DOI: 10.11992/tis.201710011
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180408.1625.026.html

不协调区间值决策系统的最大分布约简

尹继亮[1,2], 张楠[1,2], 童向荣[1,2], 陈曼如[1,2]
（1. 烟台大学 数据科学与智能技术山东省高校重点实验室, 山东 烟台 264005; 2. 烟台大学 计算机与控制工程学院, 山东 烟台 264005）

摘　要: 分布式约简可以保证约简前后决策系统各规则的置信度保持不变, 是属性约简的重要方法之一。最大分布式约简保持了约简前后决策系统中可信程度最大的规则不变, 提取置信度较大的规则在智能决策中具有广泛的应用价值。本文在相容关系下的不协调区间值决策系统中引入最大置信度的概念, 构造最大分布保持不变的可辨识矩阵, 并给出基于可辨识矩阵的最大分布约简算法。分析了不协调区间值决策系统的最大分布约简算法与其它约简算法之间的关系。最后, 利用 UCI 标准数据集进行了实验验证, 实验结果表明了算法的有效性。

关键词: 分布式约简; 最大分布约简; 置信度; 相容关系; 可辨识矩阵; 不协调; 区间值; 决策系统
中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)03−0469−10

中文引用格式: 尹继亮, 张楠, 童向荣, 等. 不协调区间值决策系统的最大分布约简[J]. 智能系统学报, 2018, 13(3): 469–478.
英文引用格式: YIN Jiliang, ZHANG Nan, TONG Xiangrong, et al. Maximum distribution reduction in inconsistent interval-valued decision systems[J]. CAAI transactions on intelligent systems, 2018, 13(3): 469–478.

Maximum distribution reduction in inconsistent interval-valued decision systems

YIN Jiliang[1,2], ZHANG Nan[1,2], TONG Xiangrong[1,2], CHEN Manru[1,2]
(1. Key Lab for Data Science and Intelligence Technology of Shandong Higher Education Institutes, Yantai University, Yantai 264005, China; 2. School of Computer and Control Engineering, Yantai University, Yantai 264005, China)

Abstract: Distribution reduction is one of the important methods of attribute reduction as it can guarantee consistent confidence coefficients of all decision rules before and after reduction. Maximum distributed reduction keeps the unchanged rule with the highest confidence coefficient in the decision system, and extracting a rule with a high confidence coefficient has a wide application value. This paper introduces the concept of maximum confidence coefficient for inconsistent interval-valued decision systems based on compatibility relation and proposes a maximum distribution reduction algorithm based on discernibility matrix, whereby a discernibility matrix is constructed to keep the unchanged maximum distribution. The relationship between the maximum distribution reduction algorithm in inconsistent interval-valued decision systems and other reduction algorithms was analyzed. Experiments were performed using UCI standard data sets, and the proposed algorithm proved to be effective.

Keywords: distributed reduction; maximum distributed reduction; confidence coefficient; compatibility relation; discernibility matrix; inharmonious; interval-valued; decision system

属性约简[1-7]是粗糙集理论[1-3]的核心研究内容之一, 在数据挖掘、机器学习、决策分析、智能信息处理等领域取得了诸多研究成果。属性约简的目的是删除冗余属性, 只保留使决策表某种分类特征不变的最小属性子集。差别矩阵方法是一种用于求取所有属性约简的有效方法, 该方法由 Skowron[8] 于 1982 年提出, 并将差别矩阵应用于正域约简中。诸多学者在此基础上做了大量的研究工作。Kryszkie-

收稿日期: 2017−10−16.　**网络出版日期:** 2018−04−08.
基金项目: 国家自然科学基金项目 (61403329, 61572418, 61702439, 61572419, 61502410); 山东省自然科学基金项目 (ZR2016FM42); 烟台大学研究生科技创新基金项目 (YDZD1807).
通信作者: 张楠. E-mail: zhangnan0851@163.com.

wicz[9]于 1999 年在不完备信息系统下引入广义决策保持约简的概念，并提出基于差别矩阵的广义决策保持约简方法；2007 年，邓大勇等[10]首先分析了不相容信息系统下几种约简目标之间的关系；2009 年，Miao 等[11]进一步分析了 3 种约简目标之间的关系，提出不可分辨关系保持约简以及相应的差别矩阵构造方法；Zhou 等[12]在 2011 年对现有的 13 种属性约简目标进行总结，并将所有约简目标分为 4 类，完善了现有约简目标之间的关系。

分布约简保持了信息系统约简前后每条规则置信度不变。2003 年，张文修等[13]提出了分配约简、分布约简以及最大分布约简的概念，并分别给出了基于差别矩阵的分配约简、分布约简以及最大分布约简方法；2007 年，徐伟华等[14]在优势关系下提出了两种约简概念，即分布约简和最大分布约简，同时建立了基于差别矩阵的分布和最大分布约简的具体方法。如某一段时间内的温度、湿度等区间值数据在现实环境中大量存在，它较好地表示了许多不确定类型数据，区间值决策系统是经典 Pawlak 决策系统的推广，充分地考虑了数据的不确定性，在近几年得到了广泛关注。2009 年，张楠等[15]定义了α-极大相容类的概念，提出了区间值决策系统的广义决策保持约简；2016 年，张楠等[16]在不协调区间值决策系统中提出确定性规则保持约简；2016 年，张楠等[17]讨论了不协调区间值决策系统中的知识约简并提出了分布约简的概念。

基于上述研究，文献[13]和[14]分别对等价关系和优势关系下的最大分布约简进行了研究，但未有区间值决策系统的最大分布约简讨论。置信度表示了信息系统中规则的可信程度，置信度越大，规则的可信程度越高；置信度越小，规则的可信程度越低，在实际应用中，人们往往关注置信度最大的规则。为此，本文提出了区间值决策系统的最大分布约简概念，为区间值决策系统提供了一种求取所有属性约简的新方法。

1　基本知识

1.1　区间值决策系统的粗糙近似

基于相容关系的区间值粗糙集模型是经典 Pawlak 粗糙集模型的推广，首先给出相关概念和性质。

给定区间值信息系统[15-18] $\mathrm{IS}=(U,\mathrm{AT},V,f)$，其中，$U$是有限对象集合，$U=\{x_1,x_2,\cdots,x_{|U|}\}$；AT是有限属性集合，$\mathrm{AT}=\{a_1,a_2,\cdots,a_{|\mathrm{AT}|}\}$；$V$是全体属性的值域，即$V=\bigcup\limits_{a_k\in\mathrm{AT}}V_{a_k}$，$V_{a_k}$是属性$a_k\in\mathrm{AT}$的值域；$f:U\times\mathrm{AT}\to V$是一个信息函数，它指定论域$U$中每一个对象$x_i$在属性$a_k$上的区间属性值，即对任意的$x_i\in U$，$a_k\in\mathrm{AT}$，有$f(x_i,a_k)=a_k(x_i)=[l_i^k,u_i^k]$。

如果属性集AT由条件属性集C和决策属性集D组成，$C=\{a_1,a_2,\cdots a_{|C|}\}$，$D=\{d\}$，即$C\cup D=\mathrm{AT}$；$V=V_C\cup V_D$，其中，$V_C$为条件属性值集合，$V_D$为决策属性值集合；$f:U\times C\to V_C$为区间值映射，$f:U\times D\to V_D$为单值映射，则称区间值信息系统为区间值决策系统$\mathrm{DS}=(U,C\cup D,V,f)$。

定义 1　设$\eta_1=[l_i^k,u_i^k]$和$\eta_2=[l_j^k,u_j^k]$为任意两个区间值，则区间值的交运算与并运算如下。

1) 区间值交运算为

$$\eta_1\cap\eta_2=\begin{cases}0, & (u_i^k<l_j^k)\vee(u_j^k<l_i^k)\\ [\max(l_i^k,l_j^k),\min(u_i^k,u_j^k)], & \text{其他}\end{cases}$$

2) 区间值并运算为

$$\eta_1\cup\eta_2=[\min(l_i^k,l_j^k),\max(u_i^k,u_j^k)]$$

目前，度量区间值相似度比较合理的主要方法有 Jaccard 相似率、悲观相似率和乐观相似率，本文统一采用 Jaccard 相似率来度量两个区间值的相似度。

定义 2　$\mathrm{DS}=(U,C\cup D,V,f)$为区间值决策系统，对任意的$x_i,x_j\in U$，$a_k\in\mathrm{C}$，区间值$a_k(x_i)=[l_i^k,u_i^k]$和$a_k(x_j)=[l_j^k,u_j^k]$的 Jaccard 相似率[18]$\alpha_{ij}^k$定义为

$$\alpha_{ij}^k=\frac{|[l_i^k,u_i^k]\cap[l_j^k,u_j^k]|}{|[l_i^k,u_i^k]\cup[l_j^k,u_j^k]|}$$

Jaccard 相似率为两个区间数的交集与并集长度的比值，它适合度量长度相似的两个区间数。

例 1　区间值决策系统$\mathrm{DS}=(U,C\cup D,V,f)$，如表 1 所示，其中，$U=\{x_1,x_2,\cdots,x_6\}$为对象的集合，$C=\{a_1,a_2,a_3,a_4\}$为条件属性的集合，$D=\{d\}$为决策属性。

表 1　不协调区间值决策系统

Table 1　Inconsistent interval-valued decision systems

U	a_1	a_2	a_3	a_4	d
x_1	[0.86,3.13]	[−0.20,2.23]	[−0.26,2.26]	[−0.19,2.20]	1
x_2	[−0.12,2.13]	[0.79,3.20]	[0.73,3.26]	[0.73,3.26]	2
x_3	[−0.13,2.20]	[0.86,2.95]	[−0.26,2.26]	[−0.24,2.12]	2
x_4	[−0.14,2.01]	[0.85,3.01]	[0.71,3.11]	[−0.24,2.11]	2
x_5	[−0.13,2.13]	[0.79,2.94]	[−0.26,2.23]	[−0.25,2.30]	1
x_6	[−0.13,2.13]	[0.82,3.10]	[−0.24,2.19]	[−0.24,2.11]	1

令$\eta_1=[l_1^1,u_1^1]$，$\eta_1=[l_2^1,u_2^1]$，分别计算η_1和η_2的交、并：

$$\eta_1\cap\eta_2=[0.86,3.13]\cap[-0.12,2.13]=[0.86,2.13]$$
$$\eta_1\cup\eta_2=[0.86,3.13]\cup[-0.12,2.13]=[-0.12,3.13]$$

计算η_1和η_2的 Jaccard 相似率：

$$\alpha_{12}^1=\frac{|[l_1^1,u_1^1]\cap[l_2^1,u_2^1]|}{|[l_1^1,u_1^1]\cup[l_2^1,u_2^1]|}=0.391$$

定义 3[15]　对于区间值决策系统$\mathrm{DS}=(U,C\cup$

$D,V,f)$, $a_k \in C$, $\alpha \in [0,1]$, 则关于条件属性a_k的α-相容关系定义为

$$\boldsymbol{T}^{\alpha}_{\{a_k\}} = \{(x_i,x_j)|(x_i,x_j) \in U \times U, \alpha^k_{ij} > \alpha\}$$

其中α^k_{ij}表示对象 x_i 和对象 x_j 关于属性a_k的α-Jaccard 相似度, 简称α-相似度。

关于条件属性子集$A \subseteq U$的α-相容关系定义为

$$\boldsymbol{T}^{\alpha}_{A} = \{(x_i,x_j)|(x_i,x_j) \in U \times U, \alpha^k_{ij} > \alpha, a_k \in A\}$$

性质 1[15]　对于区间值决策系统$\mathrm{DS} = (U, C \cup D, V, f)$, $A \subseteq C$, $a_k \in A$, $\alpha \in [0,1]$, $\boldsymbol{T}^{\alpha}_{\{a_k\}}$是属性$a_k$的$\alpha$-相容关系, 则关于集合$A$的相容关系为

$$\boldsymbol{T}^{\alpha}_{A} = \bigcap_{a_k \in A} \boldsymbol{T}^{\alpha}_{\{a_k\}}$$

性质 2[18]　设区间值决策系统$\mathrm{DS} = (U, C \cup D, V, f)$, $A \subseteq C$, 则$\boldsymbol{T}^{\alpha}_{A}$具有:

1) 自反性: 任意 $x_i \in U$, 则$(x_i,x_j) \in \boldsymbol{T}^{\alpha}_{A}$。

2) 对称性: 任意 x_i, $x_j \in U$, 若$(x_i,x_j) \in \boldsymbol{T}^{\alpha}_{A}$, 则$(x_j,x_i) \in \boldsymbol{T}^{\alpha}_{A}$。

3) 非传递性: 任意$x_i,x_j,x_k \in U$, 若满足$(x_i,x_k) \in \boldsymbol{T}^{\alpha}_{A}$和$(x_k,x_j) \in \boldsymbol{T}^{\alpha}_{A}$, 则$(x_i,x_j) \in \boldsymbol{T}^{\alpha}_{A}$不一定成立。

定义 4[18]　设区间值决策系统$\mathrm{DS} = (U, C \cup D, V, f)$, $A \subseteq C$, $\alpha \in [0,1]$, $\boldsymbol{T}^{\alpha}_{A}$是属性集 A 的 α-相容关系, 则关于对象 x_i 在属性集 A 下的α-相容类定义为

$$S^{\alpha}_{A}(x_i) = \{x_j|x_j \in U, (x_i, x_j) \in \boldsymbol{T}^{\alpha}_{A}\}$$

对任意$x_i \in U$, 区间值决策系统DS在阈值α下的相容类集合定义为

$$S^{\alpha}_{A}(U) = \{S^{\alpha}_{A}(x_1), S^{\alpha}_{A}(x_2), \cdots, S^{\alpha}_{A}(x_n)\}$$

其中n是论域的个数。

经典粗糙集中对象间的二元关系为等价关系, 具有自反性、传递性、对称性, 导出的等价类集合是对论域的划分, 而定义 4 中的相容类是对论域的覆盖。

定义 5　给定区间值决策系统$\mathrm{DS} = (U, C \cup D, V, f)$, $S^{\alpha}_{A}(x_i)$是在相容关系下包含x_i的相容类, 则对象集合X关于α-相容关系的上、下近似[19]分别定义为

$$\overline{\mathrm{apr}}^{\alpha}_{A}(X) = \{x_i|x_i \in U, S^{\alpha}_{A}(x_i) \cap X \neq \varnothing\}$$
$$\underline{\mathrm{apr}}^{\alpha}_{A}(X) = \{x_i|x_i \in U, S^{\alpha}_{A}(x_i) \subseteq X\}$$

集合X关于α-相容关系的正域为

$$\mathrm{POS}^{\alpha}_{A}(X) = \underline{\mathrm{apr}}^{\alpha}_{A}(X)$$

下近似是由肯定属于X的对象组成的集合, 上近似是由可能属于X的对象组成的集合, 根据上下近似的概念, 决策规则可以分为确定性规则和可能性规则。

定义 6[16]　给定区间值决策系统$\mathrm{DS} = (U, C \cup D, V, f)$, $X \subseteq U$, $A \subseteq C$, 则条件属性集A的近似分类精度定义为

$$\mu^{\alpha}_{A}(X) = \frac{|\underline{\mathrm{apr}}^{\alpha}_{A}(X)|}{|\overline{\mathrm{apr}}^{\alpha}_{A}(X)|}$$

近似分类精度表示确定性规则占可能性规则的比例, 近似分类精度越大, 区间值信息系统中确定性规则越多; 反之, 确定性规则越少。

定义 7[16]　对于区间值决策系统$\mathrm{DS} = (U, C \cup D, V, f)$, $A \subseteq C$, 决策属性 D 对U 的划分为 $U/D = \{D_1, D_2, \cdots, D_{|U/D|}\}$, 决策属性 D 关于α-相容关系的上、下近似分别定义为

$$\overline{\mathrm{apr}}^{\alpha}_{A}(D) = \{x_i|x_i \in U, D_j \in U/D, S^{\alpha}_{A}(x_i) \cap D_j \neq \varnothing\}$$
$$\underline{\mathrm{apr}}^{\alpha}_{A}(D) = \{x_i|x_i \in U, D_j \in U/D, S^{\alpha}_{A}(x_i) \subseteq D_j\}$$

决策属性 D 关于α-相容关系的正域定义为

$$\mathrm{POS}^{\alpha}_{A}(D) = \overline{\mathrm{apr}}^{\alpha}_{A}(D)$$

定义 8[16]　设区间值决策系统$\mathrm{DS} = (U, C \cup D, V, f)$, $A \subseteq C$, 决策属性 D 对U 的划分为$U/D = \{D_1, D_2, \cdots, D_{|U/D|}\}$, 则在决策属性 D 下, 关于条件属性集 A 的近似分类精度定为

$$\mu^{\alpha}_{A}(D) = \frac{|\underline{\mathrm{apr}}^{\alpha}_{A}(D)|}{|\overline{\mathrm{apr}}^{\alpha}_{A}(D)|}$$

定义 5 和定义 6 是关于集合X的上、下近似和近似分类精度, 而定义 7 和定义 8 是关于决策属性D的上、下近似和近似分类精度。

定义 9[13]　对于区间值决策系统$\mathrm{DS} = (U, C \cup D, V, f)$, 对任意$x_i, x_j \in U$, 且$i \neq j$, 若 x_i 和x_j具有α-相容关系, 且满足$d(x_i) = d(x_j)$, 则称$x_i \in U$是关于属性集$A \subseteq C$的α-协调对象; 否则称为α-不协调对象。

若存在一个对象$x_i \in U$是关于$A \subseteq C$的α-不协调对象, 那么称DS为不协调区间值决策表, 否则称为协调区间值决策表。

例 2　如表 1 所示的区间值决策系统, 令$\alpha = 0.6$, 则相容关系$\boldsymbol{T}^{0.6}_{C}$为

$$\boldsymbol{T}^{0.6}_{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

根据相似率布尔矩阵, 计算阈值$\alpha = 0.6$下的相容类集合为

$$S^{0.6}_{C}(U) = \{S^{0.6}_{C}(x_1), S^{0.6}_{C}(x_2), \cdots, S^{0.6}_{C}(x_6)\}$$

式中: $S^{0.6}_{C}(x_1) = \{x_1\}$, $S^{0.6}_{C}(x_2) = \{x_2\}$, $S^{0.6}_{C}(x_3) = S^{0.6}_{C}(x_5) = S^{0.6}_{C}(x_6) = \{x_3, x_5, x_6\}$, $S^{0.6}_{C}(x_4) = \{x_4\}$。

$U/D = \{\{x_1, x_5, x_6\}, \{x_2, x_3, x_4\}\}$ 为决策属性 D 对 U 的划分, 计算决策属性 D 关于相容关系$\boldsymbol{T}^{0.6}_{C}$的上、下近似:

$$\overline{\mathrm{apr}}^{0.6}_{C}(D) = U,\ \underline{\mathrm{apr}}^{0.6}_{C}(D) = \{x_1, x_2, x_4\}$$

计算条件属性集 C 的近似分类精度:

$$\mu^{0.6}_{C}(D) = \frac{|\underline{\mathrm{apr}}^{\alpha}_{C}(D)|}{|\overline{\mathrm{apr}}^{\alpha}_{C}(D)|} = \frac{3}{6} = 0.5$$

1.2 区间值决策系统的分布约简

文献[16]提出不协调区间决策系统的分布约简。

定义 10 设区间值决策系统 $\mathrm{DS}=(U,C\cup D,V,f)$, $A\subseteq U$, $U/D=\{D_1,D_2,\cdots,D_{|U/D|}\}$, 则 $x_i\in U$ 对应的概率分布定义为

$$\mu_A^\alpha(x_i)=(D(D_1/S_A^\alpha(x_i)),D(D_2/S_A^\alpha(x_i)),\cdots,D(D_j/S_A^\alpha(x_i)),\cdots,D(D_{|U|}/S_A^\alpha(x_i)))$$

式中 $D(D_j/S_A^\alpha(x_i))=\dfrac{|D_j\cap S_A^\alpha(x_i)|}{|S_A^\alpha(x_i)|}$, $j\leqslant|U/D|$。

定义 11 设区间值决策系统 $\mathrm{DS}=(U,C\cup D,V,f)$, $U=\{x_1,x_2,\cdots,x_{|U|}\}$, 则对任意 $1\leqslant i,j\leqslant|U|$:

$$\mathrm{DM}_D^\alpha(i,j)=\begin{cases}\{a_k|a_k\in C\wedge\alpha_{ij}^k<\alpha\}, & \mu_A^\alpha(x_i)\neq\mu_A^\alpha(x_j)\\ \varnothing, & \mu_A^\alpha(x_i)=\mu_A^\alpha(x_j)\end{cases}$$

式中: $\mathrm{DM}_D^\alpha(i,j)$为基于$\alpha$-相容类的分布约简可辨识矩阵$\mathbf{DM}_D^\alpha$第$i$行$j$列的元素, $\mathbf{DM}_D^\alpha$简称为分布可辨识矩阵, 其中 $i,j=1,2,\cdots,|U|$, $\varnothing$表示空集。

定义 12 设区间值决策系统$\mathrm{DS}=(U,C\cup D,V,f)$, $C=\{a_1,a_2,\cdots,a_{|C|}\}$, $\mathrm{DM}_D^\alpha(i,j)$表示可辨识矩阵中第 i 行 j 列的元素, 基于α-相容类的分布可辨识函数定义为与$a_1,a_2,\cdots,a_{|C|}$相对应的$|C|$个布尔变量$\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_{|C|}$的布尔函数为

$$f_D^\alpha(C)(\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_{|C|})=\wedge\{\vee\mathrm{DM}_D^\alpha(i,j):\mathrm{DM}_D^\alpha(i,j)\neq\varnothing\}$$

此函数简称分布可辨识函数。这里的$\vee\mathrm{DM}_D^\alpha(i,j)$表示满足$a\in\mathrm{DM}_D^\alpha(i,j)$的全体布尔变量$\bar{a}$的析取式。

利用分配率和吸收率将 $f_D^\alpha(C)$转化为$h_D^\alpha(C)(\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_m)=(\wedge\theta_1)\vee\cdots\vee(\wedge\theta_l),\theta_k\subseteq C,k=1,2,\cdots,l$, θ_k中每一个属性元素只出现一次。

定理 1 设区间值决策系统$\mathrm{DS}=(U,C\cup D,V,f)$, $h_D^\alpha(C)$是分布可辨识函数$f_D^\alpha(C)$的形式转化, 若$A\subseteq C$是分布约简, 当且仅当 A 是$h_D^\alpha(C)$的一个蕴含项。

基于可辨识矩阵的分布约简算法 (distribution reduction algorithm based on discernibility matrix, DRADM) 描述如下。

算法 1 DRADM

输入 区间值决策系统DS, 阈值 α。

输出 区间值决策系统的所有分布保持约简结果。

1) 计算区间值决策系统DS的在阈值α下的相容类集合$S_C^\alpha(U)$;

2) 根据每个对象对应的相容类, 计算每个对象相对于每一个决策类的概率分布$\mu_C^\alpha(x_i)$;

3) 根据每个对象的可信度不同构造分布约简可辨识矩阵$\mathbf{DM}_D^\alpha$;

4) 由可辨识矩阵$\mathbf{DM}_D^\alpha$计算分布约简可辨识函数$f_D^\alpha(C)$;

5) 利用分配率和吸收率将$f_D^\alpha(C)$转化为$h_D^\alpha(C)$, $h_D^\alpha(C)$中每一个蕴含项为一个分布保持的约简。

例 3 如表 1 所示的区间值决策系统, 令$\alpha=0.6$, 根据例 2 可知相似布尔矩阵以及相容类集合。

计算每个对象对应的概率分布:

$\mu_C^{0.6}(x_1)=\{1,0\}$, $\mu_C^{0.6}(x_2)=\{0,1\}$,

$\mu_C^{0.6}(x_3)=\{2/3,1/3\}$, $\mu_C^{0.6}(x_4)=\{0,1\}$,

$\mu_C^{0.6}(x_5)=\{2/3,1/3\}$, $\mu_C^{0.6}(x_6)=\{2/3,1/3\}$。

计算分布保持约简可辨识矩阵:

$$\mathbf{DM}_D^{0.6}(6,6)=\begin{bmatrix}\varnothing & & & & & \\ a_1,a_2,a_3,a_4 & \varnothing & & & & \\ a_1,a_2 & a_3,a_4 & \varnothing & & & \\ a_1,a_2,a_3 & \varnothing & a_3 & \varnothing & & \\ a_1,a_2 & a_3,a_4 & \varnothing & a_3 & \varnothing & \\ a_1,a_2 & a_3,a_4 & \varnothing & a_3 & \varnothing & \varnothing\end{bmatrix}$$

计算分布约简可辨识函数:

$$f_D^{0.6}(C)(\bar{a}_1,\bar{a}_2,\bar{a}_3,\bar{a}_4)=a_3\wedge(a_2\vee a_1)$$

转化后的分布约简可辨识函数为

$$h_D^{0.6}(C)(\bar{a}_1,\bar{a}_2,\bar{a}_3,\bar{a}_4)=(a_1\wedge a_3)\vee(a_2\wedge a_3)$$

因此分布保持约简为$\{a_1,a_3\}$和$\{a_2,a_3\}$。

2 区间值决策系统的最大分布约简

本节在区间值决策系统中引入最大规则置信度的概念, 提出了不协调区间值决策系统的最大分布约简算法。

定义 13 设区间值决策系统$\mathrm{DS}=(U,C\cup D,V,f)$, $A\subseteq C$, $U/D=\{D_1,D_2,\cdots,D_{|U/D|}\}$, 则$x_i\in U$对应的最大概率分布定义为

$$m_A^\alpha(x_i)=D(D_{j_0}/S_A^\alpha(x_i))=\max_{j\leqslant q}D(D_j/S_A^\alpha(x_i))$$

$x_i\in U$对应的最大分布为

$$\gamma_A^\alpha(x_i)=\{D_j|D(D_j/S_A^\alpha(x_i))=D(D_{j_0}/S_A^\alpha(x_i))\}$$

若对任意的 $x_i\in U$, 有$\gamma_A^\alpha(x_i)=\gamma_C^\alpha(x_i)$, 称 A 是 DS 中基于α-相容关系的最大分布协调集, 简称最大分布协调集。若 A 是最大分布协调集, 且 A 的任意真子集都不是最大分布协调集, 那么称 A 是 DS 中基于α-相容关系的最大分布相对约简, 简称最大分布约简。

定义 14 设区间值决策系统$\mathrm{DS}=(U,C\cup D,V,f)$, $A\subseteq C$, $x_i\in U$, 若属性子集 A 满足:

1) $\gamma_A^\alpha(x_i)=\gamma_C^\alpha(x_i)$;

2) 任意$B\subset A$, 满足$\gamma_B^\alpha(x_i)\neq\gamma_C^\alpha(x_i)$;

那么称属性子集 A 为区间值决策系统基于相容关系的最大分布约简。DS 的所有约简集合记为α-Reduct, 所有约简的交集称为 DS 的核, 记为α-Core。

定理 2 设区间值决策系统$\mathrm{DS}=(U,C\cup D,V,f)$, $A\subseteq C$, 则 A 是最大分布协调集当且仅当任意$x_i,x_j\in U$, 当$\gamma_C^\alpha(x_i)=\gamma_C^\alpha(x_j)$, 有$S_A^\alpha(x_i)\neq S_A^\alpha(x_j)$。

证 明 记 $J(S_A^\alpha(x_i))=\{S_C^\alpha(x_j)|S_C^\alpha(x_j)\subseteq S_A^\alpha(x_i)\}$。

“$\Rightarrow$”: 设 A 是最大分布协调集, 对任意 x_i, $x_j\in U$,

假设$S_A^\alpha(x_i)=S_A^\alpha(x_j)$, 有$(x_i,x_j)\in \boldsymbol{T}_A^\alpha$, 即$\gamma_A^\alpha(x_i)=\gamma_A^\alpha(x_j)$, 又因为$\gamma_A^\alpha(x_i)=\gamma_C^\alpha(x_i)$和$\gamma_A^\alpha(x_j)=\gamma_C^\alpha(x_j)$成立, 那么$\gamma_C^\alpha(x_i)=\gamma_C^\alpha(x_j)$, 这与$\gamma_C^\alpha(x_i)\neq\gamma_C^\alpha(x_j)$矛盾, 从而任意$x_i,x_j\in U$, 当$\gamma_C^\alpha(x_i)=\gamma_C^\alpha(x_j)$时, 有$S_A^\alpha(x_i)\neq S_A^\alpha(x_j)$。

"$\Leftarrow$": 对任意$x_i,x_j\in U$, 当$S_A^\alpha(x_i)=S_A^\alpha(x_j)$, 有$\gamma_C^\alpha(x_i)=\gamma_C^\alpha(x_j)$, 对于任意的$D_{j_0}\in\gamma_C^\alpha(x_i)$, 有$D_{j_0}\in\gamma_C^\alpha(x_j)$。由于$S_A^\alpha(x_i)=\cup\{S_C^\alpha(x_j)|S_C^\alpha(x_j)\in J(S_A^\alpha(x_i))\}$, 于是对任意的$k\leqslant q$, 有

$$D(D_k/S_A^\alpha(x_i))=$$
$$\frac{\sum\{|D_k\cap S_C^\alpha(x_j)|:S_C^\alpha(x_j)\in J(S_A^\alpha(x_i))\}}{|S_A^\alpha(x_i)|}=$$
$$\sum\left\{\frac{|D_k\cap S_C^\alpha(x_j)|}{|S_C^\alpha(x_j)|}\times\frac{|S_C^\alpha(x_j)|}{|S_A^\alpha(x_i)|}:S_C^\alpha(x_j)\in J(S_A^\alpha(x_i))\right\}\leqslant$$
$$\sum\left\{\frac{|D_{j_0}\cap S_C^\alpha(x_j)|}{|S_C^\alpha(x_i)|}\times\frac{|S_C^\alpha(x_j)|}{|S_A^\alpha(x_i)|}:S_C^\alpha(x_j)\in J(S_A^\alpha(x_i))\right\}=$$
$$\frac{|D_{j_0}\cap S_A^\alpha(x_i)|}{|S_A^\alpha(x_i)|}=D(D_{j_0}/S_A^\alpha(x_i))$$

故$D_{j_0}\in\gamma_C^\alpha(x_i)$, 从而$\gamma_A^\alpha(x_i)=\gamma_C^\alpha(x_i)$。

另一方面, 任意的$D_{j_0}\in\gamma_A^\alpha(x_i)$, 若$D_{j_0}\neq\gamma_C^\alpha(x_i)$, 则任意的$S_C^\alpha(x_j)\in J(S_A^\alpha(x_i))$, 由$\gamma_C^\alpha(x_i)=\gamma_C^\alpha(x_j)$可得$m_C^\alpha(x_j)>D(D_{j_0}/S_C^\alpha(x_j))$。取$D_{k_0}\in\gamma_C^\alpha(x_j)$, 则

$$D(D_{k_0}/S_A^\alpha(x_i))=$$
$$\sum\left\{\frac{|D_{k_0}\cap S_C^\alpha(x_j)|}{|S_C^\alpha(x_j)|}\times\frac{|S_C^\alpha(x_j)|}{S_A^\alpha(x_i)}:S_C^\alpha(x_j)\in J(S_A^\alpha(x_i))\right\}=$$
$$\sum\left\{m_C^\alpha(x_j)\times\frac{|S_C^\alpha(x_j)|}{|S_A^\alpha(x_i)|}:S_C^\alpha(x_j)\in J(S_A^\alpha(x_i))\right\}>$$
$$\sum\left\{D(D_{j_0}/S_C^\alpha(x_j))\times\frac{|S_{\mathrm{AT}}^\alpha(x_j)|}{|S_A^\alpha(x_i)|}:S_C^\alpha(x_j)\in J(S_A^\alpha(x_i))\right\}=$$
$$\sum\left\{\frac{|D_{j_0}\cap S_C^\alpha(x_j)|}{|S_C^\alpha(x_j)|}\times\frac{|S_C^\alpha(x_j)|}{|S_A^\alpha(x_i)|}:S_C^\alpha(x_j)\in J(S_A^\alpha(x_i))\right\}=$$
$$\frac{|D_{j_0}\cap S_A^\alpha(x_i)|}{|S_A^\alpha(x_i)|}=D(D_{j_0}/S_A^\alpha(x_i))$$

与$D_{j_0}\in\gamma_A^\alpha(x_i)$矛盾, 因此$D_{j_0}\in\gamma_C^\alpha(x_i)$, 于是有$\gamma_A^\alpha(x_i)\subseteq\gamma_C^\alpha(x_i)$。

因此, 证明了对任意$x_i\in U$, $\gamma_A^\alpha(x_i)=\gamma_C^\alpha(x_i)$, 即集合$A$是最大分布协调集。定理得证。

定义 15　设区间值决策系统$\mathrm{DS}=(U,C\cup D,V,f)$, $U=\{x_1,x_2,\cdots,x_{|U|}\}$, 则对任意$i\geqslant 1, j\leqslant|U|$:

$$\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)=\begin{cases}\{a_k|a_k\in C\wedge\alpha_{ij}^k<\alpha\}, & \gamma_A^\alpha(x_i)\neq\gamma_A^\alpha(x_j)\\ \varnothing, & \gamma_A^\alpha(x_i)=\gamma_A^\alpha(x_j)\end{cases}$$

$\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)$为基于$\alpha$-相容类的最大分布约简可辨识矩阵$\mathbf{DM}_{D\mathrm{Max}}^\alpha$第$i$行$j$列的元素, $\mathbf{DM}_{D\mathrm{Max}}^\alpha$简称为最大分布可辨识矩阵, 其中$i,j=1,2,\cdots,|U|$, $\varnothing$表示空集。

基于α-相容类的最大分布可辨识矩阵是一个相对于主对角线对称的矩阵, 在进行运算时只需考虑其上三角或下三角部分即可。

定理 3　设区间值决策系统$DS=(U,C\cup D,V,f)$, $A\subseteq C$, 则A是最大分布协调集当且仅当任意$x_i,x_j\in U$, 当$\gamma_C^\alpha(x_i)\neq\gamma_C^\alpha(x_j)$时, 有$\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)\cap A\neq\varnothing$。

证明　"$\Rightarrow$": 设A是最大分布协调集, 对于任意的$x_i,x_j\in U$, 假设存在$\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)$使$\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)\cap A\neq\varnothing$, 则存在$S_C^\alpha(x_i)$和$S_C^\alpha(x_j)$, 有$\gamma_C^\alpha(x_i)\neq\gamma_C^\alpha(x_j)$, 由定理 1 得$S_A^\alpha(x_i)\neq S_A^\alpha(x_j)$, 从而存在$a_k\in A$, 满足$\alpha_{ij}^k<\alpha$, 因此存在$a_k\in\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)$, 即$\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)\cap A\neq\varnothing$。

"$\Leftarrow$": 假设存在$x_i,x_j\in U$, 满足$\gamma_C^\alpha(x_i)\neq\gamma_C^\alpha(x_j)$, 且$\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)\cap A\neq\varnothing$, 则对任意$a_k\in A$, 有$a_k\notin\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)$, $\alpha_{ij}^k>\alpha$, 因此$(x_i,x_j)\in\boldsymbol{T}_A^\alpha$。假设$x_i$、$x_j$对应的$\alpha$-相容类分别为$S_A^\alpha(x_i)$和$S_A^\alpha(x_j)$, 则有$S_A^\alpha(x_i)=S_A^\alpha(x_j)$, 由定理 1 得$A$不是最大分布协调集。定理得证。

定义 16　设区间值决策系统$\mathrm{DS}=(U,C\cup D,V,f)$, $C=\{a_1,a_2,\cdots,a_{|C|}\}$, $\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)$表示最大分布可辨识矩阵中第$i$行$j$列的元素, 基于$\alpha$-相容类的最大分布可辨识函数为与$a_1,a_2,\cdots,a_m$相对应$|C|$个布尔变量$\bar{a}_{|C|}$的布尔函数: $f_D^\alpha(C)_{\mathrm{Max}}(\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_{|C|})=\wedge\{\vee\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j):\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)\neq\varnothing\}$, 为基于$\alpha$-相容类的最大分布约简可辨识函数, 简称最大分布可辨识函数。$\vee\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)$表示满足$a\in\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)$的全体布尔变量$\bar{a}$的析取式。

利用分配率和吸收率将$f_D^\alpha(C)_{\mathrm{Max}}$转化为$(\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_{|C|})=(\wedge\theta_1)\vee\cdots\vee(\theta_l)$, $\theta_k\subseteq C, k=(\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_{|C|})=(\wedge\theta_1)\vee\cdots\vee(\theta_l),\theta_k\subseteq C,k=1,2,\cdots,l$, θ_k中每一个属性元素只出现一次。

定理 4　设区间值决策系统$DS=(U,C\cup D,V,f)$, $h_D^\alpha(C)_{\mathrm{Max}}$是可辨识函数$f_D^\alpha(C)_{Max}$的形式转化, 若$A$是最大分布约简, 当且仅当$A$是$h_D^\alpha(C)_{\mathrm{Max}}$的一个蕴含项。

证明　"$\Rightarrow$": 假设θ是$h_D^\alpha(C)_{\mathrm{Max}}$的一个蕴含项, 则存在$\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)\cap\theta\neq\varnothing$, 通过定理 2 得知$\theta$是其中一个最大分布约简。

"$\Leftarrow$": 根据定义 16 可得$h_D^\alpha(C)_{\mathrm{Max}}(\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_m)=(\wedge\theta_1)\vee\cdots\vee(\theta_l),\theta_k\subseteq C,k=1,2,\cdots,l$, 若在$\theta$中去掉一个元素形成$\theta'$, 则存在$S_C^\alpha(x_i)$和$S_C^\alpha(x_j)$满足$\gamma_C^\alpha(x_i)\neq\gamma_C^\alpha(x_j)$, 使得$\mathrm{DM}_{D\mathrm{Max}}^\alpha(i,j)\cap\theta'\neq\varnothing$, 故$\theta'$不是最大分布约简, 从而$\theta$是其中一个最大分布约简。定理得证。

基于差别矩阵的分布约简算法 (maximum distribution reduction algorithm based on discernibility matrix, MDRADM) 描述如算法 2。

算法 2　MDRADM

输入　区间值决策系统 DS, 阈值α。

输出　区间值决策系统的所有最大分布保持约简结果。

1) 计算区间值决策系统 DS 在阈值α下的相容类集合$S_C^\alpha(U)$。

2) 根据每个对象对应的相容类，计算每个对象相对于每一个决策类的概率分布$\mu_C^{\alpha}(x_i)$。

3) 根据每个对象的概率分布，计算所对应的最大分布$\gamma_C^{\alpha}(x_i)$。

4) 根据每个对象的可信度不同构造最大分布约简可辨识矩阵$\mathbf{DM}_{D\mathrm{Max}}^{\alpha}$。

5) 由可辨识矩阵$\mathbf{DM}_{D\mathrm{Max}}^{\alpha}$计算最大分布约简可辨识函数$f_D^{\alpha}(C)_{\mathrm{Max}}$。

6) 利用分配率和吸收率将U转化为$h_D^{\alpha}(C)_{\mathrm{Max}}$，$h_D^{\alpha}(C)_{\mathrm{Max}}$中每一个蕴含项为一个最大分布保持的约简。

算法 2 是通过可辨识矩阵求得区间值决策表的所有最大分布保持约简，因此算法在最坏情况下的时间复杂度为$O(|C|^{|U|^2})$，$|C|$为条件属性的个数，$|U|$为对象的个数。

例 4　如表 1 所示的区间值决策系统，令$\alpha=0.6$，根据例 2 可知相似布尔矩阵以及相容类。

计算决策属性D对U划分：

$$U/D=\{D_1,D_2\}=\{\{x_1,x_5,x_6\},\{x_2,x_3,x_4\}\}。$$

计算每个对象对应的概率分布：

$$\mu_C^{0.6}(x_1)=\{1,0\},\ \mu_C^{0.6}(x_2)=\{0,1\},$$
$$\mu_C^{0.6}(x_3)=\{2/3,1/3\},\ \mu_C^{0.6}(x_4)=\{0,1\},$$
$$\mu_C^{0.6}(x_5)=\{2/3,1/3\},\ \mu_C^{0.6}(x_6)=\{2/3,1/3\}。$$

计算每个对象对应的最大分布：

$$\gamma_C^{0.6}(x_1)=\{D_1\},\ \gamma_C^{0.6}(x_2)=\{D_2\},$$
$$\gamma_C^{0.6}(x_3)=\{D_1\},\ \gamma_C^{0.6}(x_4)=\{D_2\},$$
$$\gamma_C^{0.6}(x_5)=\{D_1\},\ \gamma_C^{0.6}(x_6)=\{D_1\}。$$

计算最大分布约简可辨识矩阵：

$$\mathbf{DM}_{D\mathrm{Max}}^{0.6}(6,6)=\begin{bmatrix}\varnothing & & & & & \\ a_1,a_2,a_3,a_4 & \varnothing & & & & \\ \varnothing & a_3,a_4 & \varnothing & & & \\ a_1,a_2,a_3 & \varnothing & a_3 & \varnothing & & \\ \varnothing & a_3,a_4 & \varnothing & a_3 & \varnothing & \\ \varnothing & a_3,a_4 & \varnothing & a_3 & \varnothing & \varnothing\end{bmatrix}$$

计算最大分布约简可辨识函数：

$$f_D^{0.6}(C)_{\mathrm{Max}}(\bar{a}_1,\bar{a}_2,\bar{a}_3,\bar{a}_4)=a_3$$

因此，最大分布保持约简结果为$\{a_3\}$。

例 5　如表 1 所示的区间值决策系统，令α分别为0.4,0.5,0.6,0.7，则分布保持约简结果为

$$h_D^{0.4}(C)(\bar{a}_1,\bar{a}_2,\bar{a}_3,\bar{a}_4)=a_1\vee a_4$$
$$h_D^{0.5}(C)(\bar{a}_1,\bar{a}_2,\bar{a}_3,\bar{a}_4)=(a_1\wedge a_3)\vee(a_2\wedge a_3)$$
$$h_D^{0.6}(C)(\bar{a}_1,\bar{a}_2,\bar{a}_3,\bar{a}_4)=(a_1\wedge a_3)\vee(a_2\wedge a_3)$$
$$h_D^{0.7}(C)(\bar{a}_1,\bar{a}_2,\bar{a}_3,\bar{a}_4)=(a_1\wedge a_3)\vee(a_2\wedge a_3)$$

最大分布保持约简结果为

$$h_D^{0.4}(C)_{\mathrm{Max}}(\bar{a}_1,\bar{a}_2,\bar{a}_3,\bar{a}_4)=a_1\vee a_4$$
$$h_D^{0.5}(C)_{\mathrm{Max}}(\bar{a}_1,\bar{a}_2,\bar{a}_3,\bar{a}_4)=a_3$$
$$h_D^{0.6}(C)_{\mathrm{Max}}(\bar{a}_1,\bar{a}_2,\bar{a}_3,\bar{a}_4)=a_3$$
$$h_D^{0.7}(C)_{\mathrm{Max}}(\bar{a}_1,\bar{a}_2,\bar{a}_3,\bar{a}_4)=a_3$$

性质 3　设区间值决策系统$\mathrm{DS}=(U,C\cup D,V,f)$，$H=h_1\vee h_2\vee\cdots\vee h_m$和$K=k_1\vee k_2\vee\cdots\vee k_n$分别是分布约简和最大分布约简结果，则在阈值$\alpha$下，对于$K$中任意一个蕴含项$k_j$，$H$中存在一个蕴含项$h_i$满足$h_i\supseteq k_j$。

3　实验验证与分析

本节对提出的最大分布约简算法进行实验验证，实验包括两部分：1) 比较最大分布约简方法和其他约简方法的约简结果，验证了性质 3 的正确性；2) 比较了最大分布保持、分布保持和正域保持 3 种约简算法的约简效率。采用 UCI 标准测试集进行实验。实验环境为 PC 机，操作系统为 Windows 7 旗舰版 64 位；内存为 6.0 GB DDR3，CPU 为 Intel i5-3470。

实验选取 8 组标准 UCI 数据集，对缺失数据通过将对应属性下占多数属性值进行替换，对名词性数据采用{0,1}替换，对连续型数据采用等频分割[19]的方法，所有数据预处理均在 WEKA3.6 进行，数据集信息如表 2 所示，$|U|$表示对象数，|AT|表示条件属性数，$|D|$表示决策属性将对象分类个数。

表 2　UCI 数据集信息

Table 2　UCI data sets information

数据集	$\|U\|$	|AT|	$\|D\|$
BLOGGER	100	4	2
Fertility	100	9	2
Teaching Assistant Evaluation	151	4	3
QualitativeBankruptcy	250	6	2
User Knowledge Modeling	258	4	4
Liverdisorders	345	6	2
Auto MPG	398	6	3
Mammographic Mass	961	4	2

由于表格有限，表 3 ~ 10 中数据集名称均为相应数据集名称的缩写。

由于采用的 UCI 数据集都是单值数据，因此需将单值数据转换为区间值数据，单值数据转换为区间值数据的方法在文献[19]中已经描述，先将该方法改进，引进阈值λ，该值可调节振幅，即区间值的长度。

表 3　约简结果对比 (λ= 2.4,α= 0.4)
Table 3　Comparison of reduction results (λ= 2.4,α= 0.4)

数据集	PRADM	DRADM	MDRADM
BLO	{1, 2, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
FER	{1, 3, 7, 8, 9}	{1, 2, 3, 6, 7, 8, 9}	{1, 3, 7, 9}
TAE	{4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
QB	集合 1	{1, 2, 3, 4, 5, 6}	{1, 2, 6}
UKM	集合 2	{1, 2, 4, 5}	{1, 2, 4, 5}
LD	{5, 6}	{1, 2, 3, 4, 5, 6}	{1, 2, 3, 4, 5, 6}
AM	集合 3	{1, 2, 3, 4, 5, 6, 7}	{1, 2, 3, 4, 5, 6, 7}
MM	{1}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}

表 4　约简结果对比 (λ= 2.4,α= 0.5)
Table 4　Comparison of reduction results(λ= 2.4,α= 0.5)

数据集	PRADM	DRADM	MDRADM
BLO	{1, 2, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
FER	集合 4	集合 5	集合 6
TAE	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
QB	集合 7	{1, 2, 3, 4, 5, 6}	{1, 2, 6}
UKM	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 4, 5}
LD	{1, 2, 3, 5, 6}	{1, 2, 3, 4, 5, 6}	{1, 2, 3, 4, 5, 6}
AM	{1, 3, 7}	{1, 3, 4, 5, 6, 7}	{1, 3, 4, 5, 6, 7}
MM	{12, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}

表 5　约简结果对比 (λ= 2.4,α= 0.6)
Table 5　Comparison of reduction results (λ= 2.4,α= 0.6)

数据集	PRADM	DRADM	MDRADM
BLO	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
FER	集合 8	集合 9	集合 10
TAE	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
QB	{1, 5, 6}	{1, 2, 3, 4, 5, 6}	{1, 2, 3, 4, 5, 6}
UKM	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
LD	{1, 2, 3, 4, 5, 6}	{1, 2, 3, 4, 5, 6}	{1, 2, 3, 4, 5, 6}
AM	{1, 2, 3, 4, 5, 6, 7}	{1, 2, 3, 4, 5, 6, 7}	{1, 2, 3, 4, 5, 6, 7}
MM	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}

设区间值决策系统$(U, C\cup D, V, f)$, 对任意的$x_i\in U$, $a_t(x_i)$为x_i在属性 t 上的取值$U/D=\{D_1, D_2, \cdots, D_{|U/D|}\}$, $D_k\in U/D$, 则单值性数据转换为区间值数据的振幅为

$$\sigma_t^k=\sqrt{\frac{1}{|D_k|-1}\sum_{x_i\in D_k}(a_t(x_j)-\bar{a}_t^k)^2}$$

式中$\bar{a}_t^k=\dfrac{\sum\limits_{x_j\in D_k}a_t(x_j)}{|D_k|}$。

表 6　约简结果对比 (λ= 2.4,α= 0.7)
Table 6　Comparison of reduction results (λ= 2.4,α= 0.7)

数据集	PRADM	DRADM	MDRADM
BLO	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
FER	{3}	{3}	{3}
TA	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
QB	集合 11	集合 12	集合 13
UKM	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
LD	{1, 2, 3, 4, 5, 6}	{1, 2, 3, 4, 5, 6}	{1, 2, 3, 4, 5, 6}
AM	{1, 2, 3, 4, 5, 6, 7}	{1, 2, 3, 4, 5, 6, 7}	{1, 2, 3, 4, 5, 6, 7}
MM	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}

表 7　约简结果对比 (λ= 3.5,α= 0.4)
Table 7　Comparison of reduction results (λ= 3.5,α= 0.4)

数据集	PRADM	DRADM	MDRADM
BLO	集合 14	{1, 3, 4}	{4}
FER	集合 15	{2, 6, 7, 9}	{2, 6, 7, 9}
TAE	集合 16	{2, 3, 5}	{2, 3, 5}
QB	集合 17	{2, 6}	{2, 6}
UKM	集合 18	{5}	{5}
LD	集合 19	{1, 2, 3, 4, 5, 6}	{6}
AM	集合 20	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
MM	{1}	{1, 2, 4, 5}	{1, 2, 4, 5}

表 8　约简结果对比 (λ= 3.5,α= 0.5)
Table 8　Comparison of reduction results (λ= 3.5,α= 0.5)

数据集	PRADM	DRADM	MDRADM
BLO	集合 21	{1, 3, 4, 5}	{1, 3, 4, 5}
FER	{1, 7, 8, 9}	{1, 2, 3, 6, 7, 8, 9}	{1, 3, 6, 8, 9}
TAE	{4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
QB	集合 22	{2, 6}	{2, 6}
UKM	集合 23	{2, 5}	{2, 5}
LD	{6}	{1, 2, 3, 4, 5, 6}	{1, 2, 3, 4, 5, 6}
AM	集合 24	{1, 2, 3, 4, 5, 6, 7}	{1, 2, 3, 4, 5, 6, 7}
MM	{1}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}

区间值的左右区间分别为

$$l_i^t=a_t(x_i)-\lambda\bar{a}_t^k$$
$$u_i^t=a_t(x_i)+\lambda\bar{a}_t^k$$

式中λ为调节区间值长度的值。

3.1　约简结果对比

在本节中, 讨论了最大分布约简与其他约简方法之间的关系[20], 选取正域保持约简算法 (PRADM)[17]和分布保持约简算法 (DRADM)。λ分别取2.4和

3.5, α分别取0.4、0.5、0.6、0.7, 共进行了 8 组实验, 实验结果如表 3~10 所示, 其中: 集合 1=集合 7=集合 17=集合 19=集合 22=集合 28={{1}, {2}, {3}, {4}, {5}, {6}}; 集合 2=集合 14=集合 16=集合 18=集合 21=集合 23={{1}, {2}, {3}, {4}, {5}}; 集合 3=集合 20=集合 24={{1}, {2}, {3}, {4}, {5}, {6}, {7}}; 集合 4=集合 6={1, 3, 4, 5, 6, 7, 8, 9}; 集合 5=集合 8=集合 9=集合 10=集合 26=集合 27={1, 2, 3, 4, 5, 6, 7, 8, 9}; 集合 11=集合 12=集合 13={{1, 3, 5},{2, 3, 5},{3, 4, 5},{1, 2, 3, 4, 6},{1, 5, 6},{2, 5, 6},{3, 5, 6}{4, 5, 6}}; 集合 15={{1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}}; 集合 25={1, 2, 3, 4, 5, 7, 8, 9}。

表 9　约简结果对比 (λ= 3.5,α= 0.6)

Table 9　Comparison of reduction results (λ= 3.5,α= 0.6)

数据集	PRADM	DRADM	MDRADM
BLO	{1, 2, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
FER	集合 25	集合 26	集合 27
TA	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
QB	集合 28	{1, 2, 3, 4, 5, 6}	{1, 2, 6}
UKM	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
LD	{1, 2, 3, 5, 6}	{1, 2, 3, 4, 5, 6}	{1, 2, 3, 4, 5, 6}
AM	{1, 2, 3, 4, 5, 6, 7}	{1, 2, 3, 4, 5, 6, 7}	{1, 2, 3, 4, 5, 6, 7}
MM	{1, 2, 3, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}

表 10　约简结果对比 (λ= 3.5,α= 0.7)

Table 10　Comparison of reduction results (λ= 3.5,α= 0.7)

数据集	PRADM	DRADM	MDRADM
BLO	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
FER	{3}	{3}	{3}
TAE	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
QB	{1, 5, 6}	{1, 2, 3, 4, 5, 6}	{1, 2, 3, 4, 5, 6}
UKM	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
LD	{1, 2, 3, 4, 5, 6}	{1, 2, 3, 4, 5, 6}	{1, 2, 3, 4, 5, 6}
AM	{1, 2, 3, 4, 5, 6, 7}	{1, 2, 3, 4, 5, 6, 7}	{1, 2, 3, 4, 5, 6, 7}
MM	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}

表 3 ~ 6 为λ取2.4, α分别取 0.4、0.5、0.6、0.7 时, 正域保持、分布保持和最大分布保持约简算法的约简结果。实验结果表明, MDRADM 约简结果为 DRADM 约简结果的子集, 即验证了性质 3 的正确性, 而 PRADM 约简结果和 MDRADM 约简结果没有明显关系。这是因为, 当正域为空时, 正域约简结果为条件属性中任意一个属性, 故 PRADM 的约简结果和 MDRADM 的约简结果不存在包含关系。当λ= 3.5,α= 0.4时, 对于大部分数据集, DRADM 的约简结果最短, Fertility 数据集则在λ= 2.4, α= 0.7时最短, 但 Liverdisorders 数据集在任何阈值下均没有冗余属性。

3.2　约简效率对比

本节选取 Mammographic Mass 数据集, 对比两个算法随对象数量的增加耗时变化情况。图 1 ~ 5 为λ取2.4, α分别取 0.4、0.5、0.6、0.7、0.8 时, 3 个算法的时间耗费情况; 横坐标表示 Mammographic Mass 数据集的对象数量, 纵坐标表示运行时间, 单位为 s。

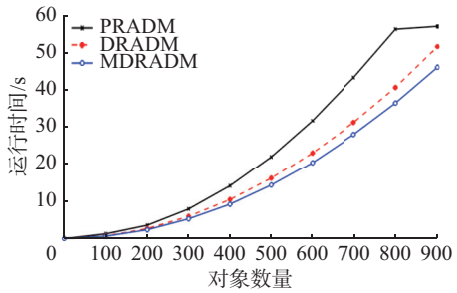

图 1　约简效率对比 (α= 0.4)

Fig. 1　Comparison of reduction efficiency (α= 0.4)

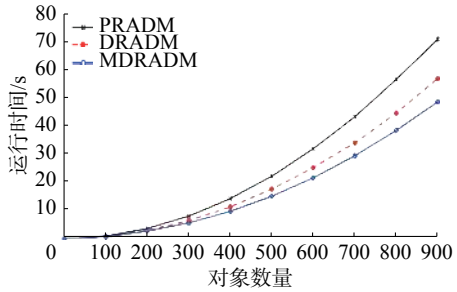

图 2　约简效率对比 (α= 0.5)

Fig. 2　Comparison of reduction efficiency (α= 0.5)

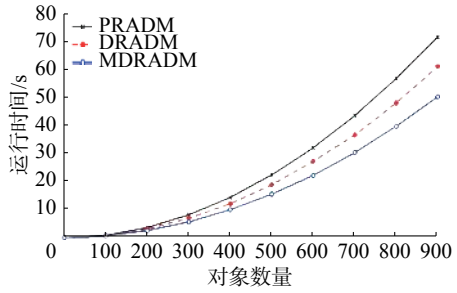

图 3　约简效率对比 (α= 0.6)

Fig. 3　Comparison of reduction efficiency (α= 0.6)

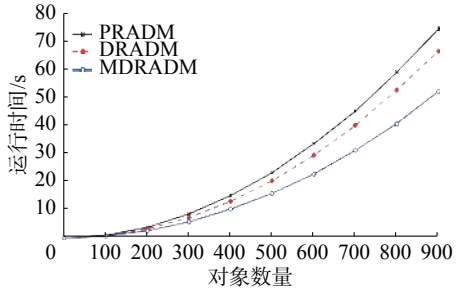

图 4　约简效率对比 (α= 0.7)

Fig. 4　Comparison of reduction efficiency (α= 0.7)

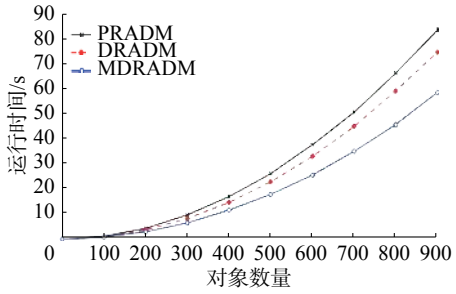

图 5　约简效率对比 (α= 0.8)

Fig. 5　Comparison of reduction efficiency (α= 0.8)

图 1 ~ 5 中虚线表示 PRADM 随着对象数量增加运行时间变化曲线, 空心圆点实线表示 MDRADM 随着对象数量增加运行时间变化曲线, 交叉点实线表示 DRADM 随着对象数量增加运行时间变化曲线。实验结果表明, 在对象数较少情况下, 由于差别矩阵较简单, PRADM、DRADM 和 MDRADM 运行时间几乎没有差别, 但随着对象数量的增加, 3 种算法的运行时间差异越来越明显; 由于 MDRADM 差别元素是 DRADM 差别元素的一个子集, PRADM 的差别矩阵为非对称矩阵, 故 MDRADM 的运行时间小于 PRADM 和 DRADM 运行时间。当α分别取 0.5、0.6、0.7、0.8 时, Mammographic Mass 数据集随着对象的增加, 3 个算法的耗时差距增大, 这是由于随着对象的增加差别矩阵愈加复杂, 计算量越大造成的; 当α = 0.4时, 也呈现这样的趋势, 但当对象数达到 900 时, 利用吸收率和结合律运算的差别矩阵较简单, 造成时间增长率减小。当λ取3.5, α分别取 0.4、0.5、0.6、0.7、0.8 时, 3 个算法的时间耗费情况跟λ取2.4时的折线图大致相同, 所以本文不作详细描述。

4　结束语

属性约简是粗糙集理论研究的热点问题之一, 在实际应用中具有重要意义, 主要作用有: 1) 提取更加泛化的规则; 2) 针对应用中的海量数据, 能够压缩数据集规模。分布保持约简能够保持信息系统在约简前后置信度不变, 而人们往往只关注置信度最大的规则, 具有广泛的应用价值。

本文在相关研究成果的基础上, 在不协调区间值决策系统中提出最大分布约简的概念, 构造了基于可辨识矩阵的最大分布约简算法, 该算法保持了在知识约简前后各个规则的最大置信度不变。实验选取 8 组 UCI 数据集将本文算法与已有的两种约简算法的约简结果和效率进行对比。实验结果表明, 分布约简包含最大分布约简, 并且最大分布约简算法比其他两种算法具有更高的效率。由于本文提出的算法是在可辨识矩阵基础上的, 其时间和空间复杂度较高, 不利于在实际应用中推广, 故提出高效率的约简算法是未来研究方向之一。

参考文献:

[1] PAWLAK Z. Rough sets[J]. International journal of computer & information sciences, 1982, 11(5): 341–356.

[2] PAWLAK Z. Rough sets: theoretical aspects of reasoning about data[M]. Boston: Kluwer Academic Publishers, 1992.

[3] 王国胤, 姚一豫, 于洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009, 32(7): 1229–1246.
WANG Guoyin, YAO Yiyu, YU Hong. A survey on rough set theory and applications[J]. Chinese journal of computers, 2009, 32(7): 1229–1246.

[4] QIAN Yuhua, LIANG Jiye, PEDRYCZ W, et al. Positive approximation: an accelerator for attribute reduction in rough set theory[J]. Artificial intelligence, 2010, 174(9/10): 597–618.

[5] WANG Feng, LIANG Jiye, QIAN Yuhua. Attribute reduction: A dimension incremental strategy[J]. Knowledge-based systems, 2013, 39: 95–108.

[6] CHEN Hongmei, LI Tianrui, RUAN Da, et al. A rough-set based incremental approach for updating approximations under dynamic maintenance environments[J]. IEEE transactions on knowledge and data engineering, 2013, 25(2): 274–284.

[7] HU Qinghua, YU Daren, XIE Zongxia. Information-preserving hybrid data reduction based on fuzzy-rough techniques[J]. Pattern recognition letters, 2006, 27(5): 414–423.

[8] SKOWRON A, RAUSZER C. The discernibility matrices and functions in information systems[M]//SŁOWIŃSKI R. Intelligent Decision Support. Dordrecht: Springer, 1992, 11: 331–362.

[9] KRYSZKIEWICZ M. Rough set approach to incomplete information systems[J]. Information sciences, 1998, 112(1/2/3/4): 39–49.

[10] 邓大勇, 黄厚宽, 李向军. 不一致决策系统中约简之间的比较[J]. 电子学报, 2007, 35(2): 252–255.
DENG Dayong, HUANG Houkuan, LI Xiangjun. Comparison of various types of reductions in inconsistent systems[J]. Acta electronica sinica, 2007, 35(2): 252–255.

[11] MIAO Duoqian, ZHAO Yan, YAO Yiyu, et al. Relative reducts in consistent and inconsistent decision tables of the Pawlak rough set model[J]. Information sciences, 2009, 179(24): 4140–4150.

[12] ZHOU Jie, MIAO Duoqian, PEDRYCZ W, et al. Analysis of alternative objective functions for attribute reduction in complete decision tables[J]. Soft computing, 2011, 15(8): 1601–1616.

[13] 张文修, 米据生, 吴伟志. 不协调目标信息系统的知识约简[J]. 计算机学报, 2003, 26(1): 12–18.
ZHANG Wenxiu, MI Jusheng, WU Weizhi. Knowledge reductions in inconsistent information systems[J]. Chinese journal of computers, 2003, 26(1): 12–18.

[14] 徐伟华, 张文修. 基于优势关系下不协调目标信息系统的分布约简[J]. 模糊系统与数学, 2007, 21(4): 124–131.
XU Weihua, ZHANG Wenxiu. Distribution reduction in inconsistent information systems based on dominance relations[J]. Fuzzy systems and mathematics, 2007, 21(4): 124–131.

[15] 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.
ZHANG Nan, MIAO Duoqian, YUE Xiaodong. Approaches to knowledge reduction in interval-valued information systems[J]. Journal of computer research and development, 2010, 47(8): 1362–1371.

[16] 张楠, 许鑫, 童向荣, 等. 不协调区间值决策系统的知识约简[J]. 小型微型计算机系统, 2017, 38(7): 1585–1589.
ZHANG Nan, XU Xin, TONG Xiangrong, et al. Knowledge reduction in inconsistent interval-valued decision systems[J]. Journal of Chinese computer systems, 2017, 38(7): 1585–1589.

[17] 张楠, 许鑫, 童向荣, 等. 不协调区间值决策系统的分布约简[J]. 计算机科学, 2017, 44(9): 78–82, 104.
ZHANG Nan, XU Xin, TONG Xiangrong, et al. Distribution reduction in inconsistent interval-valued decision systems[J]. Computer science, 2017, 44(9): 78–82, 104.

[18] 刘鹏惠, 陈子春, 秦克云. 区间值信息系统的决策属性约简[J]. 计算机工程与应用, 2009, 45(28): 148–150, 229.
LIU Penghui, CHEN Zichun, QIN Keyun. Decision attribute reduction of interval-valued information system [J]. Computer engineering and applications, 2009, 45(28): 148–150, 229.

[19] ZHANG Xiao, MEI Changlin, CHEN Degang, et al. Multi-confidence rule acquisition and confidence-preserved attribute reduction in interval-valued decision systems[J]. International journal of approximate reasoning, 2014, 55(8): 1787–1804.

[20] 史德容, 徐伟华. 区间值模糊决策序信息系统的分布约简[J]. 计算机科学与探索, 2017, 11(4): 652–658.
SHI Derong, XU Weihua. Distribution reduction in interval-valued fuzzy decision ordered information systems[J]. Journal of frontiers of computer science and technology, 2017, 11(4): 652–658.

作者简介:

尹继亮, 男, 1994 年生, 硕士研究生, 主要研究方向为粗糙集、数据挖掘与机器学习。

张楠, 男, 1979 年生, 讲师, 主要研究方向为粗糙集、认知信息学与人工智能。

童向荣, 男, 1975 年生, 教授, 主要研究方向为多 Agent 系统、分布式人工智能与数据挖掘。