基于三支决策的差别矩阵属性约简算法

宋姝璇, 张宇红, 万仁霞, 苗夺谦

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100176

基于三支决策的差别矩阵属性约简算法

宋姝璇[1]　张宇红[1]　万仁霞[1]　苗夺谦[2]
1 北方民族大学数学与信息科学学院　银川 750021
2 同济大学计算机科学与技术系　上海 201804
(wanrx1022@126.com)

摘　要　属性约简是粗糙集理论研究的核心内容之一，也是粗糙集理论的重要组成部分。该方法旨在减少冗余信息，提取出最具代表性和关键性质的属性集合。在属性约简的过程中，差别矩阵通常用于度量属性之间的关系，通过分析差别矩阵，研究者可以识别那些在描述系统行为方面贡献相似信息的属性，从而进行属性约简。基于三支决策的差别矩阵属性约简算法从差别矩阵的属性出发，首先刻画核以外的属性重要度，并以三支决策理论为基础构建一种新的属性约简方法。算法将传统概率粗糙集的上、下近似划分为三支决策中的正域、负域、边界域，基于不同的区域给出了决策规则，并通过决策损失函数来控制三支决策阈值。与同类算法相比，所提算法可以得到更为简洁的约简集和决策规则，且具有更小的时间复杂度。

关键词：三支决策；阈值；差别矩阵；重要度；属性约简
中图分类号　TP391

Attribute Reduction of Discernibility Matrix Based on Three-way Decision

SONG Shuxuan[1], ZHANG Yuhong[1], WAN Renxia[1] and MIAO Duoqian[2]
1 College of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China
2 Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

Abstract　Attribute reduction is one of the core contents of the study of rough set theory and a crucial component of the theory itself. This approach aims to minimize redundant information and extract a set of attributes that is both representative and pivotal. In the process of attribute reduction, a difference matrix is commonly employed to measure the relationships between attributes. By analyzing the difference matrix, researchers can identify attributes that contribute similar information in describing the system's behavior, facilitating the process of attribute reduction. The three-decision-based difference matrix attribute reduction algorithm, starting from the attributes of the difference matrix, characterizes the importance of attributes beyond the core. It establishes a novel approach to attribute reduction based on the three-decision theory, dividing the upper and lower approximations of traditional probability rough sets into the positive region, negative region, and boundary region within the framework of three decisions. The proposed algorithm provides decision rules based on different regions and controls the three decision thresholds through a decision loss function. Compared to similar algorithms, it yields more concise reduction sets and decision rules, and has a lower time complexity.

Keywords　Three-way decision, Threshold value, Discernibility matrix, Importance degree, Attribute reduction

粗糙集理论[1](Rough Sets Theory)是波兰学者 Pawlak 于 1982 年提出的一种数据分析方法，已成为强有力的处理模糊和不确定知识的工具[2]。近年来，粗集理论已被成功运用于气象分析、农作物识别[3]、模式识别[4]及机器学习[5]等多个领域。

在决策表中进行属性约简是粗糙集理论研究的核心内容之一[6]，其主要思想是在保持分类能力不变的前提下，消除信息系统(决策表)中不必要的知识，导出最终的决策或分类规则。经典的属性约简算法包括：一般性属性约简算法、基于遗传算法的约简算法[7]、基于差别矩阵的约简算法[8]、基于属性依赖度的约简算法[9]及基于条件信息熵的约简算法[10]。

差别矩阵是信息处理中一项重要的数学方法，国内外许多学者对基于差别矩阵的属性约简方法进行了积极有益的探索和研究。文献[11]提出基于差别矩阵的特征选择算法，该算法利用有效信息更新技术来缩短计算时长。文献[12]则介绍了一种特定类 β 分布约简算法，该算法同样基于差别矩阵，不同之处在于，该算法可以保持特定类的 β 分布在约简前后的不变性，而且在效率上胜过了全类算法。文献[13]探讨了

基金项目：国家自然科学基金(62066001)；宁夏自然科学基金(2021AAC03203)；宁夏科技领军人才项目(2022GKLRLX08)；北方民族大学研究生创新项目(YCX23088)
This work was supported by the National Natural Science Foundation of China (62066001), Natural Science Foundation of Ningxia, China (2021AAC03203), Ningxia Science and Technology Leading Talent Project (2022GKLRLX08) and Graduate Innovation Project of North Minzu University (YCX23088).
通信作者：万仁霞(wanrx1022@126.com)

一种基于等价关系的属性约简方法，适用于粒协调决策形式的背景。这个算法能够更加便捷有效地获取知识。文献[14]在广义图像模糊软集合中使用了可调节的加权软差别矩阵来解决决策问题，提升了操作的灵活性。文献[15]通过约简目标构建β分布的约简差别矩阵和差别函数来减少原始信息的损失。文献[16]提出基于差别矩阵多特定类的不完备决策系统广义决策约简算法，该算法可以解决多特定类的约简问题。文献[17]研究基于多特定类序决策表的下近似属性约简方法，将多特定类序决策表约简退化为单特定类的序决策表约简。

三支决策理论[18]是著名学者姚一豫教授在粒计算和粗糙集理论的基础上提出的，其核心思想是将对象域分为接受域、不确定域和拒绝域3个部分，从而实现分而治之的策略。该理论的研究主要集中在不确定域的处理方面[19]。三支决策已经发展成为一种独立而有效的信息处理方法，并在多个领域中发挥着重要的作用，包括但不限于聚类[20]、股票投资评估[21]、信用卡评价[22]以及城市垃圾管理[23]。

本文从差别矩阵的属性出发，利用核以外的属性依赖度刻画核以外的属性重要度，并以三支决策理论为基础，构建属性约简方法，提出了基于三支决策的差别矩阵属性约简算法。

1 相关概念

定义1[24] 一个决策表S可以表达为$S=\langle U,C,D,V,f\rangle$。$U$为论域，即系统中所有对象的有限非空集，其中$C\cup D=A, C\cap D=\emptyset$。$A$为属性集合，其子集$C$和$D$分别为条件属性与决策属性，$C=\{a_1,a_2,\cdots,a_t\}$为条件属性集合，$D=\{d_1,d_2\cdots,d_t\}$为决策属性集合；$V=\cup V_a$是属性值的集合，$V_a$是属性$a\in A$的值域；$f$为一个映射函数：$U\times A\to V$，其定义了$U$中每一个对象的属性值，即$\forall a\in A, x\in U, f(x,a)\in V_a$。

定义2[24] 决策属性D对条件属性C的依赖度定义为：$\gamma_C(D)=|POS_C(D)|/|U|$。这里，$POS_C(D)$表示$D$的$C$正域，$|\cdot|$表示集合的基数。

定义3[1] 决策表$S=\langle U,C,D,V,f\rangle$，条件属性子集$B\subset C$关于$D$的重要度定义为：$sig_B(D)=\gamma_C(D)-\gamma_{C-B}(D)$。

定义4[24] 对于$E\in X$，令E为X的一族等价关系，且$d\in E$，若$IND(E)=IND(E-\{d\})$，则称d为E中不必要的；否则d为E中必要的。

定义5[24] 设$U=\{x_1,x_2\cdots,x_n\}$为论域，决策表S的差别矩阵$\boldsymbol{M}_s$是$n\times n$的矩阵，其任一元素$m_{ij}=\{a\in C: f(x_i,a)\neq f(x_j,a)$且$\omega(x_i,x_j)=1\}$，其中：

$$\omega(x_i,x_j)=\begin{cases}1,\ x_i\in POS_c(D)\text{且}\ x_j\notin POS_c(D)\\ 1,\ x_i\notin POS_c(D)\text{且}\ x_j\in POS_c(D)\\ 1,\ (x_i,x_j)\in POS_c(D)\text{且}(x_i,x_j)\notin ind(D)\\ 0,\ \text{其他}\end{cases}$$

该差别矩阵既适用于相容决策表，也适用于不相容决策表。显然，差别矩阵$\boldsymbol{M}_s$是对角线为$\emptyset$的对称矩阵。

定义6[25] 设X为论域U的子集，即$X\subseteq U$，给定一对阈值(α,β)，且满足$0\leqslant\beta<\alpha\leqslant 1$，则$X$的$(\alpha,\beta)$的正域、负域、边界域可以定义为：

$$POS_{(\alpha,\beta)}(X)=\{x\in U\mid Pr(X\mid[x]_R)\geqslant\alpha\}$$

$$BND_{(\alpha,\beta)}(X)=\{x\in U\mid \beta<Pr(X\mid[x]_R)<\alpha\}$$

$$NEG_{(\alpha,\beta)}(X)=\{x\in U\mid Pr(X\mid[x]_R)\leqslant\beta\}$$

其中，$[x]_R$为X关于R的等价类；$Pr(X\mid[x]_R)$表示任何一个对象在属于$[x]_R$条件下属于X的条件概率。

由于不可分辨关系本质上就是等价关系，因此对于条件属性子集，有下面的结论。

定理1 条件属性子集$A, B\subset C$。若$\frac{U}{IND(B)}\subseteq\frac{U}{IND(A)}$，则$sig(B)>sig(A)$。

证明：$[x]_B=\{y_1\in U\mid(x,y_1)\in IND(B)\}$，$[x]_A=\{y_2\in U\mid(x,y_2)\in IND(A)\}$。

由于$\frac{U}{IND(B)}\subseteq\frac{U}{IND(A)}$，则$[x]_B\subseteq[x]_A$。于是$POS_{C-A}(D)\subseteq POS_{C-B}(D)$，所以得出：

$$sig(B)=\gamma_C(D)-\gamma_{C-B}(D)>\gamma_C(D)-\gamma_{C-A}(D)=sig(A)$$

2 阈值对$(\boldsymbol{\alpha},\boldsymbol{\beta})$的计算

从上述的讨论可以看出，阈值对(α,β)是确定正域、负域、边界域的关键，本文采取下述方法来计算α和β的值。

对于一个集合Ω，首先构造出一个状态集合，即$\Omega=\{X_1,X_2\}$，其中，X_1和X_2是互补关系的两种状态，即对象属于X_1或者属于X_2。由分类结果的3个域，构造出一个决策动作集$D=\{\alpha_P,\alpha_N,\alpha_B\}$，其中$\alpha_P,\alpha_N,\alpha_B$分别代表将一个对象分类到概率粗糙集上的正域、负域、边界域的决策动作。不同的决策动作代表着不同的分类结果，可能出现的6种损失函数见表1。其中，第一列函数表示第一个对象属于集合X时，采取动作$\alpha_P,\alpha_N,\alpha_B$带来的损失函数，记为$\lambda_{PX_1},\lambda_{NX_1},\lambda_{BX_1}$；第二列函数表示一个对象属于集合$\sim X$时，采取动作$\alpha_P,\alpha_N,\alpha_B$带来的损失函数，记为$\lambda_{PX_2},\lambda_{NX_2},\lambda_{BX_2}$。

表1 损失函数

Table 1 Loss Function

	X(正例)	$\sim X$(负例)
α_P	$\lambda_{PX_1}=\lambda(\alpha_P\mid X)$	$\lambda_{PX_2}=\lambda(\alpha_P\mid\sim X)$
α_N	$\lambda_{NX_1}=\lambda(\alpha_N\mid X)$	$\lambda_{NX_2}=\lambda(\alpha_N\mid\sim X)$
α_B	$\lambda_{BX_1}=\lambda(\alpha_B\mid X)$	$\lambda_{BX_2}=\lambda(\alpha_B\mid\sim X)$

对于$[x]_R$中的对象，采取不同的动作所产生的损失表示如下：

$$R(\alpha_P\mid[x]_R)=\lambda_{PX_1}Pr(X\mid[x]_R)+\lambda_{PX_2}Pr(\sim X\mid[x]_R)$$

$$R(\alpha_N\mid[x]_R)=\lambda_{NX_1}Pr(X\mid[x]_R)+\lambda_{NX_2}Pr(\sim X[x]_R)$$

$$R(\alpha_B\mid[x]_R)=\lambda_{BX_1}Pr(X\mid[x]_R)+\lambda_{BX_2}Pr(\sim X\mid[x]_R)$$

按照贝叶斯决策论的最小风险决策准则，可以得到如下规则。

(P)若$R(\alpha_P\mid[x]_R)\leqslant R(\alpha_B\mid[x]_R)$，且存在：$R(\alpha_P\mid[x]_R)\leqslant R(\alpha_N\mid[x]_R)$，则$x\in POS_{(\alpha,\beta)}(X)$。

(N)若$R(\alpha_N\mid[x]_R)\leqslant R(\alpha_P\mid[x]_R)$，且存在：$R(\alpha_N\mid[x]_R)\leqslant R(\alpha_B\mid[x]_R)$，则$x\in NEG_{(\alpha,\beta)}(X)$。

(B)若$R(\alpha_B\mid[x]_R)\leqslant R(\alpha_P\mid[x]_R)$，且存在：$R(\alpha_B\mid[x]_R)\leqslant R(\alpha_N\mid[x]_R)$，则$x\in BND_{(\alpha,\beta)}(X)$。

上述损失函数满足$\lambda_{PX_1}\leqslant\lambda_{BX_1}\leqslant\lambda_{NX_1}$和$\lambda_{NX_2}\leqslant\lambda_{BX_2}\leqslant\lambda_{PX_2}$。通常在学习过程中设定：$\lambda_{PX_1}=0,\lambda_{NX_2}=0$。

根据上面的讨论，可知$0\leqslant\beta<\alpha\leqslant 1$。$\alpha$和$\beta$的具体计算式如下：

$$\alpha=\frac{\lambda_{PX_2}-\lambda_{BX_2}}{(\lambda_{PX_2}-\lambda_{BX_2})+(\lambda_{BX_1}-\lambda_{PX_1})}$$

$$\beta=\frac{\lambda_{BX_2}-\lambda_{NX_2}}{(\lambda_{BX_2}-\lambda_{NX_2})+(\lambda_{NX_1}-\lambda_{BX_1})}$$

3 基于三支决策的差别矩阵属性约简算法

3.1 算法思想

首先，根据给出的决策表计算出差别矩阵 $\boldsymbol{M}_s$，将 $\boldsymbol{M}_s$ 中包含的单个属性元素作为核元素（若存在多个核元素，优先提取频数最大的元素），把 $\boldsymbol{M}_s$ 中含有核属性的项置空后获得新的差别矩阵，若此时差别矩阵为空，则算法结束，输出约简集；否则使用依赖度做差，计算除核以外的剩余条件属性重要度，按照重要度递减排序，将重要度最大的属性划入正域，重要度为零的属性划入负域，其余属性划入边界域。再考查边界域中的属性，直至边界域中属性元素循环符合属性核和约简集的要求。

3.2 算法描述

符号说明：R 表示属性约简集；C_0 表示核；Sig 表示属性重要度；M_S 表示差别矩阵；(α,β)表示阈值；(POS)表示正域；(NEG)表示负域；(BND)表示边界域。

算法 1 基于三支决策的差别矩阵属性约简算法

输入：决策表 S=(U,C,D,V,f)

输出：决策表 S 的核 $\mathrm{C_0}$ 和相对约简集 R

具体算法如下：

Step1 计算决策表 S 的差别矩阵 $\mathbf{M_s}$。

Step2 选出差别矩阵 $\mathbf{M_s}$ 中只有单个 $\mathrm{a_k}$ 的属性元素。选取频数最大的 $\mathrm{a_k}$。求得 $\mathrm{C_0=a_k}$，$\mathrm{R=C_0}$。

Step3 将差别矩阵 $\mathbf{M_s}$ 中包含核属性 $\mathrm{a_k}$ 的元素项置空，若 $\mathrm{M_s}=\varnothing$，则转 Step5，否则转 Step4。

Step4 计算元素 $\mathrm{a_k}\in(\mathrm{C-C_0})$的重要度，并按照重要度 $\mathrm{Sig(a_k,C-C_0,D)}$ 值的递减顺序进行以下步骤：

(1)计算阈值(α,β)。

(2)选择使 $\mathrm{Sig(a_k,C-C_0,D)}$ 最大的属性 $\mathrm{a_k}$ 划入正域(POS)；$\mathrm{Sig(a_k,C-C_0,D)}$ 为 0 的属性 $\mathrm{a_k}$ 划入负域(NEG)；$\mathrm{Sig(a_k,C-C_0,D)}$为其他的属性 $\mathrm{a_k}$ 划入边界域(BND)。

(3)将差别矩阵 $\mathbf{M_s}$ 中含有更新后的正域(POS)中属性元素 $\mathrm{a_k}$ 的项置空，若 $\mathbf{M_s}=\varnothing$，则转到 Step5，否则转到 Step4。

(4)计算边界域中元素的条件概率 $\mathrm{Pr(X_1|[x]_R)}$。

(5)将正域(POS)中的属性元素 $\mathrm{a_k}$ 从差别矩阵中剔除，若 $\mathbf{M_s}=\varnothing$，则转到 Step5，否则转(4)。

Step5 输出 $\mathrm{C_0}$ 和 R。

本算法中，Step4(1)判决元素的三支域归属，即：

当 $Pr(X_i|[x]_R)>\alpha$ 时，将满足条件的元素 X_i 所在属性划入正域；

当 $Pr(X_i|[x]_R)<\beta$ 时，将满足条件的元素 X_i 所在属性划入负域；

当 $\beta<Pr(X_i|[x]_R)<\alpha$ 时，将满足条件的元素 X_i 所在属性划入边界域。

4 实例分析与算法比较

4.1 实例分析

为了验证新算法的有效性，下面通过所提新算法来求表 2 决策表的核和相对约简集。具体求解过程如下：

Step1 计算阈值。在学习过程中通常设定 $\lambda_{PX_1}=0$，$\lambda_{NX_2}=0$。在满足上述不等式条件下，设值为：$\lambda_{PX_1}=0$，$\lambda_{BX_1}=5$，$\lambda_{NX_1}=9.5$，$\lambda_{PX_2}=9.5$，$\lambda_{BX_2}=4$，$\lambda_{NX_2}=0$。于是得到阈值参数为$(\alpha,\beta)=(0.55,0.45)$。

Step2 根据差别矩阵的定义，求得表 2 决策表的差别矩阵 $\boldsymbol{M}_s$ 是一个 14×14 的对称矩阵(此处为上三角矩阵)，如图 1 所示。

Step3 选出差别矩阵 $\boldsymbol{M}_s$ 中只有单个 a_k 且频数最大的属性元素，求得核 $C_0=\{a_5\}$，得到此时的相对约简集 $R=C_0$。

表 2 决策表

Table 2 Decision table

U	a_1	a_2	a_3	a_4	a_5	a_6	D
x_1	2	1	1	2	4	2	N
x_2	1	1	1	1	4	2	N
x_3	1	1	2	1	3	1	P
x_4	2	2	1	2	2	2	P
x_5	2	3	2	2	2	2	P
x_6	2	3	2	1	2	2	N
x_7	1	3	2	1	4	1	P
x_8	2	3	2	1	2	1	N
x_9	1	3	2	1	4	1	P
x_{10}	3	2	2	2	2	2	P
x_{11}	2	1	2	1	1	1	P
x_{12}	1	3	2	1	2	2	P
x_{13}	2	1	2	1	4	1	P
x_{14}	3	2	2	2	3	2	N

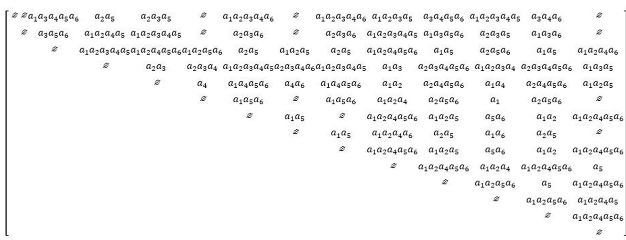


图 1 决策表的差别矩阵

Fig. 1 Difference matrix of the decision table

Step4 将差别矩阵 $\boldsymbol{M}_s$ 中包含 a_5 的项置空，如图 2 所示。

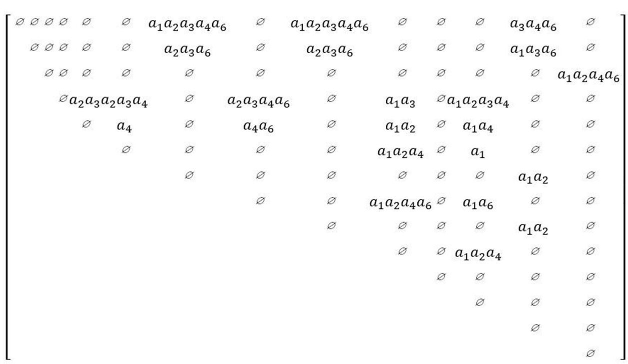


图 2 决策表中含 a_5 属性的组合项置空后的差别矩阵

Fig. 2 Difference matrix of the decision table with combination items containing attribute a_5 set to null

Step5 对 $a_k\in(C-C_0)=\{a_1,a_2,a_3,a_4,a_6\}$，计算：

$$Sig(a_1,C-C_0,D)=\gamma_C(D)-\gamma_{\{C-C_0\}-\{a_1\}}(D)\approx0.357$$

$$Sig(a_2,C-C_0,D)=\gamma_C(D)-\gamma_{\{C-C_0\}-\{a_2\}}(D)\approx0.357$$

$Sig(a_3, C-C_0, D)=\gamma_C(D)-\gamma_{\{C-C_0\}-\{a_3\}}(D)=0$

$Sig(a_4, C-C_0, D)=\gamma_C(D)-\gamma_{\{C-C_0\}-\{a_4\}}(D)\approx 0.143$

$Sig(a_6, C-C_0, D)=\gamma_C(D)-\gamma_{\{C-C_0\}-\{a_6\}}(D)=0$

显然,可以得出如下关系:

$$\begin{aligned} Sig(a_1, C-C_0, D) &= Sig(a_2, C-C_0, D) \\ &> Sig(a_4, C_0, D) \\ &> Sig(a_3, C_0, D) \\ &= Sig(a_6, C_0, D) \end{aligned}$$

(1)选择使 $Sig(a_k, C-C_0, D)$最大的属性 a_k 划入正域(POS);$Sig(a_k, C_0, D)$为 0 的属性 a_k 划入负域(NEG);$Sig(a_k, C_0, D)$为其他的属性 a_k 划入边界域(BND)。显然,属性 a_1,a_2 放入正域,属性 a_4 放入边界域。

(2)将正域(POS)中包含的属性元素 a_1、a_2 从差别矩阵 $\boldsymbol{M}_s$ 中置空。

(3)决策属性 D 下的条件划分为:

$U/D=\{D_1, D_2\}$

其中:

$D_1=\{x_1, x_2, x_6, x_8, x_{14}\}$

$D_2=\{x_3, x_4, x_5, x_7, x_9, x_{10}, x_{11}, x_{12}, x_{13}\}$

已知 a_4 中子集 $A_4=\{x_1, x_4, x_5, x_10, x_{14}\}$,则条件概率为:

$Pr(A_4 | D_1)=1/3$

$Pr(A_4 | D_2)=2/5$

根据定义 6 可得,在阈值为$(\alpha,\beta)=(0.55, 0.45)$的条件下,对于决策概念 D_1,D_2 的概率粗糙三支近似分别为:

$Pr[POS_{(\alpha,\beta)}(A_4)]=1/3$

$Pr[BND_{(\alpha,\beta)}(A_4)]=1/5$

$Pr[NEG_{(\alpha,\beta)}(A_4)]=0$

同理得到 a_3 和 a_6 中的条件概率如下:

$A_3=\{x_3, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}\}$,则条件概率为:

$Pr[POS_{(\alpha,\beta)}(A_3)]=0$

$Pr[BND_{(\alpha,\beta)}(A_3)]=0$

$Pr[NEG_{(\alpha,\beta)}(A_3)]=3/5$

$A_6=\{x_1, x_2, x_4, x_5, x_6, x_10, x_{12}, x_{14}\}$,则条件概率为:

$Pr[POS_{(\alpha,\beta)}(A_6)]=8/9$

$Pr[BND_{(\alpha,\beta)}(A_6)]=1/5$

$Pr[NEG_{(\alpha,\beta)}(A_6)]=1/5$

综上,a_3,a_6 属性存在处于负域范围内的元素,且 a_3 属性的全部元素均在负域中,a_4 属性中不存在处于负域内的元素,且正域条件概率大于边界域条件概率,因而,将 a_4 属性作为约简集中的一员。

(4)将差别矩阵 $\boldsymbol{M}_s$ 中含有 a_4 属性的元素项置空后,发现 $\boldsymbol{M}_s=\emptyset$。

Step6 输出条件属性的核 C_0 及最终的相对约简集 $R=\{a_5, a_1, a_2, a_4\}$。

以 $R=\{a_5, a_1, a_2, a_4\}$为条件属性,得到以下 7 条决策规则:

(1)$a_1=1\wedge a_2=1\wedge a_4=1\wedge a_5=3\Rightarrow P$

(2)$a_1=2\wedge a_2=3\wedge a_4=2\wedge a_5=2\Rightarrow P$

(3)$a_1=1\wedge a_2=3\wedge a_4=1\wedge a_5=4\Rightarrow P$

(4)$a_1=2\wedge a_2=1\wedge a_4=2\wedge a_5=4\Rightarrow N$

(5)$a_1=2\wedge a_3=1\wedge a_4=2\wedge a_5=3\Rightarrow P$

(6)$a_1=1\wedge a_3=1\wedge a_4=1\wedge a_5=4\Rightarrow P$

(7)$a_1=1\wedge a_3=3\wedge a_4=1\wedge a_5=4\Rightarrow P$

4.2 算法比较

为了进一步展现算法性能,本文选取基于差别矩阵属性约简的最新研究成果文献[29]和文献[30]作为比较对象。

对于表 2 决策表,文献[29]给出的基于 Markov Blanket 的双向搜索属性约简算法来寻求约简集和决策规则,得到相对约简集为 $R=\{a_4, a_3, a_5, a_6, a_1, a_2\}$,以此约简集为条件属性得到 12 条规则。文献[30]给出的基于决策多重粒度空间的悲观上近似约简算法求得相对约简集为 $R=\{a_1, a_4, a_2, a_3, a_5, a_6\}$,以约简集为条件属性得到 13 条规则,具体如表 3 所列。显然,本文所提算法得到的约简集较文献[29]算法及文献[30]算法得到的约简集更为简洁。此外,本文提出的算法在生成决策时较对比算法也更为精简。如表 3 所列,以本文算法得出的相对约简集生成的决策规则数目为 7 条,少于文献[29]算法得到的决策规则数目 12 条和文献[30]算法得到的决策规则数目 13 条。这表明本文提出的算法在提供简明而有效的决策规则方面取得了成功。

表 3 算法决策规则对比表

Table 3 Comparison of decision rules of algorithms

1	$a_1=1\wedge a_2=1\wedge a_4=1\wedge a_5=3\Rightarrow P$	$a_4=2\wedge a_3=2\wedge a_5=2\wedge a_6=2\wedge a_1=1\wedge a_2=1\Rightarrow P$	$a_1=1\wedge a_4=2\wedge a_2=1\wedge a_3=1\wedge a_5=4\wedge a_6=2\Rightarrow N$
2	$a_1=2\wedge a_2=3\wedge a_4=2\wedge a_5=2\Rightarrow P$	$a_4=2\wedge a_3=2\wedge a_5=3\wedge a_6=2\wedge a_1=1\wedge a_2=1\Rightarrow N$	$a_1=1\wedge a_4=2\wedge a_2=1\wedge a_3=1\wedge a_5=4\wedge a_6=1\Rightarrow N$
3	$a_1=1\wedge a_2=3\wedge a_4=1\wedge a_5=4\Rightarrow P$	$a_4=2\wedge a_3=2\wedge a_5=3\wedge a_6=2\wedge a_1=2\wedge a_2=1\Rightarrow P$	$a_1=2\wedge a_4=2\wedge a_2=2\wedge a_3=1\wedge a_5=2\wedge a_6=2\Rightarrow P$
4	$a_1=2\wedge a_2=1\wedge a_4=2\wedge a_5=4\Rightarrow N$	$a_4=2\wedge a_3=1\wedge a_5=4\wedge a_6=2\wedge a_1=2\wedge a_2=1\Rightarrow P$	$a_1=2\wedge a_4=3\wedge a_2=2\wedge a_3=2\wedge a_5=2\wedge a_6=2\Rightarrow P$
5	$a_1=2\wedge a_3=1\wedge a_4=2\wedge a_5=3\Rightarrow P$	$a_4=1\wedge a_3=2\wedge a_5=4\wedge a_6=2\wedge a_1=2\wedge a_2=1\Rightarrow N$	$a_1=2\wedge a_4=3\wedge a_2=2\wedge a_3=1\wedge a_5=2\wedge a_6=2\Rightarrow N$
6	$a_1=1\wedge a_3=1\wedge a_4=1\wedge a_5=4\Rightarrow P$	$a_4=3\wedge a_3=1\wedge a_5=3\wedge a_6=2\wedge a_1=1\wedge a_2=1\Rightarrow P$	$a_1=1\wedge a_4=3\wedge a_2=2\wedge a_3=1\wedge a_5=4\wedge a_6=1\Rightarrow P$
7	$a_1=1\wedge a_3=3\wedge a_4=1\wedge a_5=4\Rightarrow P$	$a_4=2\wedge a_3=2\wedge a_5=3\wedge a_6=1\wedge a_1=2\wedge a_2=1\Rightarrow P$	$a_1=2\wedge a_4=3\wedge a_2=2\wedge a_3=1\wedge a_5=2\wedge a_6=1\Rightarrow N$
8		$a_4=2\wedge a_3=2\wedge a_5=2\wedge a_6=2\wedge a_1=1\wedge a_2=2\Rightarrow P$	$a_1=1\wedge a_4=3\wedge a_2=2\wedge a_3=1\wedge a_5=4\wedge a_6=1\Rightarrow P$
9		$a_4=2\wedge a_3=1\wedge a_5=4\wedge a_6=2\wedge a_1=1\wedge a_2=2\Rightarrow P$	$a_1=3\wedge a_4=2\wedge a_2=2\wedge a_3=2\wedge a_5=2\wedge a_6=2\Rightarrow P$
10		$a_4=1\wedge a_3=2\wedge a_5=4\wedge a_6=2\wedge a_1=1\wedge a_2=2\Rightarrow N$	$a_1=2\wedge a_4=1\wedge a_2=2\wedge a_3=1\wedge a_5=1\wedge a_6=1\Rightarrow P$
11		$a_4=3\wedge a_3=1\wedge a_5=3\wedge a_6=2\wedge a_1=1\wedge a_2=2\Rightarrow P$	$a_1=1\wedge a_4=3\wedge a_2=2\wedge a_3=1\wedge a_5=2\wedge a_6=1\Rightarrow P$
12		$a_4=2\wedge a_3=2\wedge a_5=3\wedge a_6=1\wedge a_1=1\wedge a_2=2\Rightarrow P$	$a_1=2\wedge a_4=1\wedge a_2=2\wedge a_3=1\wedge a_5=4\wedge a_6=1\Rightarrow P$
13			$a_1=3\wedge a_4=2\wedge a_2=2\wedge a_3=2\wedge a_5=3\wedge a_6=2\Rightarrow N$

从复杂度分析角度来看,假设数据集样本数为$|U|$,条件属性数为$|C|$,数据分区 P 中的元素数量为$|P|$,则本文算法

的时间复杂度为 $O(|U|^2|C|)$。文献[29]中模糊差别矩阵产生最多$|U|^2$个元素项，计算每个属性的频率和判断是否有Markov Blanket时都要遍历所有的矩阵项，故其总的时间复杂度为 $O(|U|^2|C|^2)$。同理得出文献[30]的时间复杂度为 $O(\prod_{P\in P}|P|+|U|^2|C|^2)$。故本文的时间复杂度得到显著降低，这是因为本文算法不需要计算冗余属性的重要度，而是直接利用三支决策中的三域计算与阈值来判断，从而有效提高了决策效率。

结束语 属性约简是减少冗余特征的有效方法，而减少冗余属性是提高分类性能及降低分类成本的有效手段。本文对粗糙集理论属性约简算法进行研究，提出了基于三支决策的差别矩阵属性约简算法。算法首先利用差别矩阵来剔除无用信息，并通过核属性缩小了搜索范围，进一步排除了冗余特征。而对于核属性以外的条件属性则，通过三支决策的决策规则得到正域、负域与边界域的划分，并通过决策损失函数给出了属性约简决策阈值的计算依据。与同类算法相比，本文所提出的算法不仅可以得到更为简洁的约简集与简明有效的决策规则，而且因省去了考虑冗余属性重要度的计算，在时间复杂度方面也得到明显降低，从而有效地提高了决策效率。但在属性约简中，可能会面临不确定性的问题，未来的研究可以考虑如何进一步提高算法的鲁棒性，使其在处理不同类型和规模的数据时都能取得好的性能。

参考文献

[1] PAWLAK Z. Rough Set[J]. International Journal of Parallel Programming, 1982, 11, 341-356.

[2] LI F, YU Y T, XIAO J. Correlation Analysis of Haze and Meteorological Elements Based on the Rough Set Model[J]. Journal of Nanchang University(Natural Science Edition), 2019, 43(2): 187-192.

[3] ACHARJYA D P, RATHI R. An integrated fuzzy rough set and real coded genetic algorithm approach for crop identification in smart agriculture[J]. Multimedia Tools and Applications, 2022, 81(24): 35117-35142.

[4] GUESGEN H W. Using Rough Sets to Improve Activity Recognition Based on Sensor Data[J]. Sensors, 2020, 20(6): 1-10.

[5] HOSSAIN M T, WATAADA J, HERMANA M, et al. Supervised Machine Learning in Electrofacies Classification: A Rough Set Theory Approach[J]. Journal of Physics: Conference Series, 2020, 1529(5): 052048.

[6] WANG L H, WU G F. Attribute Reduction and Information Granularity[J]. Journal of Systemics, Cybernetics and Informatics, 2003, 1(1): 32-37.

[7] GAO K, TAN Y J, PAN W. Rough Set Knowledge Reduction Algorithm based on Improved Chaos Genetic Algorithm[C]//东北大学、IEEE 新加坡工业电子分会. 第 28 届中国控制与决策会议论文集(上). 2016: 5.

[8] HE Y, HE D. Discernibility Matrix-Based Attribute Reduction Algorithm of Decision Table[J]. Advanced Materials Research, 2012, 1639.

[9] ZHOU Y, YANG X J, XU Y. Research on Dependency Algorithm for Attribute Reduction[J]. Computer Engineering and Applications, 2004(4): 78-79, 223.

[10] LI J L. Improved Attribute Reduction Algorithm Based on Conditional Information Entropy[J]. Journal of North University of China(Natural Science Edition), 2014, 35(6): 709-713.

[11] QIAN W B, XIONG C Z, WANG Y L. A ranking-based feature selection for multi-label classification with fuzzy relative discernibility[J]. Applied Soft Computing, 2021, 102(1).

[12] HAN S Z, ZHANG N, ZHANG Z X. Specific Class β-Distribution Reduction for Interval-Valued Decision Systems[J]. Journal of Shandong University(Natural Science Edition), 2020, 55(11): 66-77.

[13] ZHANG X H, MI J S, LI M Z. Attribute Reduction and Rule Fusion in the Context of Granular Coordinated Decision Forms [J]. Journal of Intelligent Systems, 2019, 14(6): 1138-1143.

[14] KHAN M J, KUMAM P, LIU P D, et al. An adjustable weighted soft discernibility matrix based on generalized picture fuzzy soft set and its applications in decision making[J]. Journal of Intelligent & Fuzzy Systems, 2019, 38(2).

[15] LI L T, ZHANG N, TONG X R, et al. Beta Distribution Reduction for Interval-Valued Decision Systems Based on Difference Matrix[J/OL]. Computer Applications: 1-11[2021-02-09].

[16] TANG Y K, ZHANG N, TONG X R, et al. Generalized Decision Reduction for Multi-Specific Class in Incomplete Decision Systems[J]. Journal of Intelligent Systems, 2019, 14(6): 1199-1208.

[17] YU T Y, ZHANG N, YUE X D, et al. Approximate Reduction Based on Multi-Specific Class in Sequential Decision Tables[J]. Computer Science, 2019, 46(10): 242-251.

[18] LI Y N, QI J J, SUN B Z, et al. Three-Decision Theory Domain Method [M]. Beijing: Science Press, 2019.

[19] LI F, YU Y T, XIAO J. Correlation Analysis of Haze and Meteorological Elements Based on the Rough Set Model[J]. Journal of Nanchang University(Natural Science Edition), 2019, 43(2): 187-192.

[20] HONG Y, YUN C, PAWAN L, et al. A three-way cluster ensemble approach for large-scale data[J]. International Journal of Approximate Reasoning, 2019, 115.

[21] TANG G L, CHICLANA F, LIU P D. A decision-theoretic rough set model with q -rung orthopair fuzzy information and its application in stock investment evaluation [J]. Applied Soft Computing Journal, 2020, 91.

[22] LI Z W, HUANG D. A three-way decision method in a fuzzy condition decision information system and its application in credit card evaluation[J]. Granular Computing, 2020, 5(4).

[23] LUO C, JU Y B, GIANNAKIS M, et al. A novel methodology to select sustainable municipal solid waste management scena-rios from three-way decisions perspective[J]. Journal of Cleaner Production, 2021, 280(5): 124312.

[24] ZHANG W X, WU W Z, LIANG J, et al. Rough Set Theory and Methods [M]. Beijing: Science Press, 2001.

[25] YAO Y Y. Three-way decisions with probabilistic rough sets[J]. Information Sciences, 2010, 180: 341-353.

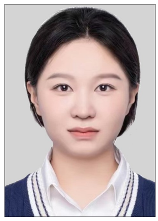
[26] YAO Y Y. The superiority of three-way decision inprobabilistic rough set models[J]. Information Sciences, 2011, 181: 1080-1096.

[27] LIU D, LI T R, MIAO D, et al. Three-way decisions and granular computing[M]. Bejing: Science Publishing, 2013, 332.

[28] FAN X, CHEN H M. Stepwise Optimized Feature Selection Algorithm Based on Difference Matrix and mRMR[J]. Computer Science, 2020, 47(1): 87-95.

[29] HU D Y, ZHOU J, GAO C. Fuzzy Difference Matrix Attribute Reduction Algorithm Based on Markov Blanket[J]. Fuzzy Systems and Mathematics, 2022, 36(6): 54-63.

[30] TAN A H, WU W Z, LI J J, et al. Reduction foundation with multigranulation rough sets using discernibility[J]. Artificial Intelligence Review: An International Science and Engineering Journal, 2020, 53(4).



SONG Shuxuan, born in 2000. Her main research interests include three-way decision and rough set, data mining and pattern recognition.

WAN Renxia, born in 1975, Ph.D, professor, Ph.D supervisor. His main research interests include information systems, data mining, knowledge learning, and granular computing.