chronic kidney disease
图 2. Chronic kidney disease

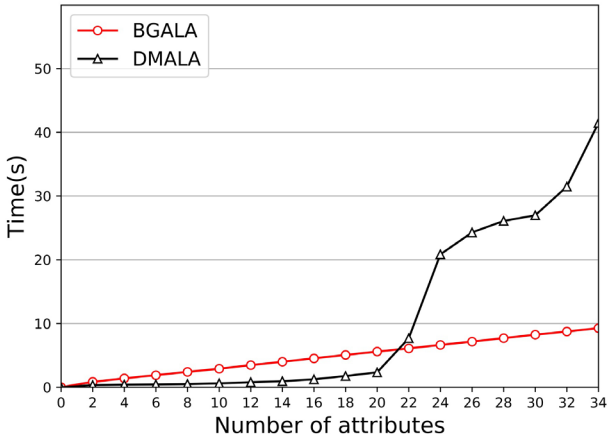

Figure 3. Dermatology
图 3. Dermatology

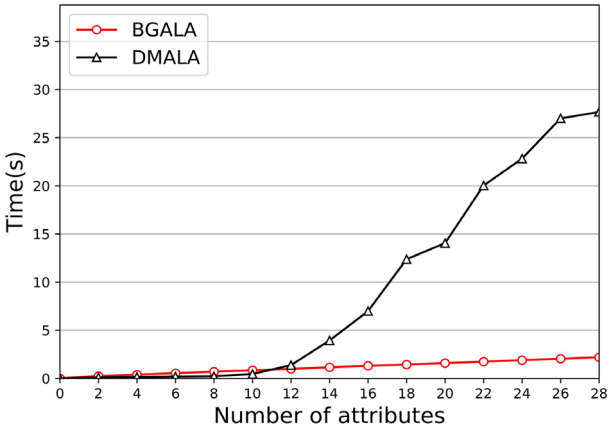

Figure 4. Flags
图 4. Flags

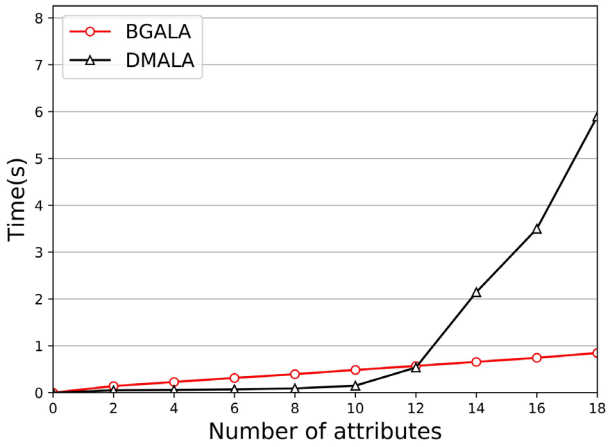

Figure 5. Lymphography
图 5. Lymphography

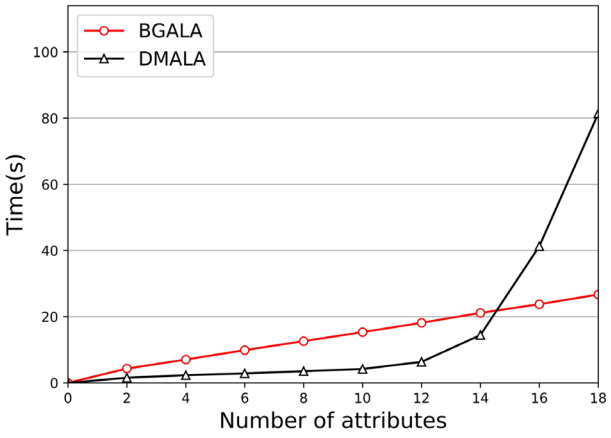

Figure 6. Stalog (vehicle)
图 6. Stalog (vehicle)

基金项目

烟台大学研究生科技创新基金(YDYB2023)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] Zhang, J.B., Li, T.R., Ruan, D. and Liu, D. (2012) Neighborhood Rough Sets for Dynamic Data Mining. *International Journal of Intelligent Systems*, **27**, 317-342. https://doi.org/10.1002/int.21523

[3] Huang, H., Meng, F.Z., Zhou, S.H., Jiang, F. and Manogaran, G. (2019) Brain Image Segmentation Based on FCM Clustering Algorithm and Rough Set. *IEEE Access*, **7**, 12386-12396. https://doi.org/10.1109/ACCESS.2019.2893063

[4] Sai, Y., Yao, Y.Y. and Zhong, N. (2001) Data Analysis and Mining in Ordered Information Tables. 2001 *IEEE International Conference of Data Mining*, San Jose, 29 November-2 December 2001, 497-504. https://doi.org/10.1109/ICDM.2001.989557

[5] Swiniarski, R.W. and Skowron, A. (2003) Rough Set Method in Feature Selection and Recognition. *Pattern Recognition Letters*, **24**, 833-849. https://doi.org/10.1016/S0167-8655(02)00196-4

[6] 唐玉凯, 张楠, 童向荣, 张小峰. 不完备决策系统下的多特定类广义决策约简[J]. 智能系统学报, 2019, 14(6): 1199-1208. http://dx.doi.org/10.11992/tis.201905059

[7] 许鑫. 连续值信息系统的不确定性度量[J]. 计算机科学与应用, 2017, 7(4): 388-397. https://doi.org/10.12677/CSA.2017.74047

[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): 681-684.

[9] Greco, S., Matarazzo, B. and Slowinski, R. (1998) A New Rough Set Approach to Multicriteria and Multiattribute Classification. 1998 *International Conference on Rough Sets and Current Trends in Computing*, Warsaw, 22-26 June 1998, 60-67. https://doi.org/10.1007/3-540-69115-4_9

[10] 袁修久, 何华灿. 优势关系下广义决策约简和上近似约简[J]. 计算机工程与应用, 2006, 42(5): 4-7. http://dx.chinadoi.cn/10.3321/j.issn:1002-8331.2006.05.002

[11] 徐伟华, 张晓燕, 张文修. 优势关系下不协调目标信息系统的下近似约简[J]. 计算机工程与应用, 2009, 45(16): 66-69, 76. http://dx.chinadoi.cn/10.3778/j.issn.1002-8331.2009.16.018

[12] Qian, Y.H., Dang C.Y., Liang, J.Y. and Tang, D.W. (2009) Set-Valued Ordered Information Systems. *Information Sciences*, **179**, 2809-2832. https://doi.org/10.1016/j.ins.2009.04.007

[13] Qian, Y.H., Liang, J.Y. and Dang C.Y. (2008) Interval Ordered Information Systems. *Computers & Mathematics with Applications*, **56**, 1994-2009. https://doi.org/10.1016/j.camwa.2008.04.021

[14] Skowron, A. and Rauszer, C. (1992) The Discernibility Matrices and Functions in Information Systems. In: Słowiński, R., Ed., *Intelligent Decision Support*, Springer, Dordrecht, 331-362. https://doi.org/10.1007/978-94-015-7975-9_21

[15] Hu, Q.H., Yu, D.R. and Guo, M.Z. (2010) Fuzzy Preference Based Rough Sets. *Information Sciences*, **180**, 2003-2022. https://doi.org/10.1016/j.ins.2010.01.015

[16] 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

[17] 赵立威, 张楠, 张中喜. 基于特征粒的序决策系统快速约简研究[J]. 山西大学学报(自然科学版), 2020, 43(4): 897-905.

[18] Du, W.S. and Hu, B.Q. (2014) Approximate Distribution Reducts in Inconsistent Interval-Valued Information Systems. *Information Sciences*, **271**, 93-114. https://doi.org/10.1016/j.ins.2014.02.070