山西大学学报(自然科学版)48(1):158—168,2025
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024138

分布优势数据集的矩阵增量属性约简算法

景运革,吉新如,王鹏岭,王宝丽*
(运城学院 数学与信息技术学院,山西 运城 044000)

摘 要:传统属性约简算法不能有效解决动态数据属性约简问题,寻求高效动态数据属性约简算法是目前人工智能领域研究的热点。本文在动态分布优势数据集中引入矩阵优势条件熵和优势矩阵,探讨基于优势条件熵的矩阵增量属性约简方法。首先,定义了分布数据集的优势矩阵和优势条件熵;其次,通过分析分布数据集添加对象的过程,提出了优势矩阵的增量更新原理和融合机制;然后,给出了基于优势条件熵的矩阵增量约简方法。最后,利用6组UCI(University of California Irvine)优势数据集进行实验,用于验证增量属性约简算法的高效性。实验结果表明:与非增量属性约简算法相比,由增量属性约简算法计算约简的运行时间缩短了85.6%。所以,本文所给出的矩阵增量属性约简算法是求解动态分布优势数据集属性约简的快速有效方法。

关键词:分布优势数据集;增量技术;属性约简;优势条件熵;优势数据集矩阵
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2025)01-0158-11

A Matrix-based Incremental Reduction Approach for Distributed Dominant Data Set

JING Yunge, JI Xinru, WANG Pengling, WANG Baoli*
(*School of Mathematics and Information Technology, Yuncheng University, Yuncheng* 044000*, China*)

Abstract: Traditional attribute reduction algorithms are inefficient to deal with dynamic decision systems, and seeking efficient reduction approach of dynamic data is a research hotspot in the field of artificial intelligence. The paper studies dominant conditional entropy-based incremental attribute reduction approach by introducing dominant conditional entropy and dominant matrix into the dynamic distributed dominant datasets. Firstly, the dominant matrix and dominant conditional entropy of distributed dominant data set are defined. Secondly, an incremental learning mechanism and fusion mechanism of dominant matrix are proposed by analyzing the process of the adding some objects into the distributed dominant data set. Then, a matrix-based incremental attribute reduction approach based on dominant conditional entropy is presented. Finally, experiments on six UCI datasets were conducted to validate the efficiency of the incremental attribute reduction approach. The experimental results showed that the reduction time of incremental attribute reduction approach was reduced by an average of 85.6% compared with the non-incremental reduction approach. Therefore, the proposed matrix-based incremental attribute reduction method is quick and effective in solving reduction of dynamic distributed dominant datasets.

Key words: distributed dominant data set; incremental technology; attribute reduction; dominant conditional entropy; dominant data set matrix

收稿日期:2024-06-14;**接受日期**:2024-10-24
基金项目:国家自然科学基金(61703363);山西省基础研究计划项目(201801D121148);运城学院数据挖掘与工业智能应用科研创新团队(YCXYTD-202402);运城学院科研项目(321701)
作者简介:景运革(1970—),男,山西运城人,博士,教授,研究方向为粒计算。E-mail:jyg701022@163.com
* **通信作者**:王宝丽(WANG Baoli),E-mail:pollycomputer@163.com
引文格式:景运革,吉新如,王鹏岭,等.分布优势数据集的矩阵增量属性约简算法[J].山西大学学报(自然科学版),2025,48(1):158-168. DOI:10.13451/j.sxu.ns.2024138.

0 引言

波兰科学研究者Pawlak提出的粗糙集理论是一种新的数学工具[1],主要用来处理不一致、不精确和不相容的数据问题。该理论在处理上述问题时,不需要任何先验知识,且处理效率较高,所以很多学者已经把它运用到数据挖掘和知识发现等领域。

属性约简是粗糙集理论中的一个重要研究内容,它通过删除决策表中的容冗信息,减少数据特征,提高数据计算效率,是数据预处理的重要工具。一些学者在经典粗糙集模型上提出了一些属性约简算法[2-7],并利用它们对数据进行预处理,取得了一定的效果。但是由于经典粗糙集模型对没有偏序关系的数据集通过上下近似集来逼近目标决策,不能有效处理具有偏序关系的数据集。

为了克服这个缺陷,Greco等提出了优势粗糙集模型[8]来解决具有偏序关系的数据挖掘等问题。通过向上联合或者向下联合来逼近目标决策。为了进一步完善优势粗糙集模型,许多研究者对优势粗糙集模型进行了扩展。梁美社等为了有效处理直觉模糊偏序关系数据规则获取问题,提出了广义优势直觉模糊粗糙集模型[9]。Yang等提出了一种新的优势粗糙集模型用来处理不完备区间值信息系统的问题[10]。上述的工作可以有效处理静态偏序数据挖掘和知识发现问题,但是在处理动态偏序数据挖掘问题时,效率不是很高。

增量学习方法能够有效利用先前的结果,减少重复计算量,极大地提高计算性能,特别适应动态数据环境。很多学者提出了大量增量方法,并利用这些方法解决动态数据近似集的更新[11-14]和属性约简的更新问题[15-19]。在增量属性约简方面:Liang等针对多个对象增加到数据集中,探讨了信息熵的更新机制,提出了群增量属性约简算法[20]。Jing等针对多个对象增加到数据集中,探讨了知识粒度的更新机制,设计了基于知识粒度的属性约简算法[21]。桑彬彬等针对单个对象增加到优势数据集中,分析了优势条件信息熵的更新原理,提出了基于信息熵属性约简方法[22]。Shu等针对不完备信息系统中属性变化的问题,提出了基于正区域的增量属性约简算法[23]。Zeng等针对一些对象增加到混合数据中如何快速更新约简问题,提出了增量属性约简算法[24]。陈宝国等针对不完备有序数据动态变化如何快速更新约简问题,提出了混合有序数据增量约简算法[25]。Sang等针对优势邻域粗糙集数据更新问题,提出了基于优势条件熵的增量属性约简算法[26]。

矩阵运算由于表达直观,计算简单,操作方便,是数据预处理的有效工具。为了解决分布优势数据集动态增加导致的属性约简计算时间长,效率低下的问题,本文将优势矩阵与优势条件熵相结合,提出分布优势数据集的矩阵增量属性约简算法,该算法首先定义了分布数据集的优势矩阵和优势条件熵,并将优势矩阵引入优势条件熵中,然后分析分布数据集添加对象的过程,给出优势矩阵的增量更新原理和融合机制,进而设计分布优势数据集的矩阵增量属性约简算法,最后通过UCI数据集实验验证了分布优势数据集的矩阵增量属性约简算法的高效性。

1 优势粗糙集的基本理论

在本节,我们将介绍优势粗糙集的相关理论和知识[27]。

1.1 优势粗糙集的基础知识

定义1 假设$DS=(U,A,V,f)$为决策数据集必须满足下列条件:

(1)$U=\{u_1,u_2,\cdots,u_n\}$;

(2)$A=C\cup D$;

(3)$V=\bigcup\limits_{a\in C\cup D}V_a$;

(4)$f:U\times A\to V$。

其中U为对象集,$A=C\cup D$为属性集,且C为条件属性集,D为决策属性集,V是属性值,f是信息函数。

定义2 假设$DS=(U,A,V,f)$为决策数据集,对于$\forall a\in A,u_i\in U,u_j\in U$,都满足公式$f(u_j,a)\geqslant f(u_i,a)$,说明在条件属性$a$下,对象$u_j$优于对象$u_i$,则优势关系表示为:$u_j\geqslant_a u_i$,另外,如果$f(u_j,a)\leqslant f(u_i,a)$,则劣势关系表示为:$u_j\leqslant_a u_i$。

定义3 假设$DS=(U,A,V,f)$为决策数据集,$B\subseteq C$对于$\forall a\in A,u_i\in U,u_j\in U$,则$B$的劣势集$[u_i]_B^{\leqslant}$和优势集$[u_i]_B^{\geqslant}$定义如下:

$$\begin{cases}D_B^{\leqslant}=[u_i]_B^{\leqslant}=\{u_j|(u_i,u_j)\in U\times U,u_j\leqslant_a u_i\},\\ D_B^{\geqslant}=[u_i]_B^{\geqslant}=\{u_j|(u_i,u_j)\in U\times U,u_j\geqslant_a u_i\}。\end{cases}\tag{1}$$

定义4 假设$DS=(U,A,V,f)$为决策数据集,对象集U按照D的偏序关系划分的决策类为$Cl=\{cl_1,cl_2,\cdots,cl_n\}$,则$Cl$的向下联合$cl_s^{\leqslant}$和向上联合$cl_s^{\geqslant}$被定义如下:

$$\begin{cases}cl_s^{\leqslant}=\bigcup\limits_{t\leqslant s}cl_t\\ cl_s^{\geqslant}=\bigcup\limits_{t\geqslant s}cl_t\end{cases},\text{其中}\forall t,s\in\{1,2,\cdots,n\}。\tag{2}$$

定义5 假设$DS=(U,A,V,f)$为决策数据集,$B\subseteq C$,对象集U按照B划分为$U/B=\{[u_1],[u_2],\cdots,[u_m]\}$,那么数据集的优势熵被定义如下[28]:

$$DH_B^{\geqslant}(U)=\frac{1}{|U|}\sum_{i=1}^{m}\log\frac{\left|[u_i]_B^{\geqslant}\right|}{|U|}。\tag{3}$$

在上式中,$|\cdot|$代表$\cdot$基数。

定义6 假设$DS=(U,A,V,f)$为决策数据集,$B\subseteq A,C\subseteq A$,条件属性$B$关于决策属性$D$的优势条件熵被定义如下:

$$DH_{D|B}^{\geqslant}(U)=\frac{1}{|U|}\sum_{i=1}^{m}\log\frac{\left|[u_i]_B^{\geqslant}\cap[u_i]_D^{\geqslant}\right|}{\left|[u_i]_B^{\geqslant}\right|}。\tag{4}$$

定义7 假设$DS=(U,A,V,f)$为决策数据集,$B\subseteq A,u_i,u_j\in U$,那么条件属性$B$的优势矩阵$\boldsymbol{M}_B$的元素$m_{ij}$被定义如下:

$$m_{ij}=\begin{cases}1 & f(u_j,B)\geqslant f(u_i,B),\\ 0 & f(u_j,B)<f(u_i,B)。\end{cases}\tag{5}$$

定义8 假设$DS=(U,A,V,f)$为决策数据集,D为决策属性,$B\subseteq A,u_i,u_j\in U$,那么条件属性$B\cup D$的优势矩阵$\boldsymbol{N}_{B\cup D}$的元素$n_{ij}$被定义如下:

$$n_{ij}=\begin{cases}1 & f(u_j,B\cup D)\geqslant f(u_i,B\cup D),\\ 0 & f(u_j,B\cup D)<f(u_i,B\cup D)。\end{cases}\tag{6}$$

1.2 分布优势决策数据集的相关概念

定义9 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DS_i$为分布决策数据集,且$DDS_i=(U_i,A_i,V_i,f_i)$为相互独立的子分布决策数据集,则$DDS$为分布决策数据集,必须满足下列条件:

(1)$U=\bigcup_{i=1}^{m}U_i$;

(2)$U_j\bigcap U_k\neq\varnothing,\forall j,k\in\{1,2,\cdots,m\}$且$j\neq k$;

(3)$A=A_i$;

(4)$V=\bigcup_{i=1}^{m}V_i$;

(5)$f=f_i$。

定义10 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,$DDS_i=(U_i,A_i,V_i,f_i)$和

$DDS_j=(U_j,A_j,V_j,f_j)$为两个相互独立的子分布优势决策数据集,$n_i=|U_i|$,$n_j=|U_i|$,$u_i\in U_i$及$u_j\in U_j$,则DDS_i和DDS_j的优势关系矩阵为$\boldsymbol{M}_{ij}^{C}=\left(m_{ij}^{lk}\right)_{n_i\times n_j}$,其元素被定义如下:

$$m_{ij}^{lk}=\begin{cases}1 & f(u_j,C)\geqslant f(u_i,C)\\ 0 & f(u_j,C)<f(u_i,C)\end{cases},l=(1,2,\cdots,n_i),k=(1,2,\cdots,n_j)。\tag{7}$$

定义11 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,$DDS_i=(U_i,A_i,V_i,f_i)$为相互独立的子分布优势决策数据集,$n=\sum_{i=1}^{m}|U_i|$,则$DDS$的优势关系矩阵为$\left(\boldsymbol{M}_{ij}^{C}\right)_{n\times n}=\left(m_{ij}\right)_{n\times n}$,其元素被定义如下:

$$m_{ji}^{kl}=\begin{cases}1 & f(u_j,C)\geqslant f(u_i,C)\\ 0 & f(u_j,C)<f(u_i,C)\end{cases},l=(1,2,\cdots,n_i),k=(1,2,\cdots,n_j)。\tag{8}$$

根据定义10和定义11可以得到两个相互独立的子分布优势决策数据集的优势关系矩阵,在此基础上,可以得到分布优势决策数据集的优势关系矩阵如定义12和定义13。

定义12 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,且$DDS_i=(U_i,A_i,V_i,f_i)$为相互独立的子分布优势决策数据集,$n=\sum_{i=1}^{m}|U_i|$,则$DDS$的优势关系矩阵为$(\boldsymbol{M}_{ij}^{C})_{n\times n}=(m_{ij})_{n\times n}$,被定义如下:

$$(\boldsymbol{M}_{ij}^{C})_{n\times n}=\begin{bmatrix}M_{11} & M_{12} & \cdots & M_{1m}\\ M_{21} & M_{22} & \cdots & M_{2m}\\ \vdots & \vdots & \ddots & \vdots\\ M_{m1} & M_{m2} & \cdots & M_{mm}\end{bmatrix}。\tag{9}$$

定义13 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,且$DDS_i=(U_i,A_i,V_i,f_i)$为相互独立的子分布优势决策数据集,$n=\sum_{i=1}^{m}|U_i|$,则$DDS$的优势关系矩阵为$(\boldsymbol{M}_{ij}^{C\cup D}$中$)_{n\times n}=(m'_{ij})_{n\times n}$,被定义如下:

$$(\boldsymbol{M}_{ij}^{C\cup D}\text{中})_{n\times n}=\begin{bmatrix}M'_{11} & M'_{12} & \cdots & M'_{1m}\\ M'_{21} & M'_{22} & \cdots & M'_{2m}\\ \vdots & \vdots & \ddots & \vdots\\ M'_{m1} & M'_{m2} & \cdots & M'_{mm}\end{bmatrix}。\tag{10}$$

根据定义6、定义12和定义13可以得到分布优势决策数据集的优势条件熵如定义14。

定义14 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,$DDS$的优势关系矩阵为$(\boldsymbol{M}_{ij}^{C}$中$)_{n\times n}$,$B\subseteq A,C\subseteq A$,则分布优势决策数据集的优势条件熵被定义如下:

$$DH_{D|B}^{\geqslant}(U)=\frac{1}{|U|}\sum_{i=1}^{m}\log\frac{\sum_{j=1}^{n}\mathrm{sum}(m'_{ji})}{\sum_{j=1}^{n}\mathrm{sum}(m_{ji})},\tag{11}$$

其中,sum()为求优势矩阵元素的和。

根据定义14可以得到分布优势决策数据集的内、外部重要度如定义15和定义16。

定义15 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,$B\subseteq C,\forall a\in B$,则$a$的内部重要度被定义如下:

$$Sig_{\bigcup_{i=1}^{m}U_i}^{\mathrm{inter}}(a,C,D)=DH_{D|(B-\{a\})}^{\geqslant}\left(\bigcup_{i=1}^{m}U_i\right)-DH_{D|B}^{\geqslant}\left(\bigcup_{i=1}^{m}U_i\right)。\tag{12}$$

定义16 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,$B\subseteq C,\forall a\in B$,则$a$的

外部重要度被定义如下:

$$Sig^{\text{outer}}_{\bigcup_{i=1}^{m}U_i}(a,B,D)=DH^{\geqslant}_{D|B}\left(\bigcup_{i=1}^{m}U_i\right)-DH^{\geqslant}_{D|(B\cup\{a\})}\left(\bigcup_{i=1}^{m}U_i\right)。\tag{13}$$

定义 17 假设 $DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$ 为分布优势决策数据集,$B\subseteq C$,B 是 DDS 的约简必须满足下列条件:

(1) $DH^{\geqslant}_{D|B}\left(\bigcup_{i=1}^{m}U_i\right)=DH^{\geqslant}_{D|C}\left(\bigcup_{i=1}^{m}U_i\right)$,

(2) $\forall a\in B, DH^{\geqslant}_{D|(B-\{a\})}\left(\bigcup_{i=1}^{m}U_i\right)\neq DH^{\geqslant}_{D|C}\left(\bigcup_{i=1}^{m}U_i\right)$。

1.3 分布优势数据集的属性约简算法

根据上述优势粗糙集及分布优势数据集的基本理论和知识,提出了分布优势数据集的属性约简算法,算法详细描述过程如算法1所述。

算法 1 分布优势数据集的属性约简算法

输入:分布决策数据集 $DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$。

输出:分布决策数据集 DDS 的约简 $RED_{\bigcup_{i=1}^{m}U_i}$。

Let $RED_{\bigcup_{i=1}^{m}U_i}\leftarrow\varphi$;

For $1\leqslant k\leqslant|U|$ do

Compute $Sig^{\text{inter}}_{\bigcup_{i=1}^{m}U_i}(a_k,C,D)$

If $Sig^{\text{inter}}_{\bigcup_{i=1}^{m}U_i}(a_k,C,D)>0$

Let $RED_{\bigcup_{i=1}^{m}U_i}\leftarrow RED_{\bigcup_{i=1}^{m}U_i}\cup a_k$

End if

End for

Let $B\leftarrow RED_{\bigcup_{i=1}^{m}U_i}$

Compute $DH^{\geqslant}_{D|B}\left(\bigcup_{i=1}^{m}U_i\right)$ and $DH^{\geqslant}_{D|C}\left(\bigcup_{i=1}^{m}U_i\right)$

While a_k

For a_k in $(C-B)$ do

Compute $Sig^{\text{outer}}_{\bigcup_{i=1}^{m}U_i}(a_k,C,D)$

$a=\max(Sig^{\text{outer}}_{\bigcup_{i=1}^{m}U_i}(a_k,C,D))$ // $\max(\bullet)$是计算最大值的函数求

$B\leftarrow B\cup a$

End for

End while

$RED_{\bigcup_{i=1}^{m}U_i}\leftarrow B$

Print $RED_{\bigcup_{i=1}^{m}U_i}$。

2 分布优势数据集的增量属性约简算法

在本节,当一些对象增加到分布优势数据集中,我们将讨论分布优势数据集的优势关系矩阵更新机制和原理,设计分布优势数据集的增量属性约简算法。

2.1 优势矩阵增量更新机制

定义 18 假设 $DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$ 为分布优势决策数据集,新增分布优势决策数据集 $DDS_X=(U_X,A,V,f)$。假设对象集 $U_X=\{u_{n+1},u_{n+2},\cdots,u_{n+t}\}$ 被增加到分布优势决策数据集中,则 U_x 的在条件属性 C 下的优势关系矩阵 $(\boldsymbol{Q}^{C}_{U_X})_{t\times t}=(q_{ij})_{t\times t}$ 定义如下:

$$q_{ij}=\begin{cases}1 & f(u_j,C)\geqslant f(u_i,C)\\ 0 & f(u_j,C)<f(u_i,C)\end{cases},1\leqslant i,j\leqslant t。\tag{14}$$

定义19 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,新增分布优势决策数据集$DDS_X=(U_X,A,V,f)$。假设对象集$U_X=\{u_{n+1},u_{n+2},\cdots,u_{n+t}\}$被增加到分布优势决策数据集中,则$U_x$的在属性$C\cup D$下的优势关系矩阵$(\boldsymbol{Q}'^{C\cup D}_{U_x})_{t\times t}=(q'_{ij})_{t\times t}$定义如下:

$$q'_{ij}=\begin{cases}1 & f(u_j,C\cup D)\geqslant f(u_i,C\cup D)\\ 0 & f(u_j,C\cup D)<f(u_i,C\cup D)\end{cases},1\leqslant i,j\leqslant t。\tag{15}$$

定义20 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$分布优势决策数据集,新增分布优势决策数据集$DDS_X=(U_X,A,V,f)$。假设对象集$U_X=\{u_{n+1},u_{n+2},\cdots,u_{n+t}\}$被增加到子分布优势数据集$DDS_i=(U_i,A_i,V_i,f_i)$中,$n_i=|U_i|$,则$DS$和$DDS_i$在条件属性$C$下的优势关系矩阵$\boldsymbol{W}^{C}_{ij}=(w^{tk}_{ij})_{t\times n_i}$定义如下:

$$w_{ij}=\begin{cases}1 & f(u_{n+j},C)\geqslant f(u_i,C)\\ 0 & f(u_{n+j},C)<f(u_i,C)\end{cases},1\leqslant j\leqslant t,1\leqslant i\leqslant n_i。\tag{16}$$

定义21 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,新增分布优势决策数据集$DDS_X=(U_X,A,V,f)$。假设对象集$U_X=\{u_{n+1},u_{n+2},\cdots,u_{n+t}\}$被增加到子分布优势数据集$DDS_i=(U_i,A_i,V_i,f_i)$中,$n_i=|U_i|$,则$DS$和$DDS_i$在属性$C\cup D$下的优势关系矩阵$\boldsymbol{W}'^{C\cup D}_{ij}=(w'^{tk}_{ij})_{t\times n_i}$定义如下:

$$w'_{ij}=\begin{cases}1 & f(u_{n+j},C\cup D)\geqslant f(u_i,C\cup D)\\ 0 & f(u_{n+j},C\cup D)<f(u_i,C\cup D)\end{cases}1\leqslant j\leqslant t,1\leqslant i\leqslant n_i。\tag{17}$$

定义22 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,新增分布优势决策数据集$DDS_X=(U_X,A,V,f)$。假设对象集$U_X=\{u_{n+1},u_{n+2},\cdots,u_{n+t}\}$被增加到子分布优势数据集$DDS_i=(U_i,A_i,V_i,f_i)$中,$n_i=|U_i|$,则$DDS_i$和$DS$在条件属性$C$下的优势关系矩阵$\boldsymbol{H}^{C}_{ij}=(h^{kt}_{ij})_{n_i\times t}$定义如下:

$$h_{ij}=\begin{cases}1 & f(u_i,C)\geqslant f(u_{n+j},C)\\ 0 & f(u_i,C)<f(u_{n+j},C)\end{cases},1\leqslant j\leqslant t,1\leqslant i\leqslant n_i。\tag{18}$$

定义23 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,新增分布优势决策数据集$DDS_X=(U_X,A,V,f)$。假设对象集$U_X=\{u_{n+1},u_{n+2},\cdots,u_{n+t}\}$被增加到子分布优势数据集$DDS_i=(U_i,A_i,V_i,f_i)$中,$n_i=|U_i|$,则$DDS_i$和$DS$在属性$C\cup D$下的优势关系矩阵$\boldsymbol{H}'^{C\cup D}_{ij}=(h'^{kt}_{ij})_{n_i\times t}$定义如下:

$$h'_{ij}=\begin{cases}1 & f(u_i,C\cup D)\geqslant f(u_{n+j},C\cup D)\\ 0 & f(u_i,C\cup D)<f(u_{n+j},C\cup D)\end{cases},1\leqslant j\leqslant t,1\leqslant i\leqslant n_i。\tag{19}$$

根据定义18—定义23,可以得到增加对象后的分布决策数据集的优势关系矩阵如定理1和定理2。

定理1 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,新增分布优势决策数据集$DDS_X=(U_X,A,V,f)$。假设对象集$U_X=\{u_{n+1},u_{n+2},\cdots,u_{n+t}\}$被增加到子分布优势数据集$DDS_i=(U_i,A_i,V_i,f_i)$中,$n_i=|U_i|$,则增加对象集后的分布优势数据集在条件属性$C$下的优势关系矩阵$(\boldsymbol{Z}^{C}_{ij})_{(n+t)\times(n+t)}=(z^{kt}_{ij})_{(n+t)\times(n+t)}$为:

$$(\boldsymbol{Z}^{C}_{ij})_{(n+t)\times(n+t)}=\begin{bmatrix}M_{11} & M_{12} & \cdots & M_{1i} & H_{1i} & \cdots & M_{1m}\\ M_{21} & M_{22} & \cdots & M_{2i} & H_{2i} & \cdots & M_{2m}\\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ M_{m1} & M_{m2} & \cdots & M_{mi} & H_{mi} & \cdots & M_{mm}\\ W_{11} & W_{12} & \cdots & W_{1i} & Q_{1i} & \cdots & W_{mi}\end{bmatrix}。\tag{20}$$

定理2 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,新增分布优势决策数据集$DDS_X=(U_X,A,V,f)$。假设对象集$U_X=\{u_{n+1},u_{n+2},\cdots,u_{n+t}\}$被增加到子分布优势数据集$DDS_i=(U_i,A_i,V_i,f_i)$中,$n_i=|U_i|$,则增加对象集后的分布优势数据集在属性$C\cup D$下的优势关系矩阵$(Z'^{C\cup D}_{ij})_{(n+t)\times(n+t)}=(z'^{kt}_{ij})_{(n+t)\times(n+t)}$为:

$$(Z'^{C\cup D}_{ij})_{(n+t)\times(n+t)}=\begin{bmatrix} M'_{11} & M'_{12} & \cdots & M'_{1i} & H'_{1i} & \cdots & M'_{1m} \\ M'_{21} & M'_{22} & \cdots & M'_{2i} & H'_{2i} & \cdots & M'_{2m} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ M'_{m1} & M'_{m2} & \cdots & M'_{mi} & H'_{mi} & \cdots & M'_{mm} \\ W'_{11} & W'_{12} & \cdots & W'_{1i} & Q'_{1i} & \cdots & W'_{mi} \end{bmatrix}。 \tag{21}$$

根据定理1—定理2,可以得到增加对象后的分布决策数据集的优势条件熵如定理3。

定理3 假设$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$为分布优势决策数据集,新增分布优势决策数据集$DDS_X=(U_X,A,V,f)$,$B\subseteq A,C\subseteq A$。假设对象集$U_X=\{u_{n+1},u_{n+2},\cdots,u_{n+t}\}$被增加到子分布优势数据集$DDS_i=(U_i,A_i,V_i,f_i)$中。则分布优势决策数据集$DDS$增加对象集$U_X$后的优势条件熵为:

$$DH^{\geqslant}_{D|B}\left(\bigcup_{i=1}^{m}U_i\cup U_X\right)=\frac{1}{|U|}\sum_{i=1}^{m}\log\frac{\sum_{j=1}^{n}\mathrm{sum}(z'_{ji})}{\sum_{j=1}^{n}\mathrm{sum}(z_{ji})}。 \tag{22}$$

2.2 分布优势数据集的矩阵增量约简算法

当一些对象被增加到分布优势数据集后,利用算法1所计算的分布优势数据的约简$RED_{\bigcup_{i=1}^{m}U_i}$及优势关系矩阵,并根据上述的增量更新机制,本文提出了分布优势数据集的增量属性约简算法,算法的具体过程如下所述。

算法2 分布优势数据集的增量属性约简算法

输入:分布决策数据集$DDS=(U,A,V,f)=\bigcup_{i=1}^{m}DDS_i$,$RED_{\bigcup_{i=1}^{m}U_i}$,增加对象集$U_X$。

输出:分布决策数据集DDS增加了对象集U_X后的约简$RED_{\bigcup_{i=1}^{m}U_i\cup U_X}$。

Let $B\leftarrow RED_{\bigcup_{i=1}^{m}U_i}$

Compute $DH^{\geqslant}_{D|RED_U}(\bigcup_{i=1}^{m}U_i\cup U_X)$ and $DH^{\geqslant}_{D|C}(\bigcup_{i=1}^{m}U_i\cup U_X)$

While $DH^{\geqslant}_{D|B}(\bigcup_{i=1}^{m}U_i\cup U_X)\neq DH^{\geqslant}_{D|C}(\bigcup_{i=1}^{m}U_i\cup U_X)$

For a_k in $(C-B)$ do

Compute $Sig^{\mathrm{outer}}_{\bigcup_{i=1}^{m}U_i\cup U_X}(a_k,C,D)$

$a=\max(Sig^{\mathrm{outer}}_{\bigcup_{i=1}^{m}U_i\cup U_X}(a_k,C,D))$ // $\max(\bullet)$是计算最大值的函数求:

$B\leftarrow B\cup a$

End for

End while

$RED_{\bigcup_{i=1}^{m}U_i\cup U_X}\leftarrow B$

Print $RED_{\bigcup_{i=1}^{m}U_i\cup U_X}$

3 仿真实验结果分析

在本节,为了验证分布优势数据集的增量属性约简算法的有效性,主要从分布优势数据集的选取、计算约简的运行时间和约简分类的精确度,及仿真实验的软硬件环境的选取及实验方案的设计等方面来对算法1(非增量算法)和算法2(增量算法)进行测试。

3.1 仿真实验数据集和实验环境

在实验过程中,本节将从机器学习网站上选取6个UCI优势数据集作为实验的数据,6个UCI优势数据集的详细情况叙述如表1所示,其中3个数据集属性值的类型是整型数据,1个数据集属性值的类型是浮点型数据,2个数据集属性值的类型是整型数据和字符串数据混合组成,由于个别数据集在实验中不能直接处理,所以需要对上述数据集进行预处理。对于数据集中存在少量缺失值问题,我们在实验中直接删除缺失值;对于数据集中条件属性和决策属性存在字符串的值,我们按照字符串的优劣赋予大小不同的数值,相同的字符串赋予相同的数值。另外,本文编写代码对预处理过的数据集按照决策属性类型对相对应的数据进行比较分析,实验结果验证了它们是具有优势关系的数据集。仿真实验硬件环境为:Intel(R)Core(TM)i5-5200U CPU(Central Processing Unit)@ 2.2 GHz,内存2.0 GB。仿真实验软件环境为:Eclipse 3.7,操作系统是Win10 64位。

表1 6组实验数据的基本信息

Table 1 The basic information of 6 experimental datasets

优势数据集	对象集/个	条件属性/个	决策属性/个
Spectf	267	45	3
Dermatology	366	34	6
Car	1 728	6	4
Wine	178	13	3
BCW	699	9	2
Postoperative	90	8	3

3.2 属性约简结果比较

为了验证算法2计算性能的有效性,本文做了大量仿真实验。为了实验操作方便,本文假设分布优势数据集由3个子优势数据集组成。实验过程如下:本文把表1中的每个优势数据集中的对象集按50%,50%比例分成两个优势数据集,把其中一个优势数据集分成3个子优势数据集,看成分布优势数据集,然后把另一个优势数据集添加到任意1个子优势数据集中,然后利用算法2和算法1去计算分布优势数据集的属性约简和运行时间。为了让运行时间具有一定的可靠性和稳定性,所以本文把5次计算的时间取平均值作为计算约简的运行时间。属性约简结果和运行时间的结果如表2所示。

在表2中,分布优势数据集的算法1与算法2所获得约简的结果基本上是相同的,但是算法1的运行时间远远大于算法2的运行时间,说明了分布优势数据集的增量属性约简算法在处理动态分布优势数据集的约简问题上具有较强的计算性能。

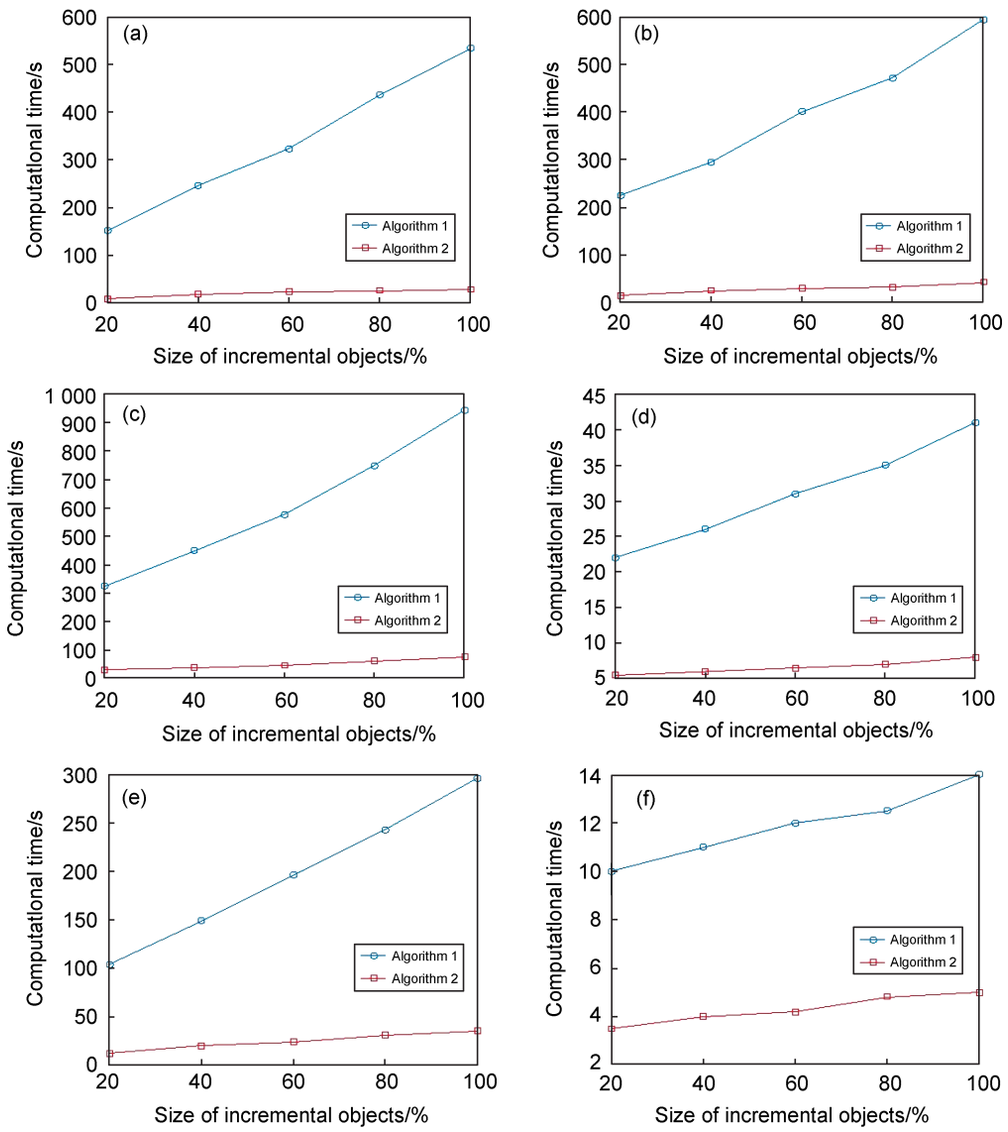
表2 算法1与算法2在6组数据集上的属性约简结果和运行时间对比

Table 2 Comparison of reduction results and running time between algorithm 1 and algorithm 2 on 6 datasets

数据集	算法1		算法2	
	约简结果	时间/s	约简结果	时间/s
Spectf	1,2,3,4,5,6,8,10,12,13,14,15,16,19,20,21,22,23,24,25,26,28,29,30,31,32,33,35,36.34,38,39,40,41,42,43,44	534	1,2,3,4,5,6,8,10,12,13,14,15,16,19,20,21,22,23,24,25,26,28,29,30,31,32,33,35,36.34,38,39,40,41,42,43,44	27
Dermatology	1,2,3,4,5,6,8,9,10,11,12,13,14,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,32,33,34	694	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,17,19,20,21,22,23,24,25,26,27,28,29,32,33,34	42
Car	1,3,4,5,6	944	2,3,4,5,6	77
Wine	1,2,3,5,6,7,8,9,10,11,12,13	41	1,2,3,5,6,7,8,9,10,11,12,13	8
BCW	1,2,3,4,5,8,9	296	1,2,3,4,5,6,9	35
Postoperative	1,2,3,4,5,7,8	14	1,2,3,4,5,7,8	5

3.3 运行时间比较

为了验证算法2在运行时间上优于算法1,本文做了大量仿真实验。实验具体过程如下:在3.2节分布优势数据集的基础上,把另一优势数据集的对象集按20%,40%,60%,80%,100%比例分成5份,然后依次把每一份数据增加到任意1个子优势数据集中。实验结果如图1所示。



注:6组数据集:(a) Spectf; (b) Dermatology; (c) Car; (d) Wine ; (e) BCW; (f) Postoperative。每个子图纵轴表示运行时间(单位为 s),横轴表示对象增加的百分比(单位为 %),算法1的运行时间用蓝色折线表示,算法2的运行时间用红色折线表示。

图1 算法1和算法2在6组数据集上增加了不同比例数据集下的运行时间比较图

Fig. 1 Comparisons of execution time between algorithm 1 and algorithm 2 under different data sets when adding different ratio objects on 6 data sets

在图1中,随着增加对象数量的增加,算法1和算法2计算约简的时间也会增加,算法2运行时间增长幅度较小,而算法1运行时间增长幅度更大,说明算法2(分布优势数据的增量属性约简算法)能极大提高动态分布优势数据集属性约简的效率。

3.4 仿真实验分类精确度比较

通过大量实验验证算法2和算法1所计算约简的有效性,本文做了大量仿真实验。实验具体过程如下:利用3.2节获得分布优势数据集的属性约简,然后通过10折交叉验证法及贝叶斯分类法来计算算法1和算法2所获得约简的分类精确度。实验结果如表3所示。

表3 算法1与算法2在6组数据集上的分类精确度对比

Table 3 Comparison of classification accuracy between algorithm 1 and algorithm 2 on 6 datasets

数据集	算法1/%	算法2/%
Spectf	96.86	96.86
Dermatology	94.25	94.45
Car	92.55	92.95
Wine	94.98	94.98
BCW	95.12	95.32
Postoperative	72.45	72.45

在表3中,算法1和算法2所获得约简的分类精确度基本上是相同的,说明分布优势数据集的增量属性约简算法可以有效处理分布优势数据集添加多个对象时动态更新属性约简的问题。

4 结束语

本文以动态分布优势数据集为研究对象,利用分布优势条件信息熵为属性约简的度量单位,探讨了分布优势数据集优势关系矩阵的增量更新原理,提出了分布优势数据集的增量属性约简方法。最后采用UCI数据集进行了大量实验并对测试结果进行比较分析,分析结果验证了分布优势数据集的增量属性约简算法能够极大地提高动态分布优势数据集属性约简的效率。今后将继续研究分布优势数据集属性和对象变化的增量属性约简算法。

参考文献:

[1] PAWLAK Z. Rough Sets[J]. *Int J Comput Inf Sci*, 1982, **11**(5): 341−356. DOI: 10.1007/bf01001956.

[2] 徐岩柏, 景运革. 多源数据矩阵增量约简算法[J]. 计算机工程与应用, 2022, **58**(3): 195−200. DOI: 10.3778/j.issn.1002-8331.2008-0188.
XU Y B, JING Y G. Matrix-based Incremental Reduction Approach of Multi-resource Data[J]. *Comput Eng Appl*, 2022, **58**(3): 195−200. DOI: 10.3778/j. issn. 1002-8331.2008-0188.

[3] LIANG J Y, WANG F, DANG C Y, *et al*. An Efficient Rough Feature Selection Algorithm with a Multi-granulation View[J]. *Int J Approx Reason*, 2012, **53**(6): 912−926. DOI: 10.1016/j.ijar.2012.02.004.

[4] WANG C Z, WANG Y, SHAO M W, *et al*. Fuzzy Rough Attribute Reduction for Categorical Data[J]. *IEEE Trans Fuzzy Syst*, 2020, **28**(5): 818−830. DOI: 10.1109/TFUZZ.2019.2949765.

[5] SANG B B, XU W H, CHEN H M, *et al*. Active Anti-noise Fuzzy Dominance Rough Feature Selection Using Adaptive K-nearest Neighbors[J]. *IEEE Trans Fuzzy Syst*, 2023, **31**(11): 3944−3958. DOI: 10.1109/TFUZZ.2023.3272316.

[6] SANG B B, YANG L, CHEN H M, *et al*. Fuzzy Rough Feature Selection Using a Robust Non-linear Vague Quantifier for Ordinal Classification[J]. *Expert Syst Appl*, 2023, **230**: 120480. DOI: 10.1016/j.eswa.2023.120480.

[7] 胡凯欣, 马建敏, 刘权芳. 对象导出三支概念格的矩阵粗糙熵约简[J]. 山西大学学报(自然科学版), 2024. DOI: 10.13451/j.sxu.ns.2024028.
HU K X, MA J M, LIU Q F. Matrix Rough Entropy-based Reductions of Object-Induced Three-Way Concept Lattices[J]. *J Shanxi Univ Nat Sci Ed*, 2024. DOI: 10.13451/j.sxu.ns.2024028.

[8] GRECO S, MATARAZZO B, SLOWINSKI R. Rough Approximation of a Preference Relation by Dominance Relations[J]. *Eur J Oper Res*, 1999, **117**(1): 63−83. DOI: 10.1016/s0377-2217(98)00127-1.

[9] 梁美社, 米据生, 赵天娜. 广义优势多粒度直觉模糊粗糙集及规则获取[J]. 智能系统学报, 2017, **12**(6): 883−888. DOI: 10.11992/tis.201706034.
LIANG M S, MI J S, ZHAO T N. Generalized Dominance-based Multi-granularity Intuitionistic Fuzzy Rough Set and Acquisition of Decision Rules[J]. *CAAI Trans Intell Syst*, 2017, **12**(6): 883−888. DOI: 10.11992/tis.201706034.

[10] YANG X B, YU D J, YANG J Y, *et al*. Dominance−based Rough Set Approach to Incomplete Interval−valued Information System[J]. *Data Knowl Eng*, 2009, **68**(11): 1331−1347. DOI: 10.1016/j.datak.2009.07.007.

[11] LUO C, LI T R, CHEN H M, *et al*. Incremental Approaches for Updating Approximations in Set-valued Ordered Information Systems[J]. *Knowl Based Syst*,

2013, **50**: 218−233. DOI: 10.1016/j.knosys.2013.06.013.

[12] LI S Y, LI T R, LIU D. Incremental Updating Approximations in Dominance-based Rough Sets Approach under the Variation of the Attribute Set[J]. *Knowl Based Syst*, 2013, **40**: 17−26. DOI: 10.1016/j.knosys.2012.11.002.

[13] JING Y G, LI T R, HUANG J F, *et al*. An Incremental Attribute Reduction Approach Based on Knowledge Granularity under the Attribute Generalization[J]. *Int J Approx Reason*, 2016, **76**: 80−95. DOI: 10.1016/j.ijar.2016.05.001.

[14] 贺青青, 马建敏, 丁娜. 形式背景上基于矩阵信息熵的矩阵熵约简[J]. 南京大学学报(自然科学), 2023, **59**(1): 98−106. DOI: 10.13232/j.cnki.jnju.2023.01.010.
HE Q Q, MA J M, DING N. Matrix Information Entropy Based Matrix Entropy Reduction in Formal Contexts [J]. *J Nanjing Univ Nat Sci*, 2023, **59**(1): 98−106. DOI: 10.13232/j.cnki.jnju.2023.01.010.

[15] QIAN J, MIAO D Q, ZHANG Z H, *et al*. Hybrid Approaches to Attribute Reduction Based on Indiscernibility and Discernibility Relation[J]. *Int J Approx Reason*, 2011, **52**(2): 212−230. DOI: 10.1016/j.ijar.2010.07.011.

[16] WANG S, LI T R, LUO C, *et al*. A Novel Approach for Efficient Updating Approximations in Dynamic Ordered Information Systems[J]. *Inf Sci*, 2020, **507**: 197−219. DOI: 10.1016/j.ins.2019.08.046.

[17] JING Y G, LI T R, FUJITA H, *et al*. An Incremental Attribute Reduction Method for Dynamic Data Mining[J]. *Inf Sci*, 2018, **465**: 202−218. DOI: 10.1016/j.ins.2018.07.001.

[18] 张义宗, 王磊, 徐阳. 属性集变化下序决策信息系统的增量属性约简算法[J]. 南京大学学报(自然科学), 2023, **59**(5): 813−822. DOI: 10.13232/j. cnki. jnju.2023.05.009.
ZHANG Y Z, WANG L, XU Y. Incremental Attribute Reduction Algorithm for Ordered Decision Information Systems with the Change of Attribute Set[J]. *J Nanjing Univ Nat Sci*, 2023, **59**(5): 813−822. DOI: 10.13232/j.cnki.jnju.2023.05.009.

[19] JING Y G, LI T R, FUJITA H, *et al*. An Incremental Attribute Reduction Approach Based on Knowledge Granularity with a Multi-granulation View[J]. *Inf Sci*, 2017, **411**: 23−38. DOI: 10.1016/j.ins.2017.05.003.

[20] LIANG J Y, WANG F, DANG C Y, *et al*. A Group Incremental Approach to Feature Selection Applying Rough Set Technique[J]. *IEEE Trans Knowl Data Eng*, 2014, **26**(2): 294−308. DOI: 10.1109/TKDE.2012.146.

[21] JING Y G, LI T R, LUO C, *et al*. An Incremental Approach for Attribute Reduction Based on Knowledge Granularity[J]. *Knowl Based Syst*, 2016, **104**: 24−38. DOI: 10.1016/j.knosys.2016.04.007.

[22] 桑彬彬, 陈留中, 陈红梅, 等. 优势关系粗糙集增量属性约简算法[J]. 计算机科学, 2020, **47**(8): 137−143. DOI: 10.11896/jsjkx.190700188.
SANG B B, CHEN L Z, CHEN H M, *et al*. Matrix-based Approach for Calculating Knowledge Granulation and Its Application in Attribute Reduction[J]. *Comput Eng Sci*, 2020, **47**(8): 137−143. DOI: 10.11896/jsjkx.190700188.

[23] SHU W H, SHEN H. Updating Attribute Reduction in Incomplete Decision Systems with the Variation of Attribute Set[J]. *Int J Approx Reason*, 2014, **55**(3): 867−884. DOI: 10.1016/j.ijar.2013.09.015.

[24] ZENG A P, LI T R, LIU D, *et al*. A Fuzzy Rough Set Approach for Incremental Feature Selection on Hybrid Information Systems[J]. *Fuzzy Sets Syst*, 2015, **258**: 39−60. DOI: 10.1016/j.fss.2014.08.014.

[25] 陈宝国, 陈磊, 邓明, 等. 基于不完备混合序信息系统的增量式属性约简[J]. 工程科学与技术, 2024, **56**(1): 65−81. DOI: 10.15961/j.jsuese.202201214.
CHEN B G, CHEN L, DENG M, *et al*. Incremental Attribute Reduction Algorithm Based on Incomplete Hybrid Order Information System[J]. *Adv Eng Sci*, 2024, **56**(1): 65−81. DOI: 10.15961/j.jsuese.202201214.

[26] SANG B B, CHEN H M, YANG L, *et al*. Incremental Feature Selection Using a Conditional Entropy Based on Fuzzy Dominance Neighborhood Rough Sets[J]. *IEEE T Fuzzy Syst*. 2022, **30**(6): 1683−1697. DOI: 10.1109/TFUZZ.2021.3064686.

[27] 刘清. Rough集及Rough推理[M]. 北京: 科学出版社, 2001.
LIU Q. Rough Set and Rough Reasoning[M]. Beijing: Science Press, 2001.

[28] HU Q H, GUO M Z, YU D R, *et al*. Information Entropy for Ordinal Classification[J]. *Sci China Inf Sci*, 2010, **53**(6): 1188−1200. DOI: 10.1007/s11432-010-3117-7.