软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(8):1761−1780 [doi: 10.13328/j.cnki.jos.004507]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

决策域分布保持的启发式属性约简方法*

马希骜[1], 王国胤[2], 于 洪[2]

[1](西南交通大学 信息科学与技术学院,四川 成都 610031)

[2](计算智能重庆市重点实验室(重庆邮电大学),重庆 400065)

通讯作者: 王国胤, E-mail: wanggy@ieee.org, http://cs.cqupt.edu.cn/wanggy

摘 要: 在决策粗糙集中,由于引入了概率阈值,属性增加或减少时,正域或者非负域有可能变大、变小或者不变,即属性的增减与决策域(正域或非负域)之间不再具有单调性.分析结果表明,现有的基于整个决策域的属性约简定义可能会改变决策域.为使决策域保持不变,引入了正域分布保持约简与非负域分布保持约简的概念.此外,决策域的非单调性使得属性约简算法必须检查一个属性集合的所有子集.为了简化算法设计,提出了正域和非负域分布条件信息量的定义,并证明其满足单调性,从而为设计决策域分布保持约简的启发式计算方法提供了理论基础.为了进一步获得最小约简,提出一种基于遗传算法的决策域分布保持启发式约简算法,并在两种单调的决策域分布条件信息量基础上构造了新算子,即修正算子,确保遗传算法找到的是约简而不是约简的超集.对比实验从分类正确率与误分类代价两个方面都反映了决策域分布保持约简定义的合理性,并且,所提出的遗传算法在大多数情况下都找到了最小约简.

关键词: 决策粗糙集模型;决策域分布保持约简;遗传算法;属性约简;启发式方法

中图法分类号: TP181

中文引用格式: 马希骜,王国胤,于洪.决策域分布保持的启发式属性约简方法.软件学报,2014,25(8):1761−1780. http://www.jos.org.cn/1000-9825/4507.htm

英文引用格式: Ma XA, Wang GY, Yu H. Heuristic method to attribute reduction for decision region distribution preservation. Ruan Jian Xue Bao/Journal of Software, 2014,25(8):1761−1780 (in Chinese). http://www.jos.org.cn/1000-9825/4507.htm

Heuristic Method to Attribute Reduction for Decision Region Distribution Preservation

MA Xi-Ao[1], WANG Guo-Yin[2], YU Hong[2]

[1](School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

[2](Chongqing Key Laboratory of Computational Intelligence (Chongqing University of Posts and Telecommunications), Chongqing 400065, China)

Corresponding author: WANG Guo-Yin, E-mail: wanggy@ieee.org, http://cs.cqupt.edu.cn/wanggy

Abstract: In decision-theoretic rough set models, since decision regions (positive region or non-negative region) are defined by allowing some extent of misclassification, the monotonicity of decision regions with respect to attribute sets does not hold. The definition of attribute reduction based on the whole decision regions may change decision regions. In order not to change decision regions, the positive region and non-negative distribution preservation reduction are introduced into decision-theoretic rough set models. Moreover, due to the non-monotonicity of decision regions, attribute reduction algorithms must search all possible subsets of an attribute set. The positive region and non-negative region distribution condition information contents are presented to facilitate the design of heuristic algorithms for decision region distribution preservation reduction. In a bid to then solve the minimum attribute reduction problem, heuristic genetic algorithm is applied to decision region distribution preservation reduction. A new modify operator is constructed by using two kinds of decision region distribution condition information contents so that genetic algorithm can find decision region

* 基金项目: 国家自然科学基金(61272060, 61379114); 重庆市自然科学基金(CSTC2013jjB40003)
收稿时间: 2013-07-13; 定稿时间: 2013-10-11

distribution preservation reduction. Experimental results verify the effectiveness of decision region distribution preservation reduction and show the efficiency of the genetic algorithm to solve the minimum attribute reduction problem.

Key words: decision-theoretic rough set model; decision region distribution preservation reduction; genetic algorithm; attribute reduction; heuristic method

粗糙集理论[1]是由波兰数学家 Pawlak 于 1982 年提出来的.它是一种处理不确定性问题的数学工具.属性约简是粗糙集理论中的核心问题之一.通过属性约简可以减少数据冗余,从而简化决策规则.一般说来,属性约简可以理解为保持给定决策表某种性质不变的最小属性集合[2].目前,粗糙集理论已被成功应用于决策支持、机器学习、数据挖掘、图像处理等众多研究领域[3−5].

经典粗糙集理论在模拟人类智能对模糊和不确定性概念进行处理时缺乏容错能力,泛化能力不强.为了解决这个问题,学者们通过在经典粗糙集模型中引入概率包含关系提出了许多概率粗糙集模型,其中具有代表性的有决策粗糙集模型 DTRS(decision-theoretic rough set model)[6]、0.5 概率粗糙集模型[7]、变精度粗糙集模型 VPRS(variable precision rough set model)[8]和贝叶斯粗糙集模型 BRS(Bayesian rough set model)[9]等.

基于概率粗糙集模型的属性约简问题已经取得了很多成果.例如,Inuiguchi[10]讨论了变精度粗糙集中几种属性约简方法的关系,并强调了分析变精度粗糙集和经典粗糙集属性约简的不同之处;Mi 等人[11]基于变精度粗糙集提出了β下近似分布约简和β上近似分布约简的概念,并通过可辨识矩阵提供了计算β下近似分布约简和β上近似分布约简的方法;Slezak 和 Ziarko[9]提出了 Bayesian 粗糙集模型,通过用概率增益函数来评估 Bayesian 粗糙集模型的分类质量,给出了该模型下的属性约简计算方法;Zhou 和 Miao[12]研究了变精度粗糙集中β区间属性约简,提出了β区间核的概念,通过建立有序可辨识矩阵构造了获得β区间约简的启发式算法;Yao 和 Zhao[2]系统地阐述了决策粗糙集下的属性约简理论,并在研究了决策区域和决策规则的单调性以及规则的覆盖度、信任度和决策风险等各种评价标准的基础上提出了一种泛化属性约简;Jia 等人[13]提出了一种决策风险最小化的属性约简方法,并设计了计算决策风险最小化属性约简的启发式算法、模拟退火算法和遗传算法.

与其他概率粗糙集模型相比,决策粗糙集通过引入贝叶斯决策步骤,给出了用于确定正域、边界域和负域阈值的方法.在决策论框架下,所需要的阈值可以通过具体问题的损失函数来计算.Yao 等人[14]进一步分析了决策粗糙集和其他概率粗糙集之间的关系,并指出:通过设置不同的损失函数,可以推导出已有的几种概率粗糙集模型;同时,Yao[15]从语义角度出发,研究了粗糙集 3 个域在概率意义下的语义解释,即在决策粗糙集中,正域所获得的规则表示接受决策,边界域所获得的规则表示延迟决策,而负域所获得的规则表示拒绝决策.因此,决策粗糙集为概率粗糙集提供了一个统一的理论框架.

目前,关于决策粗糙集理论与应用的研究受到越来越多的关注并且取得了很多研究成果.例如,Li 和 Zhou[16]根据不同决策者不同的风险偏好,给出了基于决策粗糙集的乐观决策、悲观决策与中立决策的多角度决策模型;Herbert 和 Yao[17]提出了博弈论粗糙集,并将其应用到决策粗糙集中代价损失函数的确定中,为最优概率阈值的选定提供了一条新的途径;Yu 等人[18]提出了基于决策粗糙集的自动聚类方法,将决策粗糙集中的风险函数用于对聚类过程进行评估,以此来指导子类的合并过程;Zhou 等人[19]通过考虑不同的误分类代价,提出了多类决策粗糙集模型,并将其应用到代价敏感分类中;Qian 等人[20]通过结合决策粗糙集与多个二元关系诱导的粒结构,提出了多粒度决策粗糙集模型,为多粒度粗糙集模型提供了统一的理论框架.

在决策粗糙集中,根据不同的标准,属性约简的定义主要包括正域或非负域(本文后面统称为决策域)的定量与定性保持约简[13,21]、正域最大化属性约简[21]以及决策风险最小化属性约简[13]等.然而,由于在决策粗糙集中,决策域的确定引入了一对阈值,当属性减少时,正域或者非负域有可能变大、变小或者不变[21].也就是说,经典粗糙集中,决策域关于属性集合包含之间的单调性不再成立,这就使得决策域的定性等价与定量等价并不相等.因此导致在决策域的定量保持属性约简之后,决策域可能被改变;决策域的定性保持约简虽然不改变整个决策表的决策域,但是它并不能保证不改变每一个决策类的决策域.而正域最大化属性约简在可解释性上存在一定的困难,因为决策域的定义带有一定的不确定性.相比之下,决策风险最小化属性约简虽然更客观,但是它也有可能改变决策域.然而,用户通常并不希望改变决策域,属性约简的目的只是为了减少冗余,引入概率阈值的目

的是为了增强分类能力,而不是改变决策域.

因此,为了不改变决策域,第 2 节将提出(α,β)正域分布保持约简和(α,β)非负域分布保持约简的定义.与前文提到的几种属性约简定义[13,21]相比,决策域分布保持约简不改变每个决策类的决策域.第 5 节也将通过实验比较几种定义所获得约简的分类正确率与误分类代价,说明决策域分布保持约简是一种更好的选择.

另一方面,由于概率粗糙集中决策域与属性的增减之间并不存在单调性,为了获得约简,必须检查整个属性集合的所有子集,这为算法设计带来了很大的困难.为了有效降低算法设计的难度,具有单调性的启发式函数是必要的.值得注意的是,在概率粗糙集的相关研究中,并没有单调的度量函数用于属性约简,因此,大多数启发式算法实际上找到的可能是一个约简的超集(包含一个约简的属性集合).

因此,第 3 节将提出两种决策域分布条件信息量:(α,β)正域分布条件信息量和(α,β)非负域分布条件信息量,并证明其单调性.它们分别被用作计算(α,β)正域分布保持约简和(α,β)非负域分布保持约简的启发式信息.在此基础上将给出两种决策域分布保持约简核属性的定义以及求核算法.

给定一个决策表,通常存在多个约简,为了得到简洁而有效的决策规则,获得最小约简非常重要,而获取最小约简已被证明是一个 NP-hard 问题[22].目前,属性约简算法主要有两种[23]:一种是启发式的贪心算法,另一种是基于种群的随机优化方法.前者速度较快,但是通常只能找到约简而不是最小约简;后者比起前者找到最小约简的概率更大.在基于种群的随机优化方法中,通常将最小约简问题转化为组合优化问题.因此,设计合适的适应度函数非常关键.然而,这些研究大多是在经典粗糙集下进行的,在决策粗糙集中,由于决策域与属性增减之间的非单调性,在该模型下的属性约简与经典粗糙集模型下的属性约简相比具有较大的差异性[21].因此,我们有必要研究决策粗糙集模型下的最小属性约简问题.

本文第 4 节将讨论基于决策域分布保持的最小约简问题.首先,将最小约简问题转化为约束优化问题.考虑到遗传算法[24]在求解最小约简问题时的优良表现,本文提出基于遗传算法的正域分布保持约简算法以及非负域分布保持约简算法.适应度函数的最优值是否能够描述最优解,是算法成功与否的关键.理论分析说明,本文提出的适应度函数能够保证将最小约简问题转换为适应度函数最大化问题.此外,为了确保遗传算法能够搜索到决策域分布保持约简,在遗传算法中加入修正算子.修正算子通过第 3 节定义的两种决策域分布条件信息量,将决策域分布保持约简的超集转换为决策域分布保持约简.决策域分布条件信息量的单调性也保证了修正算子的正确性.第 5 节给出的在 UCI 数据集[25]上的实验结果验证了遗传算法找到最小约简的有效性.

1 决策粗糙集模型

本节介绍了决策粗糙集的相关概念[6,26].

一个决策表表示为一个四元组:$S=(U,At=C\cup D,\{V_a|a\in At\},\{I_a|a\in At\})$,其中,$U=\{x_1,x_2,\ldots,x_n\}$是非空有限对象集,$C$是条件属性集,$D$是决策属性集,$C\cap D=\varnothing$,$V_a$是属性$a\in At$的值的非空集合,$I_a:U\to V_a$是一个信息函数,它指定$U$中每个对象的属性值.一般来说,$V$和$I$可省略,决策表简记为$S=(U,At=C\cup D)$.

给定一个决策表$S=(U,At=C\cup D)$,$B\subseteq At$,不可分辨关系定义为

$$IND(B)=\{(x,y)\in U\times U:I_a(x)=I_a(y),\forall a\in B\}.$$

$IND(B)$是U上的一个等价关系,它形成U的一个划分,记为$U/IND(B)$.给定一个对象$x\in U$,$[x]_B$表示包含x的B等价类,即,$[x]_B=\{y\in U:(x,y)\in IND(B)\}$.

给定一个对象x,假设状态集$\Omega=\{D_1,D_2,\ldots,D_m\}$由$m$个决策类构成.$m$类分类问题可以转换成$m$个两类分类问题.具体而言,对第$j$个决策类$D_j$,$j=1,2,\ldots,m$,其状态集可表示为$\Omega_j=\{D_j,\neg D_j\}$,分别表示对象是否属于决策类$D_j$,如果对象$x$通过等价类$[x]_B$刻画,则对象$x$属于$D_j$和不属于$D_j$的条件概率分别为$P(D_j|[x]_B)=\dfrac{|[x]_B\cap D_j|}{|[x]_B|}$和$P(\neg D_j|[x]_B)=1-P(D_j|[x]_B)$.给定行为集$A=\{a_{P_j},a_{B_j},a_{N_j}\}$,其中,$a_{P_j},a_{B_j},a_{N_j}$分别表示将对象分类到正域$POS(D_j)$、边界域$BND(D_j)$和负域$NEG(D_j)$的 3 种行为.损失函数可由如下矩阵表示:

$$\begin{array}{c|ccc|ccc} & D_j & & & \neg D_j & & \\ & a_{P_j} & a_{B_j} & a_{N_j} & a_{P_j} & a_{B_j} & a_{N_j} \\ D_1 & \lambda_{P_1D_1} & \lambda_{B_1D_1} & \lambda_{N_1D_1} & \lambda_{P_1\neg D_1} & \lambda_{B_1\neg D_1} & \lambda_{N_1\neg D_1} \\ D_2 & \lambda_{P_2D_2} & \lambda_{B_2D_2} & \lambda_{N_2D_2} & \lambda_{P_2\neg D_2} & \lambda_{B_2\neg D_2} & \lambda_{N_2\neg D_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ D_j & \lambda_{P_jD_j} & \lambda_{B_jD_j} & \lambda_{N_jD_j} & \lambda_{P_j\neg D_j} & \lambda_{B_j\neg D_j} & \lambda_{N_j\neg D_j} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ D_m & \lambda_{P_mD_m} & \lambda_{B_mD_m} & \lambda_{N_mD_m} & \lambda_{P_m\neg D_m} & \lambda_{B_m\neg D_m} & \lambda_{N_m\neg D_m} \end{array}.$$

在该矩阵中,$\lambda_{P_jD_j}$,$\lambda_{B_jD_j}$ 和 $\lambda_{N_jD_j}$ 表示当一个对象属于 D_j 时采取行为 a_{P_j},a_{B_j} 和 a_{N_j} 引起的损失;$\lambda_{P_j\neg D_j}$,$\lambda_{B_j\neg D_j}$ 和 $\lambda_{N_j\neg D_j}$ 表示当一个对象不属于 D_j 时采取行为 a_{P_j},a_{B_j} 和 a_{N_j} 引起的损失.

给定一个对象关于属性 B 的等价类$[x]_B$,则采取 3 种行为的期望损失分别表示为

$$R(a_{P_j}\,|\,[x]_B)=\lambda_{P_jD_j}P(D_j\,|\,[x]_B)+\lambda_{P_j\neg D_j}P(\neg D_j\,|\,[x]_B),$$

$$R(a_{B_j}\,|\,[x]_B)=\lambda_{B_jD_j}P(D_j\,|\,[x]_B)+\lambda_{B_j\neg D_j}P(\neg D_j\,|\,[x]_B),$$

$$R(a_{N_j}\,|\,[x]_B)=\lambda_{N_jD_j}P(D_j\,|\,[x]_B)+\lambda_{N_j\neg D_j}P(\neg D_j\,|\,[x]_B).$$

根据最小风险贝叶斯决策准则,可以得到如下形式的决策规则:

(P) If $R(a_{P_j}\,|\,[x]_B)\leqslant R(a_{B_j}\,|\,[x]_B)$ and $R(a_{P_j}\,|\,[x]_B)\leqslant R(a_{N_j}\,|\,[x]_B)$, decide $x\in POS(D_j)$;

(B) If $R(a_{B_j}\,|\,[x]_B)\leqslant R(a_{P_j}\,|\,[x]_B)$ and $R(a_{B_j}\,|\,[x]_B)\leqslant R(a_{N_j}\,|\,[x]_B)$, decide $x\in BND(D_j)$;

(N) If $R(a_{N_j}\,|\,[x]_B)\leqslant R(a_{P_j}\,|\,[x]_B)$ and $R(a_{N_j}\,|\,[x]_B)\leqslant R(a_{B_j}\,|\,[x]_B)$, decide $x\in NEG(D_j)$.

因为 $P(D_j|[x]_B)+P(\neg D_j|[x]_B)=1$,考虑一组特殊的损失函数:

(c_1) $\lambda_{P_jD_j}\leqslant\lambda_{B_jD_j}<\lambda_{N_jD_j}$;$\lambda_{N_j\neg D_j}\leqslant\lambda_{B_j\neg D_j}<\lambda_{P_j\neg D_j}$.

该损失函数表示:将一个属于 D_j 的对象分类到 D_j 正域 $POS(D_j)$的损失小于等于将它分类到 D_j 边界域 $BND(D_j)$的损失,并且这两种损失都小于将它分类到 D_j 负域 $NEG(D_j)$的损失;反之,将一个不属于 D_j 的对象分类到 D_j 负域 $NEG(D_j)$的损失小于等于将它分类到 D_j 边界域 $BND(D_j)$的损失,并且这两种损失都小于将它分类到 D_j 正域 $POS(D_j)$的损失.在条件(c_1)下,决策规则(P)~规则(N)可以简化为:

(P1) If $P(D_j|[x]_B)\geqslant\alpha_j$ and $P(D_j|[x]_B)\geqslant\gamma_j$, decide $x\in POS(D_j)$;

(B1) If $P(D_j|[x]_B)\leqslant\alpha_j$ and $P(D_j|[x]_B)\geqslant\beta_j$, decide $x\in BND(D_j)$;

(N1) If $P(D_j|[x]_B)\leqslant\beta_j$ and $P(D_j|[x]_B)\leqslant\gamma_j$, decide $x\in NEG(D_j)$.

其中,参数α_j,β_j 和γ_j 可表示为

$$\begin{cases}\alpha_j=\dfrac{(\lambda_{P_j\neg D_j}-\lambda_{B_j\neg D_j})}{(\lambda_{P_j\neg D_j}-\lambda_{B_j\neg D_j})+(\lambda_{B_jD_j}-\lambda_{P_jD_j})}\\ \beta_j=\dfrac{(\lambda_{B_j\neg D_j}-\lambda_{N_j\neg D_j})}{(\lambda_{B_j\neg D_j}-\lambda_{N_j\neg D_j})+(\lambda_{N_jD_j}-\lambda_{B_jD_j})}\\ \gamma_j=\dfrac{(\lambda_{P_j\neg D_j}-\lambda_{N_j\neg D_j})}{(\lambda_{P_j\neg D_j}-\lambda_{N_j\neg D_j})+(\lambda_{N_jD_j}-\lambda_{P_jD_j})}\end{cases}\tag{1}$$

此外,对于边界区域,规则(B1)的条件表明$\alpha_j>\beta_j$.因此,我们得到条件(c_2):

(c_2) $\dfrac{(\lambda_{N_jD_j}-\lambda_{B_jD_j})}{(\lambda_{B_j\neg D_j}-\lambda_{N_j\neg D_j})}>\dfrac{(\lambda_{B_jD_j}-\lambda_{P_jD_j})}{(\lambda_{P_j\neg D_j}-\lambda_{B_j\neg D_j})}$.

条件(c_1)和条件(c_2)说明 $1\geqslant\alpha_j>\gamma_j>\beta_j\geqslant 0$.在这种情况下,经过权衡,决策规则(P1)~规则(N1)可进一步简化为:

(P2) If $P(D_j|[x]_B)\geqslant\alpha_j$,decide $x\in POS(D_j)$;

(B2) If $\beta_j<P(D_j|[x]_B)<\alpha_j$,decide $x\in BND(D_j)$;

(N2) If $P(D_j|[x]_B)\leqslant\beta_j$,decide $x\in NEG(D_j)$.

参数γ_j不再需要,规则(P2)、规则(B2)和规则(N2)保证每个对象只被分类到一个区域.

根据规则(P2)、规则(B2)和规则(N2),(α_j,β_j)-正域、(α_j,β_j)-边界域和(α_j,β_j)-负域分别定义为

$$\begin{cases} POS_B^{(\alpha_j,\beta_j)}(D_j)=\{x\in U\mid P(D_j\mid[x]_B)\geqslant\alpha_j\} \\ BND_B^{(\alpha_j,\beta_j)}(D_j)=\{x\in U\mid \beta_j<P(D_j\mid[x]_B)<\alpha_j\} \\ NEG_B^{(\alpha_j,\beta_j)}(D_j)=\{x\in U\mid P(D_j\mid[x]_B)\leqslant\beta_j\} \end{cases} \tag{2}$$

正域和边界域的并称为非负域,记为 $\neg NEG_B^{(\alpha_j,\beta_j)}(D_j)=\{x\in U\mid P(D_j\mid[x]_B)>\beta_j\}$.

在接下来的论述中,我们记$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_m)$,$\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_m)$,决策属性 D 导出的划分为$\boldsymbol{\pi}_D=\{D_1,D_2,\ldots,D_m\}$.为了简化讨论与描述,假设每一个决策类具有同样的损失函数,即,对于任意 $D_j\in\boldsymbol{\pi}_D$,有:

$$\lambda_{PP}=\lambda_{P_jD_j},\lambda_{BP}=\lambda_{B_jD_j},\lambda_{NP}=\lambda_{N_jD_j},\lambda_{PN}=\lambda_{P_j\neg D_j},\lambda_{BN}=\lambda_{B_j\neg D_j},\lambda_{NN}=\lambda_{N_j\neg D_j}.$$

2 基于决策域分布保持的属性约简

在这一节,首先分析了决策粗糙集中属性约简面临决策域变化的问题.为了解决这个问题,本文随后提出了决策域分布保持约简的概念.

2.1 存在的问题

在粗糙集理论中,一个约简可以理解为保持给定决策表某种性质不变的最小属性集合,其定义如下:

定义 1[27]. 给定一个决策表 $S=(U,At=C\cup D)$,$B\subseteq C$,用ρ来表示决策表的某种性质,则称 B 是保持性质ρ的 C 的一个约简,当且仅当满足以下条件:

(I) B 和 C 具有同样的性质ρ;

(II) 对于$\forall b\in B$,$B-\{b\}$不满足性质ρ.

在定义 1 中,条件(I)被称为充分条件,条件(II)被称为必要条件.

在经典粗糙集模型中,根据不同的标准,性质ρ通常可以替换为决策表的正域[1]、分类质量[1]和条件信息熵[28]等.在概率粗糙集模型中,性质ρ通常被替换为决策划分$\boldsymbol{\pi}_D$的正域或者非负域的定量与定性标准,其中,决策划分的正域与非负域定义如下:

定义 2. 给定一个决策表 $S=(U,At=C\cup D)$,$B\subseteq C$,则决策划分$\boldsymbol{\pi}_D$的正域和非负域分别定义为

$$\begin{cases} POS_B^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)=\{x\in U\mid \exists D_j\in\boldsymbol{\pi}_D: P(D_j\mid[x]_B)\geqslant\alpha_j\} \\ \neg NEG_B^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)=\{x\in U\mid \exists D_j\in\boldsymbol{\pi}_D: P(D_j\mid[x]_B)>\beta_j\} \end{cases} \tag{3}$$

为了简化,在接下来的讨论中我们只讨论正域,对于非负域有类似的结果.正如文献[21]所指出的:由于正域关于属性集合包含之间的单调性并不成立,因此,正域的定量等价($|POS_B^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)|=|POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)|$)并不意味着它们的定性等价 ($POS_B^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)=POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)$), 即 $|POS_B^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)|=|POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)|$ 并不意味着 $POS_B^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)=POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)$.这就导致当用$|POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)|$或者 $POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)$ 替换定义 1 中的性质ρ时,得到的约简结果有可能不相同.而在经典粗糙集中,因为正域与属性之间的单调性成立,这两种约简定义是等价的.

因此,在决策粗糙集中,当属性减少时,正域或者非负域有可能变大、变小或者不变[21].针对这个特点,Yao 和 Zhao[2]提出了一种属性约简的泛化定义,即一个约简可以理解为满足给定决策表某种性质的最小属性集合.其形式化的定义如下:

定义 3[2]. 给定一个决策表 $S=(U,At=C\cup D)$,$B\subseteq C$,用一个度量集合 $E=\{e_1,e_2,\ldots\}$来表示决策表的某种性质,则称 B 是和决策属性 D 相关的 C 的一个约简,当且仅当满足以下条件:

(I) 对于所有 $e\in E$,$e(\boldsymbol{\pi}_D|\boldsymbol{\pi}_B)\succeq e(\boldsymbol{\pi}_D|\boldsymbol{\pi}_C)$;

(II) 对于$\forall A\subset B$,有:对于所有 $e\in E$,$e(\boldsymbol{\pi}_D|\boldsymbol{\pi}_B)\succeq e(\boldsymbol{\pi}_D|\boldsymbol{\pi}_C)$不成立.

在这个定义中,对于给定的某种性质,可用一个度量来表示.该度量定义为从条件属性集的幂集 2^C 到一个偏序集合 L 的映射,即 $e:2^C\to(L,\succeq)$,其中,$\succeq$是满足自反性、非对称性和传递性的偏序关系.

根据定义3,可以将已有的一些属性约简表示出来.例如,针对决策域的非单调,Zhao等人[21]将决策粗糙集中的属性约简理解为一种全局优化问题,即一个约简被定义为 C 的所有子集中使正域最大的最小属性集合,具体 $\arg\max_{B\subseteq C}\{POS_B^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)\}$. 对于其他决策域也可以用同样的方法定义.但是,Jia 等人[13]指出:由于决策域的单调性不满足,当属性增加或删除之后,决策域变大或者变小哪一种情况更好,这在可解释性上存在一定的困难.因此, Jia 等人[13]提出了一种决策风险最小化属性约简定义,将属性约简理解为保持决策风险最小.其定义如下:

定义 4[13]. 给定一个决策表 $S=(U,At=C\cup D),B\subseteq C$,则称 B 是 C 的一个决策风险最小化属性约简,当且仅当满足如下的条件:

(I) $B=\arg\min_{B\subseteq C}\{\boldsymbol{COST}_B\}$;

(II) 对于$\forall A\subset B$,有 $\boldsymbol{COST}_A>\boldsymbol{COST}_B$.

其中,

$$\boldsymbol{COST}_B=\sum_{x_i\in POS_B^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)}(1-P_i)\cdot\lambda_{PN}+\sum_{x_j\in BND_B^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)}(P_j\cdot\lambda_{BP}+(1-P_j)\cdot\lambda_{BN})+\sum_{x_k\in NEG_B^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)}P_k\cdot\lambda_{NP}\tag{4}$$

其中,$P_i=P(D_{\max}([x_i]_B)\,|\,[x_i]_B),D_{\max}([x_i]_B)=\arg\max_{D_j\in\pi_D}\left\{\frac{|[x_i]_B\cap D_j|}{|[x_i]_B|}\right\}$.

现在我们通过一个例子来说明上述几种定义中存在的问题.

例 1:给定一个决策表 $S=(U,At=C\cup D)$,见表 1,其中,$U=\{x_1,x_2,\ldots,x_{12}\}$,$C=\{c_1,c_2,c_3,c_4,c_5,c_6\}$,$D=\{d\}$.假设所有损失函数为:$\lambda_{PP}=\lambda_{NN}=0,\lambda_{PN}=6,\lambda_{NP}=3,\lambda_{BP}=1,\lambda_{BN}=3$,那么根据公式(1),有:

$$\boldsymbol{\alpha}=(0.75,0.75,0.75),\ \boldsymbol{\beta}=(0.60,0.60,0.60).$$

Table 1 A decision table

表 1 一个决策表

U	c_1	c_2	c_3	c_4	c_5	c_6	d
x_1	1	0	0	0	0	0	1
x_2	1	1	0	0	1	1	1
x_3	1	1	0	0	1	1	1
x_4	0	0	1	1	0	1	2
x_5	1	1	0	0	0	1	2
x_6	1	1	0	0	0	1	2
x_7	1	1	0	0	1	1	2
x_8	1	0	0	0	0	1	3
x_9	1	0	1	0	1	1	3
x_{10}	1	0	0	0	0	0	3
x_{11}	1	0	1	0	1	1	2
x_{12}	1	0	0	0	0	0	3

根据公式(2)和定义 2,有:

$$POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(D_1)=\varnothing,POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(D_2)=\{x_4,x_5,x_6\},POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(D_3)=\{x_8\},POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)=\{x_4,x_5,x_6,x_8\}.$$

根据定义 1、定义 3 和定义 4,可以得到$\{c_2,c_3\}$是一个正域的定量保持属性约简,$\{c_5,c_6\}$是一个正域的定性保持属性约简,$\{c_4\}$是一个决策风险最小化属性约简,$\{c_2,c_4,c_5\}$是一个正域最大化属性约简.

根据定义 2,我们可以计算出这些不同约简的正域:

$$POS_{\{c_2,c_3\}}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)=\{x_1,x_8,x_{10},x_{12}\},$$

$$POS_{\{c_5,c_6\}}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)=\{x_4,x_5,x_6,x_8\},$$

$$POS_{\{c_4\}}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)=\{x_4\},$$

$$POS_{\{c_2,c_4,c_5\}}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{\pi}_D)=\{x_1,x_4,x_5,x_6,x_8,x_{10},x_{12}\}.$$

从以上结果可以看出,正域的定量保持属性约简和决策风险最小化属性约简都不同程度地改变了正域,正

域最大化属性约简虽然使正域扩大,但是它的可解释性存在困难[13].正域的定性保持属性约简虽然不改变整个决策表的正域,即 $POS_{\{c_5,c_6\}}^{(\alpha,\beta)}(\pi_D)=POS_C^{(\alpha,\beta)}(\pi_D)=\{x_4,x_5,x_6,x_8\}$, 但它并不能保持每一个决策类的正域不变,例如:

$$POS_{\{c_5,c_6\}}^{(\alpha,\beta)}(D_1)=\varnothing,$$
$$POS_{\{c_5,c_6\}}^{(\alpha,\beta)}(D_2)=\{x_4,x_5,x_6,x_8\},$$
$$POS_{\{c_5,c_6\}}^{(\alpha,\beta)}(D_3)=\varnothing.$$

与原始决策表每个决策类的正域分布相比,决策类 D_1 和 D_2 的正域被改变.

由例 1 可见,由于决策域关于属性集合的非单调性,上述几种属性约简的定义都有可能改变决策域.然而,实际应用中我们并不希望改变每个决策类的决策域,因为我们引入概率阈值的目的是为了增强分类能力,而不是改变决策域.属性约简的目的只是为了消除冗余属性.

此外,因为在决策粗糙集中决策域和决策风险相对于属性增减之间的非单调性,在定义 1、定义 3 和定义 4 中的必要条件必须检查属性集合 B 的所有子集,而不是对 B 中每一个属性 b 依次检查一遍子集 $B-\{b\}$[21].因而,传统的基于增加-删除法和基于直接删除法[27]的启发式约简算法可能会找到约简的超集而不是约简本身.相比之下,采取群体智能优化算法更容易求得最优解,例如遗传算法和粒子群优化算法等[13,29].值得注意的是,为了减少算法设计上的困难,开发一种具有单调性的度量函数来指导概率粗糙集中属性约简算法的设计是必要的.

2.2 决策域分布保持约简的定义

为了保证在属性约简之后每一个决策类的决策域不发生变化,我们给出了(α,β)正域分布保持约简和(α,β)非负域分布保持约简的定义.与上述几种属性约简定义相比,正域分布保持约简和非负域分布保持约简分别不改变每一个决策类的正域和非负域.

定义 5. 给定一个决策表 $S=(U,At=C\cup D),B\subseteq C$,记

$$\begin{cases} POS_B^{(\alpha,\beta)}=\{POS_B^{(\alpha_1,\beta_1)}(D_1),POS_B^{(\alpha_2,\beta_2)}(D_2),...,POS_B^{(\alpha_m,\beta_m)}(D_m)\} \\ \neg NEG_B^{(\alpha,\beta)}=\{\neg NEG_B^{(\alpha_1,\beta_1)}(D_1),\neg NEG_B^{(\alpha_2,\beta_2)}(D_2),...,\neg NEG_B^{(\alpha_m,\beta_m)}(D_m)\} \end{cases} \tag{5}$$

(1) 如果 $POS_B^{(\alpha,\beta)}=POS_C^{(\alpha,\beta)}$, 则称 B 为 S 的(α,β)正域分布保持集;如果 $POS_B^{(\alpha,\beta)}=POS_C^{(\alpha,\beta)}$ 且对于$\forall A\subset B$, $POS_A^{(\alpha,\beta)}\neq POS_C^{(\alpha,\beta)}$, 则称 B 为 S 的(α,β)正域分布保持约简.

(2) 如果 $\neg NEG_B^{(\alpha,\beta)}=\neg NEG_C^{(\alpha,\beta)}$, 则称 B 为 S 的(α,β)非负域分布保持集;如果 $\neg NEG_B^{(\alpha,\beta)}=\neg NEG_C^{(\alpha,\beta)}$ 且对于$\forall A\subset B$, $\neg NEG_A^{(\alpha,\beta)}\neq\neg NEG_C^{(\alpha,\beta)}$, 则称 B 为 S 的(α,β)非负域分布保持约简.

一个(α,β)正域(或非负域)分布保持集是保持决策表每个决策类的正域(或非负域)不变的属性集合.

例 2:续例 1.

根据定义 5 得到,$B=\{c_1,c_2,c_5,c_6\}$是一个正域分布保持约简.此外,有:

$$POS_B^{(\alpha,\beta)}(D_1)=\varnothing,$$
$$POS_B^{(\alpha,\beta)}(D_2)=\{x_4,x_5,x_6\},$$
$$POS_B^{(\alpha,\beta)}(D_3)=\{x_8\}.$$

可以看到,属性集合 B 能够保持决策表每个决策类的正域不变.不难发现,$\{c_2,c_3,c_5\}$是一个非负域分布保持约简.

3 决策域分布保持约简的启发式计算方法

这一节将给出决策域分布保持约简的启发式计算方法。在此基础上给出决策域分布保持约简核属性的定义以及求核算法.

3.1 决策域分布保持约简的启发式计算方法

在设计属性约简算法时,度量函数的单调性是非常重要的.由于在决策粗糙集中,决策域的定义引入了概率阈值,因此决策域随属性的变化之间不具备单调性,这为算法设计带来了一定的困难.因为要找到一个约简就必

须检查属性集合的所有子集,否则可能找到一个约简的超集,而检查所有子集是一个非常耗时的工作.因此,有必要开发满足单调性的启发式度量函数来简化算法设计的困难.在这一节中,我们通过条件信息量的变形提出了(α,β)正域分布条件信息量和(α,β)非负域分布条件信息量,并证明了其单调性。它们分别用于计算(α,β)正域分布保持约简和(α,β)非负域分布保持约简.

首先,让我们回顾一下条件信息量的定义.

定义 6[30]. 给定一个决策表 $S=(U,At=C\cup D),P\subseteq At,Q\subseteq At,U/IND(P)=\{X_1,X_2,\ldots,X_N\},U/IND(Q)=\{Y_1,Y_2,\ldots,Y_M\}$,$Q$相对于 P 的条件信息量定义为

$$I(Q\mid P)=\sum_{i=1}^{N}\frac{|X_i|}{|U|}\sum_{j=1}^{M}\frac{|X_i\cap Y_j|}{|X_i|}\left(1-\frac{|X_i\cap Y_j|}{|X_i|}\right) \tag{6}$$

接下来,我们通过条件信息量的变形,给出两种单调的决策域分布条件信息量.它们分别被称为(α,β)正域分布条件信息量和(α,β)非负域分布条件信息量.在此基础上,我们给出了(α,β)正域分布保持约简和(α,β)非负域分布保持约简的启发式计算方法.

给定一个决策表 $S=(U,At=C\cup D),B\subseteq C$,记论域 U 上的两个覆盖:

$$U/R_{POS_B^{(\alpha,\beta)}}=\{POS_B^{(\alpha_1,\beta_1)}(D_1),POS_B^{(\alpha_2,\beta_2)}(D_2),\ldots,POS_B^{(\alpha_m,\beta_m)}(D_m),POS_B^{(\alpha_{m+1},\beta_{m+1})}(D_{m+1})\},$$

$$U/R_{\neg NEG_B^{(\alpha,\beta)}}=\{\neg NEG_B^{(\alpha_1,\beta_1)}(D_1),\neg NEG_B^{(\alpha_2,\beta_2)}(D_2),\ldots,\neg NEG_B^{(\alpha_m,\beta_m)}(D_m),\neg NEG_B^{(\alpha_{m+1},\beta_{m+1})}(D_{m+1})\},$$

其中,$POS_B^{(\alpha_{m+1},\beta_{m+1})}(D_{m+1})=U-\bigcup_{1\leqslant j\leqslant m}POS_B^{(\alpha_j,\beta_j)}(D_j),\neg NEG_B^{(\alpha_{m+1},\beta_{m+1})}(D_{m+1})=U-\bigcup_{1\leqslant j\leqslant m}\neg NEG_B^{(\alpha_j,\beta_j)}(D_j)$.

定义 7. 给定一个决策表 $S=(U,At=C\cup D),B\subseteq C,U/IND(B)=\{X_1,X_2,\ldots,X_N\}$:

(1) 条件属性 B 相对于 $POS_C^{(\alpha,\beta)}$ 的(α,β)正域分布条件信息量定义为

$$I(R_{POS_C^{(\alpha,\beta)}}\mid B)=\sum_{i=1}^{N}\frac{|X_i|}{|U|}\sum_{j=1}^{m+1}\frac{|X_i\cap POS_C^{(\alpha_j,\beta_j)}(D_j)|}{|X_i|}\left(1-\frac{|X_i\cap POS_C^{(\alpha_j,\beta_j)}(D_j)|}{|X_i|}\right) \tag{7}$$

(2) 条件属性 B 相对于 $\neg NEG_C^{(\alpha,\beta)}$ 的(α,β)非负域分布条件信息量定义为

$$I(R_{\neg NEG_C^{(\alpha,\beta)}}\mid B)=\sum_{i=1}^{N}\frac{|X_i|}{|U|}\sum_{j=1}^{m+1}\frac{|X_i\cap \neg NEG_C^{(\alpha_j,\beta_j)}(D_j)|}{|X_i|}\left(1-\frac{|X_i\cap \neg NEG_C^{(\alpha_j,\beta_j)}(D_j)|}{|X_i|}\right) \tag{8}$$

(α,β)正域和(α,β)非负域分布条件信息量反映了条件属性的划分相对于每个决策类(α,β)正域和(α,β)非负域分布的协调程度.

为了证明(α,β)正域分布条件信息量和(α,β)非负域分布条件信息量的单调性,需要首先证明如下引理:

引理 1. 给定一个决策表 $S=(U,At=C\cup D),A_1\subseteq C,A_2\subseteq C,U/IND(A_1)=\{X_1,X_2,\ldots,X_N\}$,而 $U/IND(A_2)=\{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_{j-1},X_{j+1},\ldots,X_N,X_i\cup X_j\}$是将划分 $U/IND(A_1)$中的某两个等价块 X_i 和 X_j 合并为 $X_i\cup X_j$ 得到的新划分.那么,

(1) $I(R_{POS_C^{(\alpha,\beta)}}\mid A_2)\geqslant I(R_{POS_C^{(\alpha,\beta)}}\mid A_1)$,

$I(R_{POS_C^{(\alpha,\beta)}}\mid A_2)=I(R_{POS_C^{(\alpha,\beta)}}\mid A_1)\Leftrightarrow$

$\forall k\in\{1,2,\ldots,m+1\},\dfrac{|X_i\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_i|}=\dfrac{|X_j\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_j|}$;

(2) $I(R_{\neg NEG_C^{(\alpha,\beta)}}\mid A_2)\geqslant I(R_{\neg NEG_C^{(\alpha,\beta)}}\mid A_1)$,

$I(R_{\neg NEG_C^{(\alpha,\beta)}}\mid A_2)=I(R_{\neg NEG_C^{(\alpha,\beta)}}\mid A_1)\Leftrightarrow$

$\forall k\in\{1,2,\ldots,m+1\},\dfrac{|X_i\cap \neg NEG_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_i|}=\dfrac{|X_j\cap \neg NEG_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_j|}$.

证明:

(1) 记:

$$I(R_{POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}} \mid A_1)=\sum_{i=1}^{N}\frac{|X_i|}{|U|}\sum_{j=1}^{m+1}\frac{|X_i\cap POS_C^{(\alpha_j,\beta_j)}(D_j)|}{|X_i|}\left(1-\frac{|X_i\cap POS_C^{(\alpha_j,\beta_j)}(D_j)|}{|X_i|}\right);$$

$$I(R_{POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}} \mid A_2)=I(R_{POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}} \mid A_1)-\frac{|X_i|}{|U|}\sum_{j=1}^{m+1}\frac{|X_i\cap POS_C^{(\alpha_j,\beta_j)}(D_j)|}{|X_i|}\left(1-\frac{|X_i\cap POS_C^{(\alpha_j,\beta_j)}(D_j)|}{|X_i|}\right)-$$

$$\frac{|X_j|}{|U|}\sum_{k=1}^{m+1}\frac{|X_j\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_j|}\left(1-\frac{|X_j\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_j|}\right)+$$

$$\frac{|X_i\cup X_j|}{|U|}\sum_{k=1}^{m+1}\frac{|(X_i\cup X_j)\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_i\cup X_j|}\left(1-\frac{|(X_i\cup X_j)\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_i\cup X_j|}\right);$$

$$I_\Delta=I(R_{POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}} \mid A_2)-I(R_{POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}} \mid A_1)$$

$$=\frac{|X_i\cup X_j|}{|U|}\sum_{k=1}^{m+1}\frac{|(X_i\cup X_j)\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_i\cup X_j|}\left(1-\frac{|(X_i\cup X_j)\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_i\cup X_j|}\right)-$$

$$\frac{|X_i|}{|U|}\sum_{j=1}^{m+1}\frac{|X_i\cap POS_C^{(\alpha_j,\beta_j)}(D_j)|}{|X_i|}\left(1-\frac{|X_i\cap POS_C^{(\alpha_j,\beta_j)}(D_j)|}{|X_i|}\right)-$$

$$\frac{|X_j|}{|U|}\sum_{k=1}^{m+1}\frac{|X_j\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_j|}\left(1-\frac{|X_j\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_j|}\right)$$

$$=\frac{1}{|U|}\sum_{k=1}^{m+1}\left[|(X_i\cup X_j)\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|\left(1-\frac{|(X_i\cup X_j)\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_i\cup X_j|}\right)-\right.$$

$$\left.|X_i\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|\left(1-\frac{|X_i\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_i|}\right)-|X_j\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|\left(1-\frac{|X_j\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_j|}\right)\right]$$

$$=\frac{1}{|U|}\sum_{k=1}^{m+1}\left[\frac{|X_i\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|^2}{|X_i|}+\frac{|X_j\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|^2}{|X_j|}-\frac{|(X_i\cup X_j)\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|^2}{|X_i\cup X_j|}\right].$$

令$|X_i|=x,|X_j|=y,|X_i\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|=a_kx,|X_j\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|=b_kx$,显然有 $x>0,y>0,0\leqslant a_k\leqslant 1,0\leqslant b_k\leqslant 1$,则

$$I_\Delta=\frac{1}{|U|}\sum_{k=1}^{m+1}\left[\frac{(a_kx)^2}{x}+\frac{(b_ky)^2}{y}-\frac{(a_kx+b_ky)^2}{x+y}\right]=\frac{1}{|U|}\sum_{k=1}^{m+1}\left[\frac{(a_k-b_k)^2xy}{x+y}\right]\geqslant 0.$$

当 $a_k=b_k$ 时,即,对于$\forall k\in\{1,2,\ldots,m+1\}$,都有 $\frac{|X_i\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_i|}=\frac{|X_j\cap POS_C^{(\alpha_k,\beta_k)}(D_k)|}{|X_j|}$ 的情况下,$I_\Delta=0$.

故,引理得证.

(2) 证明与情形(1)的证明类似. □

引理 1 说明:如果将决策表条件属性集的分类进行合并,($\boldsymbol{\alpha},\boldsymbol{\beta}$)正域分布条件信息量和($\boldsymbol{\alpha},\boldsymbol{\beta}$)非负域分布条件信息量将单调不减.因此,我们得到如下的单调性定理:

定理 1. 给定一个决策表 $S=(U,At=C\cup D),A\subseteq C,B\subseteq C$ 且 $A\subseteq B$,那么,

(1) $I(R_{POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}} \mid A)\geqslant I(R_{POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}} \mid B)$;

(2) $I(R_{\neg NEG_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}} \mid A)\geqslant I(R_{\neg NEG_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}} \mid B)$.

证明:

(1) 根据引理 1,如果将决策表条件属性的分类进行合并,将使($\boldsymbol{\alpha},\boldsymbol{\beta}$)正域分布条件信息量非严格单调增加,而划分 $U/IND(A)$是可以通过将划分 $U/IND(B)$中的部分等价类合并得到的,所以有:

$$I(R_{POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}} \mid A)\geqslant I(R_{POS_C^{(\boldsymbol{\alpha},\boldsymbol{\beta})}} \mid B).$$

(2) 证明与情形(1)的证明类似. □

定理 2. 给定一个决策表 $S=(U,At=C\cup D),B\subseteq C$,那么,

(1) $I(R_{POS_B^{(\alpha,\beta)}}\mid B)=0$;

(2) $I(R_{\neg NEG_B^{(\alpha,\beta)}}\mid B)=0$.

证明:根据(α,β)正域分布条件信息量和(α,β)非负域分布条件信息量的定义直接可以得到. □

定理 3. 给定一个决策表 $S=(U,At=C\cup D),B\subseteq C$,那么,

(1) B 是 S 的一个(α,β)正域分布保持集当且仅当 $I(R_{POS_C^{(\alpha,\beta)}}\mid B)=0$;

(2) B 是 S 的一个(α,β)非负域分布保持集当且仅当 $I(R_{\neg NEG_C^{(\alpha,\beta)}}\mid B)=0$.

证明:记$\xi([x]_B)=\{[y]_C:[y]_C\subseteq[x]_B\}$,因为 $B\subseteq C$,所以$\xi([x]_B)$构成$[x]_B$的一个划分.

(1) 设 $U/R_{POS_C^{(\alpha,\beta)}}=\{POS_C^{(\alpha_1,\beta_1)}(D_1),POS_C^{(\alpha_2,\beta_2)}(D_2),...,POS_C^{(\alpha_m,\beta_m)}(D_m),POS_C^{(\alpha_{m+1},\beta_{m+1})}(D_{m+1})\}$.

必要性:因为 B 是 S 的一个(α,β)正域分布保持集,所以有 $POS_B^{(\alpha,\beta)}=POS_C^{(\alpha,\beta)}$, 因此有 $U/R_{POS_B^{(\alpha,\beta)}}=U/R_{POS_C^{(\alpha,\beta)}}$, 根据定理 2 可以得到 $I(R_{POS_C^{(\alpha,\beta)}}\mid B)=I(R_{POS_B^{(\alpha,\beta)}}\mid B)=0$.

充分性:对任意 $1\leqslant j\leqslant m$,若 $x\in POS_C^{(\alpha_j,\beta_j)}(D_j)$, 则可以得到 $[x]_B\subseteq POS_C^{(\alpha_j,\beta_j)}(D_j)$ 或 $[x]_B\cap POS_C^{(\alpha_k,\beta_k)}(D_k)\neq\varnothing$ $(1\leqslant k\leqslant m$ 且 $k\neq j)$或 $[x]_B\cap\left(U-\bigcup_{1\leqslant j\leqslant m}POS_{(\alpha_j,\beta_j)}(D_j)\right)\neq\varnothing$.

当 $[x]_B\cap POS_C^{(\alpha_k,\beta_k)}(D_k)\neq\varnothing$ 或 $[x]_B\cap\left(U-\bigcup_{1\leqslant j\leqslant m}POS_{(\alpha_j,\beta_j)}(D_j)\right)\neq\varnothing$ 时,有:

$$I(R_{POS_C^{(\alpha,\beta)}}\mid B)=\sum_{[x]_B\in U/B}\frac{|[x]_B|}{|U|}\sum_{j=1}^{m+1}\frac{|[x]_B\cap POS_C^{(\alpha_j,\beta_j)}(D_j)|}{|[x]_B|}\left(1-\frac{|[x]_B\cap POS_C^{(\alpha_j,\beta_j)}(D_j)|}{|[x]_B|}\right)>0.$$

这与 $I(R_{POS_C^{(\alpha,\beta)}}\mid B)=0$ 矛盾,因此有 $[x]_B\subseteq POS_C^{(\alpha_j,\beta_j)}(D_j)$; 又因为 $[x]_B=\bigcup\{[y]_C:[y]_C\in\xi([x]_B)\}$, 所以对于所有$[y]_C\in\xi([x]_B)$, $[y]_C\subseteq[x]_B\subseteq POS_C^{(\alpha_j,\beta_j)}(D_j)$ 成立.这就是说,对于所有$[y]_C\in\xi([x]_B)$,有 $P(D_j|[y]_C)\geqslant\alpha_j$.从而

$$\begin{aligned}P(D_j\mid[x]_B)&=\left(\sum\{|[y]_C\cap D_j|:[y]_C\in\xi([x]_B)\}\right)\Big/|[x]_B|\\&=\sum\left\{P(D_j\mid[y]_C)\cdot\frac{|[y]_C|}{|[x]_B|}:[y]_C\in\xi([x]_B)\right\}\\&\geqslant\alpha_j\sum\left\{\frac{|[y]_C|}{|[x]_B|}:[y]_C\in\xi([x]_B)\right\}=\alpha_j.\end{aligned}$$

因此, $x\in POS_B^{(\alpha_j,\beta_j)}(D_j)$.

另一方面,如果 $x\in POS_B^{(\alpha_j,\beta_j)}(D_j)$, 那么有 $[x]_B\subseteq POS_B^{(\alpha_j,\beta_j)}(D_j)$;

又因为 $I(R_{POS_C^{(\alpha,\beta)}}\mid B)=0$, 可以得到 $[x]_B\subseteq POS_C^{(\alpha_j,\beta_j)}(D_j)$ 或 $[x]_B\cap POS_C^{(\alpha_j,\beta_j)}(D_j)=\varnothing$.

当 $[x]_B\cap POS_C^{(\alpha_j,\beta_j)}(D_j)=\varnothing$ 时,因为 $[x]_B=\bigcup\{[y]_C:[y]_C\in\xi([x]_B)\}$, 所以对于所有$[y]_C\in\xi([x]_B)$,有:

$$[x]_C\cap POS_C^{(\alpha_j,\beta_j)}(D_j)=\varnothing.$$

这就是说,对于所有$[y]_C\in\xi([x]_B)$,有 $P(D_j|[y]_C)<\alpha_j$.从而

$$\begin{aligned}P(D_j\mid[x]_B)&=\left(\sum\{|[y]_C\cap D_j|:[y]_C\in\xi([x]_B)\}\right)\Big/|[x]_B|\\&=\sum\left\{P(D_j\mid[y]_C)\cdot\frac{|[y]_C|}{|[x]_B|}:[y]_C\in\xi([x]_B)\right\}\\&<\alpha_j\sum\left\{\frac{|[y]_C|}{|[x]_B|}:[y]_C\in\xi([x]_B)\right\}=\alpha_j.\end{aligned}$$

结果 $[x]_B \cap POS_B^{(\alpha_j,\beta_j)}(D_j)=\varnothing$, 这与 $[x]_B \subseteq POS_B^{(\alpha_j,\beta_j)}(D_j)$ 矛盾,因此 $[x]_B \subseteq POS_C^{(\alpha_j,\beta_j)}(D_j)$, 从而

$$x \in POS_C^{(\alpha_j,\beta_j)}(D_j).$$

这就证明了对于任意 $1\leqslant j\leqslant m$,有 $POS_B^{(\alpha_j,\beta_j)}(D_j)=POS_C^{(\alpha_j,\beta_j)}(D_j)$, 即 B 是 S 的一个(α,β)正域分布保持集.

(2) 证明与情形(1)的证明类似. □

定理 4. 给定一个决策表 $S=(U,At=C\cup D),B\subseteq C$,那么,

(1) B 中一个属性 b 相对于 $R_{POS_B^{(\alpha,\beta)}}$ 是不必要的当且仅当 $I(R_{POS_B^{(\alpha,\beta)}} \mid B-\{b\})=0$;

(2) B 中一个属性 b 相对于 $R_{\neg NEG_B^{(\alpha,\beta)}}$ 是不必要的当且仅当 $I(R_{\neg NEG_B^{(\alpha,\beta)}} \mid B-\{b\})=0$.

根据定理 4,立即得到如下推论:

推论 1. 给定一个决策表 $S=(U,At=C\cup D),B\subseteq C$,那么,

(1) B 中一个属性 b 相对于 $R_{POS_B^{(\alpha,\beta)}}$ 是必要的当且仅当 $I(R_{POS_B^{(\alpha,\beta)}} \mid B-\{b\})>0$;

(2) B 中一个属性 b 相对于 $R_{\neg NEG_B^{(\alpha,\beta)}}$ 是必要的当且仅当 $I(R_{\neg NEG_B^{(\alpha,\beta)}} \mid B-\{b\})>0$.

根据定理 3 和推论 1,可以得到如下的定理:

定理 5. 给定一个决策表 $S=(U,At=C\cup D),B\subseteq C$,那么,

(1) B 是 S 的一个(α,β)正域分布保持约简当且仅当:

(I) $I(R_{POS_C^{(\alpha,\beta)}} \mid B)=0$;

(II) 对于任意 $b\in B$,有 $I(R_{POS_C^{(\alpha,\beta)}} \mid B-\{b\})>0$.

(2) B 是 S 的一个(α,β)非负域分布保持约简当且仅当:

(I) $I(R_{\neg NEG_C^{(\alpha,\beta)}} \mid B)=0$;

(II) 对于任意 $b\in B$ 有 $I(R_{\neg NEG_C^{(\alpha,\beta)}} \mid B-\{b\})>0$.

定理 5 给出了(α,β)正域分布保持约简和(α,β)非负域分布保持约简的启发式计算方法.

3.2 决策域分布保持约简的核属性

在一个决策表中,核属性是决策表所有约简的交集.它通常可以作为属性约简算法的起点,然后利用一定的启发式信息求解属性约简,这样可以显著缩小约简算法在属性空间中的搜索范围,有效提高约简算法的运行效率.接下来,我们给出决策表中两种决策域分布保持约简核属性的形式化定义及其求核算法.

定义 8. 给定一个决策表 $S=(U,At=C\cup D)$,那么:

(1) (α,β)正域分布保持约简的核属性定义为

$$CORE_{POS^{(\alpha,\beta)}}(C)=\bigcap RED_{POS^{(\alpha,\beta)}}(C) \tag{9}$$

(2) (α,β)非负域分布保持约简的核属性定义为

$$CORE_{\neg NEG^{(\alpha,\beta)}}(C)=\bigcap RED_{\neg NEG^{(\alpha,\beta)}}(C) \tag{10}$$

其中, $RED_{POS^{(\alpha,\beta)}}(C)$ 和 $RED_{\neg NEG^{(\alpha,\beta)}}(C)$ 分别是所有(α,β)正域分布保持约简和所有(α,β)非负域分布保持约简的集合.

根据定理 3 和定义 8,可以得到如下的定理:

定理 6. 给定一个决策表 $S=(U,At=C\cup D),c\in C$,那么:

(1) $c\in CORE_{POS^{(\alpha,\beta)}}(C)$ 当且仅当 $I(R_{POS_C^{(\alpha,\beta)}} \mid C-\{c\})>0$;

(2) $c\in CORE_{\neg NEG^{(\alpha,\beta)}}(C)$ 当且仅当 $I(R_{\neg NEG_C^{(\alpha,\beta)}} \mid C-\{c\})>0$.

证明:

(1) 必要性:设 $I(R_{POS_C^{(\alpha,\beta)}} \mid C-\{c\})=0$, 根据定理 4,$c$ 在 C 中是不必要的.这与 $c\in CORE_{POS^{(\alpha,\beta)}}(C)$ 矛盾,因此有

$I(R_{POS_C^{(\alpha,\beta)}} \mid C-\{c\})>0$.

充分性:如果 $I(R_{POS_C^{(\alpha,\beta)}} \mid C-\{c\})>0$, 根据推论 1,$c$ 在 C 中是必要的,那么它一定出现在 S 的所有(α,β)正域分布保持约简中,因此, $c\in\bigcap RED_{POS^{(\alpha,\beta)}}(C)$, 即 $c\in CORE_{POS^{(\alpha,\beta)}}(C)$.

(2) 证明与情形(1)的证明类似. □

根据定理 6 和定义 8,可以得到如下的定义:

定义 9. 给定一个决策表 $S=(U,At=C\cup D)$,那么,

(1) (α,β)正域分布保持约简的核属性定义为

$$CORE_{POS^{(\alpha,\beta)}}(C)=\{c\in C \mid I(R_{POS_C^{(\alpha,\beta)}} \mid C-\{c\})>0\} \tag{11}$$

(2) (α,β)非负域分布保持约简的核属性定义为

$$CORE_{\neg NEG^{(\alpha,\beta)}}(C)=\{c\in C \mid I(R_{\neg NEG_C^{(\alpha,\beta)}} \mid C-\{c\})>0\} \tag{12}$$

定义 9 给出了两种决策域分布保持约简核属性的计算方法.根据定义 9 我们设计求核算法如下:

算法 1. 求核算法.

Input:决策表 $S=(U,At=C\cup D)$,阈值(α,β).

Output:决策域分布保持约简的核 $CORE_{T^{(\alpha,\beta)}}(C)$.

Note:T=POS 或¬NEG //T 表示决策域分布保持约简的类型

Step 1. 计算 $U/R_{T_B^{(\alpha,\beta)}}$;

Step 2. 令 $\varnothing\rightarrow CORE_{T^{(\alpha,\beta)}}(C)$;

Step 3. for 对每个 $c\in C$ do //计算核属性

计算 $I(R_{T_C^{(\alpha,\beta)}} \mid C-\{c\})$;

if $I(R_{T_C^{(\alpha,\beta)}} \mid C-\{c\})>0$ then

$CORE_{T^{(\alpha,\beta)}}(C)=CORE_{T^{(\alpha,\beta)}}(C)\cup\{c\}$;

end if

end for

Step 4. return $CORE_{T^{(\alpha,\beta)}}(C)$;

在算法 1 中,计算核 $CORE_{T^{(\alpha,\beta)}}(C)$ 需计算$|C|$次 $I(R_{T_C^{(\alpha,\beta)}} \mid C-\{c\})$, 而计算 $I(R_{T_C^{(\alpha,\beta)}} \mid C-\{c\})$ 的时间复杂度为 $O(|C||U|^2)$,所以在最坏情况下,算法 1 的时间复杂度为 $O(|C|^2|U|^2)$.

4 基于决策域分布保持的最小约简问题及启发式算法

这一节讨论了基于决策域分布保持的最小约简问题,将最小约简问题转化为约束优化问题.考虑到遗传算法在求解最小约简时所表现出的高效性,提出了基于遗传算法的决策域分布保持启发式约简算法.

4.1 最小约简问题描述

一般来说,一个决策表存在多个约简.而在实际应用中,为了得到简洁的决策规则,获取最小约简非常有意义.为了设计智能优化算法,在这一节,我们将基于决策域分布保持的最小约简问题转化为如下约束优化问题:

定义 10. 给定一个决策表 $S=(U,At=C\cup D)$,那么,

(1) 基于(α,β)正域分布保持的最小约简问题定义为:最大化 $\frac{|C|-|B|}{|C|}$, 使得:

$$\begin{cases} B\subseteq C \\ POS_B^{(\alpha,\beta)}=POS_C^{(\alpha,\beta)} \\ \forall A\subset B, POS_A^{(\alpha,\beta)}\neq POS_C^{(\alpha,\beta)} \end{cases} \tag{13}$$

(2) 基于(α,β)非负域分布保持的最小约简问题定义为:最大化$\frac{|C|-|B|}{|C|}$使得:

$$\begin{cases} B\subseteq C \\ \neg NEG_B^{(\alpha,\beta)}=\neg NEG_C^{(\alpha,\beta)} \\ \forall A\subset B,\neg NEG_A^{(\alpha,\beta)}\neq \neg NEG_C^{(\alpha,\beta)} \end{cases} \tag{14}$$

给定一个属性集合 B,如果它是定义 10 中的一个可行解,那么它是一个决策域分布保持约简;如果它是定义 10 中的一个最优解,那么它是一个最小约简.

为了简化算法设计的难度,根据定理 5,该优化问题可以等价地定义如下:

定义 11. 给定一个决策表 $S=(U,At=C\cup D)$,那么,

(1) 基于(α,β)正域分布保持的最小约简问题定义为:最大化$\frac{|C|-|B|}{|C|}$使得:

$$\begin{cases} B\subseteq C \\ I(R_{POS_C^{(\alpha,\beta)}}\mid B)=0 \\ \forall b\in B,I(R_{POS_C^{(\alpha,\beta)}}\mid B-\{b\})>0 \end{cases} \tag{15}$$

(2) 基于(α,β)非负域分布保持的最小约简问题定义为:最大化$\frac{|C|-|B|}{|C|}$使得:

$$\begin{cases} B\subseteq C \\ I(R_{\neg NEG_C^{(\alpha,\beta)}}\mid B)=0 \\ \forall b\in B,I(R_{\neg NEG_C^{(\alpha,\beta)}}\mid B-\{b\})>0 \end{cases} \tag{16}$$

4.2 基于遗传算法的决策域分布保持启发式约简算法

这一节给出了利用遗传算法求解最小约简的设计方案,其中包括染色体的表示、适应度函数的设计以及各种遗传算子的设计,然后给出了基于遗传算法的决策域分布保持启发式约简算法.

(1) 染色体的表示

采用一个定长的二进制向量来表示每个染色体,染色体的长度等于条件属性全集所含属性的个数,染色体的每个基因位和相应的条件属性对应,二进制位等于 1,表示备选解包含对应位的条件属性;二进制位等于 0,表示备选解不包含对应位的条件属性.例如,假设在决策表 S 中有 8 个条件属性$\{c_1,c_2,\ldots,c_8\}$,那么染色体 10101100 对应的可能解为$\{c_1,c_3,c_5,c_6\}$.

(2) 适应度函数

适应度函数是评价一个染色体表示的解有多好的确定性指标,它控制了种群的进化方向.在本文的算法中,适应度函数定义为

$$f(ch_B)=\begin{cases}\frac{|C|-|B|}{|C|}+Q, & \text{if } I(R_{T_C^{(\alpha,\beta)}}\mid B)=0\\ Q, & \text{otherwise}\end{cases} \tag{17}$$

其中,ch_B 表示一个染色体;条件 $I(R_{T_C^{(\alpha,\beta)}}\mid B)=0$ 表示一个染色体 ch_B 是一个决策域分布保持约简的超集(决策域分布保持集),这里,T 表示决策域分布保持约简的类型,它可以取 POS 或$\neg NEG$,分别表示正域分布保持约简和非负域分布保持约简.因为我们在设计选择算子时采用轮盘赌的方式,为了增加种群多样性避免算法过早收敛到次优解,在适应度函数中需要一个大于 0 的常量 Q,常量 Q 保证了每一个染色体都有被选择到下一代进化的可能性,在本文中,我们设 Q=0.5.显然,当 ch_B 是一个决策域分布保持约简的超集时,ch_B 中包含的属性越少,适应度函数 $f(ch_B)$的值越大;当 ch_B 是一个决策域分布保持约简的真子集时,适应度函数最小.这确保了适应度函数最大的染色体能够与最小约简相对应,由此说明了适应度函数的正确性.

(3) 选择算子

选择算子根据适应度函数的大小采取比例选择,通过轮盘赌的方式生成后代.每个个体被选择的概率为

$$P(ch_B^i)=\frac{f(ch_B^i)}{\sum_{i=1}^{N} f(ch_B^i)} \tag{18}$$

其中, $f(ch_B^i)$ 表示第 i 个染色体的适应度值,N 是种群大小.同时采用精英保存策略,即,用当前种群中的最优个体替换新种群中的最差个体,且最优个体不进行交叉和变异,精英保存策略能够确保算法最终收敛到最优解.

(4) 交叉算子

交叉算子采用有性繁殖,随机选择两个父个体进行均匀交叉,对染色体的每个基因位,首先生成一个 0~1 之间的随机数,如果随机数小于交叉概率 P_c,则将相应基因位上的两个染色体的基因进行互换.如果染色体的基因位对应决策域分布保持约简的核属性,则不进行交叉操作.

(5) 变异算子

变异算子采用均匀变异,对于每个基因位,随机生成一个 0~1 之间的随机数,如果随机数小于变异概率 P_m,那么将该位上的基因值取反,即,如果该基因位上的基因值是 0,那么将其变为 1;如果是 1,那么将其变为 0.对于核属性所对应的基因位不进行变异操作.

(6) 修正算子

在本文的遗传算法中加入了一个修正算子,由于种群中适应度值大于常量 Q 的染色体是决策域分布保持约简的一个超集,因此,修正算子的作用是将这些表示决策域分布保持约简超集的染色体修正为表示决策域分布保持约简的染色体,以保证最终得到的约简是决策域分布保持约简而不是决策域分布保持约简的超集.算法 2 给出了对染色体进行修正运算的具体描述.

算法 2. 修正算子.

Input:染色体.

Output:修正后的染色体.

Note:T=POS 或¬NEG //T 表示决策域分布保持约简的类型

Step 1. 将染色体解码为对应的属性集合 B,令 CD=B;

Step 2. for 每一个 $a\in CD$ do

计算 $I(R_{T_C^{(\alpha,\beta)}}|\{a\})$;

end for

Step 3. 根据 $I(R_{T_C^{(\alpha,\beta)}}|\{a\})$ 的大小将 CD 中的属性按降序排序;

Step 4. while $CD\neq\varnothing$ do

CD=CD−$\{a\}$,其中,a 是 CD 中的第 1 个属性;

if $I(R_{T_C^{(\alpha,\beta)}}|B-\{a\})=0$ then

B=B−$\{a\}$;

end if

end while

Step 5. 将属性集合 B 编码为对应的染色体 ch_B;

Step 6. return ch_B

(α,β)正域分布条件信息量和(α,β)非负域分布条件信息量的单调性可以保证对每个决策域分布保持约简的超集执行完算法 2 后,能够获得相应的决策域分布保持约简.

对每个染色体的修正运算的时间复杂度为 $O(|C|^2|U|^2)$,而对整个种群修正运算在最坏情况下的时间复杂度为 $O(N|C|^2|U|^2)$,其中,N 是种群大小.

(7) 算法终止条件

当种群变的稳定时算法终止,即连续 t 代最优个体的适应度值没有变化或迭代次数达到最大值.

(8) 遗传约简算法的实现

根据前面的描述,可以设计基于遗传算法的决策域分布保持启发式约简算法如下:

算法 3. 基于遗传算法的决策域分布保持启发式约简算法.

Input:决策表 $S=(U,At=C\cup D)$,阈值(α,β).

Output:决策域分布保持约简 R.

Note:$T=POS$ 或$\neg NEG$ //T 表示决策域分布保持约简的类型

Step 1. 通过算法 1 求决策域分布保持约简的核属性 $CORE_{T(\alpha,\beta)}(C)$;

Step 2. if $I(R_{T_C^{(\alpha,\beta)}} \mid CORE_{T(\alpha,\beta)}(C))=0$ then

令 $R=CORE_{T(\alpha,\beta)}(C)$;

go to Step 8;

end if

Step 3. 令 $t=1$;

Step 4. 初始化:随机生成 N 个长度为$|C|$的二进制串组成初始种群 $pop(t)$,对每个属性 $c\in C$,如果 c 是核属性,则相应的基因位初始化为 1;如果 c 不是核属性,则随机初始化为 0 或 1;

Step 5. 个体评价:评价 $pop(t)$中每个个体的适应度;

Step 6. while 终止条件不满足 do

Step 6.1. 选择:从 $pop(t)$中,运用选择算子选择出 $M/2$ 对父个体,其中,$M\geqslant N$;

Step 6.2. 交叉:对所选择的 $M/2$ 对父个体,以概率 P_c 执行交叉,形成 M 个中间个体;

Step 6.3. 变异:对 M 个中间个体,分别独立以概率 P_m 执行变异,形成 M 个候选个体;

Step 6.4. 选择:从 M 个候选个体中,运用选择算子选择出 N 个个体,组成新一代种群 $pop(t+1)$;

Step 6.5. 修正:对新种群 $pop(t+1)$,运用修正算子进行修正操作;

Step 6.6. 个体评价:评价 $pop(t+1)$中每个个体的适应度;

Step 6.7. 令 $t=t+1$;

end while

Step 7. 令 R 为最优个体对应的属性集合;

Step 8. return R;

算法 3 在最坏情况下时间复杂度为 $O(N|I\text{-max}||C|^2|U|^2)$,其中,$N$ 是种群大小,$|I\text{-max}|$是最大迭代数.

5 实验结果与分析

在这一节中,我们将通过实验比较几种属性约简定义的分类结果,并且验证遗传算法是否能够求得最小约简.数据集来自 UCI 机器学习数据库[25].在实验中,所有连续属性用等频率离散化方法进行离散.不完备数据用均值或者众数填充.所有程序基于 weka[31]用 Java 语言编写.

5.1 6种定义的比较

这一节通过分类正确率与误分类代价两个标准来评估本文第 2 节讨论的 6 种属性约简定义,并给出了各种定义下约简的平均长度;同时,也比较了两种决策域分布保持约简与原始未约简数据的分类正确率和误分类代价,说明决策域分布保持约简是一种更好的选择.

在实验中,对于每个数据集,随机生成 10 组不同的损失函数,损失函数的值在区间(0,1)范围内,分类正确率和误分类代价的均值与标准差被记录.为了简化,假设每个决策类具有同样的损失函数,损失函数满足接下来的限制条件,即,对于$\forall D_j\in\pi_D$,有$\lambda_{BP}<\lambda_{NP},\lambda_{BN}<\lambda_{PN}$ 和$\lambda_{PP}=\lambda_{NN}=0$.十折交叉验证被用于测试,两种流行的分类算法

J48[32]和 Naive Bayes[33]被用于进行比较实验.

由于决策域和决策风险的非单调性,群体智能优化算法更容易求得最优解,因此,实验中所有 6 种定义全部基于遗传算法实现.正域分布保持约简和非负域分布保持约简通过本文方法实现,分别记为 GA-PRDR 和 GA-NNRDR;决策风险最小化属性约简使用文献[13]中的算法,记为 GA-MINDC;正域最大化属性约简采用文献[29]中的算法,记为 GA-MAXPR.事实上,文献[29]给出了基于决策域保持属性约简的遗传算法框架,因此,基于正域的定性与定量保持属性约简可以通过修改文献[29]中的适应函数获得,分别记为 GA-QLPRP 和 GA-QNPRP.6 种遗传算法的详细参数见表 2,其中,P_c 和 P_m 分别表示交叉概率和变异概率.

Table 2 GA parameter settings

表 2 遗传算法参数设置

Algorithm	N	|I-max|	P_c	P_m	Select-Method	Fitness-Function
GA-PRDR	40	100	0.7	0.3	Roulette	$f(ch_B)=\begin{cases}\dfrac{\|C\|-\|B\|}{\|C\|}+0.5, & \text{if } I(R_{POS_C^{(\alpha,\beta)}}\mid B)=0\\ 0.5, & \text{otherwise}\end{cases}$
GA-NNRDR	40	100	0.7	0.3	Roulette	$f(ch_B)=\begin{cases}\dfrac{\|C\|-\|B\|}{\|C\|}+0.5, & \text{if } I(R_{\neg NEG_C^{(\alpha,\beta)}}\mid B)=0\\ 0.5, & \text{otherwise}\end{cases}$
GA-MINDC	40	100	0.7	0.3	Roulette	$f(ch_B)=\boldsymbol{COST}_B+\left(\dfrac{\|B\|}{\|C\|}\right)^3$
GA-MAXPR	40	100	Gaussian	Gaussian	Stochastic	$f(ch_B)=\begin{cases}\dfrac{\|C\|-\|B\|}{\|C\|}+\|POS_B^{(\alpha,\beta)}(\boldsymbol{\pi}_D)\|, & \text{if } POS_B^{(\alpha,\beta)}(\boldsymbol{\pi}_D)\supseteq POS_C^{(\alpha,\beta)}(\boldsymbol{\pi}_D)\\ 0, & \text{otherwise}\end{cases}$
GA-QLPRP	40	100	Gaussian	Gaussian	Stochastic	$f(ch_B)=\begin{cases}\dfrac{\|C\|-\|B\|}{\|C\|}+\|POS_B^{(\alpha,\beta)}(\boldsymbol{\pi}_D)\|, & \text{if } POS_B^{(\alpha,\beta)}(\boldsymbol{\pi}_D)=POS_C^{(\alpha,\beta)}(\boldsymbol{\pi}_D)\\ 0, & \text{otherwise}\end{cases}$
GA-QNPRP	40	100	Gaussian	Gaussian	Stochastic	$f(ch_B)=\begin{cases}\dfrac{\|C\|-\|B\|}{\|C\|}+\|POS_B^{(\alpha,\beta)}(\boldsymbol{\pi}_D)\|, & \text{if } \|POS_B^{(\alpha,\beta)}(\boldsymbol{\pi}_D)\|=\|POS_C^{(\alpha,\beta)}(\boldsymbol{\pi}_D)\|\\ 0, & \text{otherwise}\end{cases}$

(1) 分类正确率

这个部分比较了不同属性约简定义所得到约简的分类正确率.表 3 和表 4 显示了采用 J48 和 Naive Bayes 分类算法进行测试的实验结果,平均最大分类正确率加粗显示.从结果中可以发现:在两种分类模型下,正域分布保持约简和非负域分布保持约简在多数情况下所选择的属性获得了更好的分类结果;相对而言,决策风险最小化属性约简所选择的属性与两种决策域分布保持约简所选择的属性分类效果相当,并且好于正域最大化和正域的定性与定量保持约简.与原始数据分类结果(记录在第 2 列 Raw data 中)相比,两种决策域分布保持约简在多数情况下,分类正确率有所提高.

Table 3 Classification accuracy comparison with J48

表 3 J48 分类正确率比较

Data	Raw data	GA-PRDR	GA-NNRDR	GA-MINDC	GA-MAXPR	GA-QLPRP	GA-QNPRP
Horse-colic	0.845 1	**0.8511±0.0011**	**0.8511±0.0011**	0.8497±0.0021	0.7848±0.0844	0.8147±0.0522	0.8266±0.0709
Heart-statlog	0.788 9	0.8159±0.0220	**0.8219±0.0122**	0.7852±0.0250	0.7604±0.0710	0.7670±0.0605	0.7641±0.0307
Hepatitis	0.780 6	0.7858±0.0317	0.7890±0.0238	**0.8206±0.0160**	0.8019±0.0206	0.8123±0.0201	0.8071±0.0211
Sonar	0.802 9	0.7298±0.0270	**0.7317±0.0346**	0.7149±0.0345	0.7010±0.0459	0.7106±0.0372	0.7197±0.0452
Monks-1	0.822 6	**0.9597±0.0000**	**0.9597±0.0000**	0.9435±0.0260	0.8452±0.1470	0.8871±0.1386	0.8774±0.1442
Monks-3	0.934 4	**0.9344±0.0000**	**0.9344±0.0000**	**0.9344±0.0000**	0.7066±0.1938	0.7279±0.1759	0.6705±0.2025
Wdbc	0.940 2	**0.9450±0.0088**	**0.9450±0.0072**	0.9288±0.0186	0.9380±0.0105	0.9334±0.0151	0.9195±0.0180
Wpbc	0.757 6	0.7601±0.0165	**0.7621±0.0127**	0.7470±0.0178	0.7601±0.0154	0.7576±0.0136	0.7379±0.0185
Voting	0.963 2	**0.9632±0.0000**	**0.9632±0.0000**	0.9628±0.0014	0.9352±0.0651	0.9395±0.0365	0.9457±0.0403
Musk-1	0.758 4	0.7074±0.0226	0.7206±0.0230	**0.7538±0.0256**	0.7380±0.0211	**0.7538±0.0239**	0.7485±0.0211

Table 4 Classification accuracy comparison with Naïve Bayes

表 4 Naive Bayes 分类正确率比较

Data	Raw data	GA-PRDR	GA-NNRDR	GA-MINDC	GA-MAXPR	GA-QLPRP	GA-QNPRP
Horse-colic	0.788 0	0.8304±0.0039	**0.8318±0.0049**	0.8019±0.0090	0.7723±0.0597	0.7927±0.0382	0.7986±0.0552
Heart-statlog	0.829 6	0.8459±0.0155	**0.8496±0.0067**	0.8211±0.0162	0.7963±0.0752	0.7974±0.0732	0.7867±0.0511
Hepatitis	0.838 7	0.8277±0.0119	0.8155±0.0206	**0.8329±0.0102**	0.8097±0.0101	0.8265±0.0201	0.8200±0.0265
Sonar	0.778 8	0.7495±0.0209	**0.7678±0.0242**	0.7255±0.0164	0.7091±0.0334	0.7288±0.0244	0.7192±0.0259
Monks-1	0.774 2	**0.7581±0.0000**	**0.7581±0.0000**	0.7524±0.0089	0.7202±0.0622	0.7306±0.0693	0.7298±0.0638
Monks-3	0.934 4	**0.9344±0.0000**	**0.9344±0.0000**	**0.9344±0.0000**	0.7107±0.1901	0.7328±0.1700	0.6672±0.2078
Wdbc	0.935 0	**0.9443±0.0064**	0.9380±0.0143	0.9241±0.0181	0.9311±0.0112	0.9274±0.0140	0.9248±0.0163
Wpbc	0.661 6	0.7561±0.0115	**0.7687±0.0112**	0.7369±0.0271	0.7323±0.0235	0.7116±0.0250	0.7167±0.0318
Voting	0.901 1	**0.9287±0.0046**	0.9264±0.0056	0.9074±0.0093	0.9074±0.0516	0.8938±0.0188	0.9129±0.0320
Musk-1	0.699 6	**0.7113±0.0164**	0.7095±0.0173	0.6958±0.0156	0.6947±0.0149	0.6977±0.0145	0.6998±0.0115

(2) 误分类代价

这个部分比较了不同属性约简定义所得到约简的误分类代价.我们采用文献[13]给出的评价标准,即,误分类代价被定义为

$$mc=\lambda_{PN}\cdot n_{PN}+\lambda_{NP}\cdot n_{NP} \tag{19}$$

其中,n_{PN} 和 n_{NP} 分别表示误分类对象数量,即错误的肯定数量和错误的否定数量.在实验中,代价敏感分类方法被利用[31],同样地,J48 与 Naive Bayes 分类器被用作基准分类器.

表 5 和表 6 分别给出了采用 J48 和 Naive Bayes 分类算法进行测试的实验结果,平均最小误分类代价加粗显示.从实验结果可以看出:两种决策域分布保持约简和决策风险最小化属性约简在多数情况下都获得了更低的误分类代价,与原始数据的误分类代价相比,两种决策域分布保持约简在多数情况下至少有 1 种分类器能够降低误分类代价.

因此,无论是分类正确率还是误分类代价,决策域分布保持约简都是一种更好的选择.

Table 5 Misclassification cost comparison with J48

表 5 J48 误分类代价比较

Data	Raw data	GA-PRDR	GA-NNRDR	GA-MINDC	GA-MAXPR	GA-QLPRP	GA-QNPRP
Horse-colic	37.69±6.42	36.91±6.35	**36.57±6.56**	36.61±6.48	47.26±11.56	41.86±9.70	38.81±10.24
Heart-statlog	39.78±9.74	**35.29±7.21**	35.27±7.06	41.00±11.49	42.35±11.76	44.06±18.87	44.30±12.47
Hepatitis	18.97±6.43	20.55±10.46	20.02±6.71	**17.23±5.63**	20.05±7.91	18.91±7.57	19.07±6.98
Sonar	24.57±6.48	31.38±9.27	31.15±9.66	**30.99±9.49**	37.57±12.64	33.89±10.43	31.21±11.38
Monks-1	19.46±5.43	**10.22±1.98**	**10.22±1.98**	11.55±2.60	16.48±9.02	12.13±3.24	13.27±4.36
Monks-3	6.34±0.92	**6.14±1.05**	**6.14± 1.05**	6.18±1.10	19.11±12.79	19.20±11.65	20.33±13.61
Wdbc	23.02±4.95	**21.63±4.71**	21.71±6.98	26.86±7.25	23.96±7.54	23.85±8.72	31.51±13.76
Wpbc	34.94±11.02	36.41±11.69	**35.53±10.80**	37.82±13.80	37.02±12.93	38.27±13.01	40.72±13.41
Voting	10.52±3.04	10.78±3.20	**10.45±3.14**	10.75±2.85	14.20±5.23	13.89±6.21	15.30±13.12
Musk-1	76.95±34.38	88.55±41.97	84.43±38.24	**76.45±36.00**	78.30±33.67	76.63±33.60	78.19±34.45

Table 6 Misclassification cost comparison with Naïve Bayes

表 6 Naive Bayes 误分类代价比较

Data	Raw data	GA-PRDR	GA-NNRDR	GA-MINDC	GA-MAXPR	GA-QLPRP	GA-QNPRP
Horse-colic	51.01±10.14	41.50±7.50	**41.14±6.29**	46.97±9.05	54.00±14.07	48.13±11.51	47.76±15.21
Heart-statlog	32.71±8.53	30.22±6.59	**29.54±7.19**	33.66±9.82	37.37±11.98	39.00±19.70	40.14±12.91
Hepatitis	17.29±4.50	17.18±5.33	18.07±5.90	**16.63±4.52**	19.43±5.60	17.56±5.43	18.90±5.85
Sonar	27.76±6.33	30.04±10.20	**27.74±9.89**	32.06±10.76	36.04±11.64	34.17±7.99	33.14±11.10
Monks-1	21.67±5.02	22.55±5.83	22.5±5.83	**22.44±5.65**	24.71±3.73	23.09±4.37	23.61±4.43
Monks-3	6.06±1.47	**5.87±1.26**	**5.87±1.26**	5.91±1.31	18.68±12.91	19.58±12.44	21.14±14.60
Wdbc	24.41±6.71	**21.20±6.64**	24.12±11.30	28.28±8.38	25.91±7.76	27.02±8.17	28.36±9.08
Wpbc	48.04±9.91	35.82±12.36	**35.53±10.80**	39.35±12.09	36.77±13.23	40.98±12.22	38.91±10.52
Voting	25.36±7.93	19.69±5.92	20.25±6.24	23.64±9.29	**19.53±4.96**	24.17±8.05	22.20±11.14
Musk-1	97.07±32.01	89.69±40.35	**89.42±38.37**	96.76±33.49	97.63±34.38	97.49±34.00	96.63±34.97

(3) 约简平均长度

表 7 列出了用遗传算法求解每种约简的平均长度与标准差,其中,$|U|$是对象的个数,$|C|$是条件属性的个数.

从约简长度的标准差可以看出:因为在我们的算法中加入了修正算子,使得遗传算法找到的是约简本身而不是约简的超集,因此,本文算法求得的约简长度较稳定.这也是由于决策域分布保持约简的定义更严格,因为它要求每一个决策类相应的决策域都不变.

Table 7 Average length of a reduct based on GA

表 7 基于遗传算法的约简平均长度

Data	$\|U\|$	$\|C\|$	GA-PRDR	GA-NNRDR	GA-MINDC	GA-MAXPR	GA-QLPRP	GA-QNPRP
Horse-colic	268	27	9.0±0.0	9.0±0.0	14.6±1.7	11.5±2.3	11.2±1.9	11.5±4.1
Heart-statlog	270	13	10.2±0.4	10.1±0.3	11.4±1.4	10.9±4.2	10.6±4.6	8.7±4.9
Hepatitis	150	19	8.4±0.7	8.0±0.0	10.2±1.0	7.5±3.0	8.0±2.8	8.1±3.0
Sonar	208	60	7.0±0.0	6.9±0.3	20.2±1.3	20.4±1.1	21.1±1.0	20.0±1.6
Monks-1	124	6	3.0±0.0	3.0±0.0	3.3±0.5	2.7±1.2	2.7±0.9	2.9±1.0
Monks-3	122	6	4.0±0.0	4.0±0.0	4.2±0.4	1.5±0.9	1.5±0.9	1.5±0.9
Wdbc	569	30	7.4±0.5	7.6±0.5	10.7±0.8	10.6±1.0	10.3±1.7	10.5±1.4
Wpbc	198	33	7.0±0.0	7.0±0.0	11.8±0.6	11.1±1.4	11.5±0.9	11.9±0.8
Voting	435	16	12.0±0.0	12.0±0.0	13.7±2.6	8.2±6.4	11.7±5.6	9.1±5.8
Musk-1	476	166	11.0±0.0	11.0±0.0	67.2±2.6	65.8±2.6	66.2±6.9	65.7±3.7

5.2 最小属性约简

为了验证本文提出的遗传算法是否能够找到最小约简,我们选取了一组 UCI 数据集进行实验.在实验中,每个决策类的阈值设为α_j=0.8 和β_j=0.4.对于每个数据集,利用遗传算法分别对两种决策域分布保持约简实验 20 次,遗传算法参数设置见表 2.实验结果见表 8,其中,PRDPR 表示正域分布保持约简,NNRDPR 表示非负域分布保持约简,$|Core|$是核属性的个数,最小约简长度用 Min-$|R|$表示,它是通过回溯算法[34]求出的.GA-$|R|$表示遗传算法获得的约简长度.在 GA-$|R|$一列中,上标位置括号中的数字表示在这 20 次实验中获得该约简长度的实验次数.比如,在数据集 Hepatitis 上的实验结果,$8^{(17)}9^{(3)}$表示在 20 次运行中有 17 次得到的约简长度为 8,3 次得到的约简长度为 9.

Table 8 Experimental result of GA

表 8 遗传算法实验结果

Data sets	$\|U\|$	$\|C\|$	PRDPR			NNRDPR		
			$\|Core\|$	Min-$\|R\|$	GA-$\|R\|$	$\|Core\|$	Min-$\|R\|$	GA-$\|R\|$
Zoo	101	16	2	5	$5^{(20)}$	2	5	$5^{(20)}$
Lymphography	148	18	0	7	$7^{(17)}8^{(2)}9^{(1)}$	0	7	$7^{(17)}8^{(3)}$
Horse-colic	368	22	4	9	$9^{(20)}$	4	9	$9^{(20)}$
Heart-statlog	270	13	6	10	$10^{(16)}11^{(4)}$	6	10	$10^{(17)}11^{(3)}$
Hepatitis	150	19	0	8	$8^{(17)}9^{(3)}$	8	0	$8^{(17)}9^{(3)}$
Monks-1	124	6	3	3	$3^{(20)}$	3	3	$3^{(20)}$
Wdbc	569	30	0	7	$7^{(16)}8^{(4)}$	0	7	$7^{(16)}8^{(4)}$
Wpbc	198	33	0	7	$7^{(20)}$	0	7	$7^{(20)}$
Voting	435	16	11	12	$12^{(20)}$	11	12	$12^{(20)}$
SPECTF-Heart	267	44	0	7	$7^{(16)}8^{(4)}$	0	7	$7^{(16)}8^{(4)}$
Kr-vs-kp	3 196	37	27	29	$29^{(20)}$	27	29	$29^{(20)}$
Soybean	683	35	8	11	$11^{(20)}$	8	11	$11^{(20)}$

从实验结果可以看出:遗传算法每次运行可能会找到不同的约简;但在大多数运行情况下,表示遗传算法通常可以找到最小约简.比如,在数据集 Wpbc 上,20 次实验都找到了最小属性约简.这是因为本文的适应度函数能够保证最小约简与适应度函数最大化问题等价.

6 结 论

决策粗糙集是一种典型的概率粗糙集,由于引入了概率阈值,属性的增减与正域(或非负域)的变化不再似经典粗糙集理论中具备单调性.本文分析了决策粗糙集中属性约简中存在的一些问题.为了不改变决策域(正域或非负域),在决策粗糙集中引入了(α,β)正域分布保持约简和(α,β)非负域保持约简.此外,属性增减与决策域变化之间的非单调性给算法设计带来了一定的困难.为了简化算法设计,通过条件信息量的变形,提出了(α,β)正域

分布条件信息量和(α,β)非负域分布条件信息量,并证明其具有单调性,它们分别作为设计计算(α,β)正域分布保持约简和(α,β)非负域保持约简的启发式信息;同时,本文也给出了两种决策域分布保持约简的核属性以及求核算法.为了获得最小约简,本文还提出了一种基于遗传算法的启发式约简算法.在遗传算法中,我们的适应度函数能够保证最优解与最小约简相对应.另外,通过上述两种满足单调性的决策域分布条件信息量,在遗传算法中加入了一种新的算子,称为修正算子.修正算子确保遗传算法找到的是约简本身而不是约简的超集.最后,通过实验的方法从分类正确率和误分类代价说明了决策域分布保持约简与其他几种属性约简定义之间的性能差异,并且验证了遗传算法求解最小属性约简的有效性.

References:

[1] Pawlak Z. Rough Sets-Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer Academic, 1991.

[2] Yao YY, Zhao Y. Attribute reduction in decision-theoretic rough set models. Information Sciences, 2008,178(17):3356−3373. [doi: 10.1016/j.ins.2008.05.010]

[3] Shen Q, Jensen R. Rough sets, their extensions and applications. Int'l Journal of Automation and Computing, 2007,4(3):217−228. [doi: 10.1007/s11633-007-0217-y]

[4] Wu WZ, Leung Y, Shao MW. Generalized fuzzy rough approximation operators determined by fuzzy implicators. Int'l Journal of Approximation Reasoning, 2013,54(9):1388−1409. [doi: 10.1016/j.ijar.2013.05.004]

[5] Hu QH, Yu DR, Xie ZX. Numerical attribute reduction based on neighborhood granulation and rough approximation. Ruan Jian Xue Bao/Journal of Software, 2008,19(3):640−649 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/640.htm [doi: 10.3724/SP.J.1001.2008.00640]

[6] Yao YY, Wong SKM, Lingras P. A decision-theoretic rough set model. In: Ras ZW, Zemankova M, Emrich ML, eds. Proc. of the Methodologies for Intelligent Systems. New York: North-Holland, 1990. 17−24.

[7] Pawlak Z, Wong SKM, Ziarko W. Rough sets: Probabilistic versus deterministic approach. Int'l Journal of Man-Machine Studies, 1988,29(1):81−95. [doi: 10.1016/S0020-7373(88)80032-4]

[8] Ziarko W. Variable precision rough set model. Journal of Computer and System Science, 1993,46(1):39−59. [doi: 10.1016/0022-0000(93)90048-2]

[9] Slezak D, Ziarko W. Attribute reduction in the Bayesian version of variable precision rough set model. Electronic Notes in Theoretical Computer Science, 2003,82(4):263−273. [doi: 10.1016/S1571-0661(04)80724-2]

[10] Inuiguchi M. Attribute reduction in variable precision rough set model. Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2006,14(4):461−479. [doi: 10.1142/S0218488506004126]

[11] Mi JS, Wu WZ, Zhang WX. Approaches to knowledge reductions based on variable precision rough set model. Information Sciences, 2004,159(3-4):255−272. [doi: 10.1016/j.ins.2003.07.004]

[12] Zhou J, Miao DQ. β-Interval attribute reduction in variable precision rough set model. Soft Computing, 2011,15(8):1643−1656. [doi: 10.1007/s00500-011-0693-4]

[13] Jia XY, Liao WH, Tang ZM, Shang L. Minimum cost attribute reduction in decision-theoretic rough set models. Information Sciences, 2013,219(10):151−167. [doi: 10.1016/j.ins.2012.07.010]

[14] Yao YY. Probabilistic rough set approximations. Int'l Journal of Approximation Reasoning, 2008,49(2):255−271. [doi: 10.1016/j.ijar.2007.05.019]

[15] Yao YY. Two semantic issues in a probabilistic rough set model. Fundamenta Informaticae, 2010,108(3-4):1−17. [doi: 10.3233/FI-2011-422]

[16] Li HX, Zhou XZ. Risk decision making based on decision-theoretic rough set: A three-way view decision model. Int'l Journal of Computational Intelligence Systems, 2011,4(1):1−11. [doi: 10.1080/18756891.2011.9727759]

[17] Herbert JP, Yao JT. Game-Theoretic rough sets. Fundamenta Informaticae, 2011,108(3-4):267−286. [doi: 10.3233/FI-2011-423]

[18] Yu H, Chu SS, Yang DC. Autonomous knowledge-oriented clustering using decision-theoretic rough set theory. Fundamenta Informaticae, 2012,115(2-3):141−156. [doi: 10.3233/FI-2012-646]

[19] Zhou B. Multi-Class decision-theoretic rough sets. Int'l Journal of Approximation Reasoning, 2014,55(1):211−224. [doi: 10.1016/j.ijar.2013.04.006].

[20] Qian YH, Zhang H, Sang YL, Liang JY. Multigranulation decision-theoretic rough sets. Int'l Journal of Approximation Reasoning, 2014,55(1):255−237. [doi: 10.1016/j.ijar.2013.03.004]

[21] Zhao Y, Wong SKM, Yao YY. A note on attribute reduction in the decision-theoretic rough set model. In: Peter JF, Skowron A, Chan CC, Grzymala-Busse JW, Ziarko W, eds. Proc. of the Trans. on Rough Sets XIII. LNCS 6499, Heidelberg: Springer-Verlag, 2011. 260−275.

[22] Wong SKM, Ziarko W. On optimal decision rules in decision tables. Bulletin of Polish Academy of Sciences, 1985,33:693−696.

[23] Ye DY, Chen ZJ, Ma SL. A novel and better fitness evaluation for rough set based minimum attribute reduction problem. Information Sciences, 2013,222(10):413−423. [doi: 10.1016/j.ins.2012.08.020]

[24] Holland JH. Adaptation in Natural and Artificial Systems. Ann Arbor: University of Michigan Press, 1975.

[25] Frank A, Asuncion A.. UCIrvine machine learning repository. 2011. http://archive.ics.uci.edu/ml

[26] Liu D, Li TR, Li HX. A multiple-category classification approach with decision-theoretic rough sets. Fundamenta Informaticae, 2012,155(3-4):173−188. [doi: 10.3233/FI-2012-648]

[27] Yao YY, Zhao Y, Wang J. On reduct construction algorithms. In: Gavrilova ML, Tan CJK, Wang YX, Yao YY, Wang GY, eds. Proc. of the Trans. on Computational Science II. LNCS 5150, Heidelberg: Springer-Verlag, 2008. 100−117. [doi: 10.1007/978-3-540-87563-5_6]

[28] Wang GY, Yu H, Yang DC. Decision table reduction based on conditional information entropy. Journal of Computers, 2002,2(7): 759−766 (in Chinese with English abstract). [doi: 10.3321/j.issn:0254-4164.2002.07.013]

[29] Chebrolu S, Sanjeevi SG. Attribute reduction in decision-theoretic rough set models using genetic algorithm. In: Panigrahi BK, Suganthan PN, Das S, Satapathy SC, eds. Proc. of the Swarm, Evolutionary, and Memetic Computing. LNCS 7076, Heidelberg: Springer-Verlag, 2011. 307−314. [doi: 10.1007/978-3-642-27172-4_38]

[30] Liu ZH, Liu SY, Wang J. An attribute reduction algorithm based on the information quantity. Journal of Xidian University, 2003, 30(6):835−838 (in Chinese with English abstract). [doi: 10.3969/j.issn.1001-2400.2003.06.028]

[31] Hall M, Frank E, Holmes G, Pfahringer B, Recutemann P, Witten IH. The WEKA data mining software: An update. SIGKDD Explorations, 2009,11(1):10−18. [doi: 10.1145/1656274.1656278]

[32] Quinlan JR. C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann Publishers, 1993.

[33] Kantardzic M. Data Mining-Concepts, Models, Methods, and Algorithms. 2nd ed., Hoboken: Wiley-IEEE Press, 2011.

[34] Min F, Zhu W. Attribute reduction of data with error ranges and test costs. Information Sciences, 2012,211(30):48−67. [doi: 10.1016/j.ins.2012.04.031]

附中文参考文献:

[5] 胡清华,于达仁,谢宗霞.基于邻域粒化和粗糙逼近的数值属性约简.软件学报,2008,19(3):640−649. http://www.jos.org.cn/1000-9825/19/640.htm [doi: 10.3724/SP.J.1001.2008.00640]

[28] 王国胤,于洪,杨大春.基于条件信息熵的决策表约简.计算机学报,2002,25(7):759−766. [doi: 10.3321/j.issn:0254-4164.2002.07.013]

[30] 刘振华,刘三阳,王珏.基于信息量的一种属性约简算法.西安电子科技大学学报,2003,30(6):835−838. [doi: 10.3969/j.issn.1001-2400.2003.06.028]

马希骜(1985－),男,宁夏银川人,博士生,主要研究领域为决策粗糙集,数据挖掘.
E-mail: maxiao73559@163.com

于洪(1972－),女,博士,教授,CCF 会员,主要研究领域为粗糙集,智能信息处理与Web 智能,数据挖掘.
E-mail: hongyu.cqupt@gmail.com

王国胤(1970－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为粗糙集,粒计算,机器学习,数据挖掘,知识技术,认知计算.
E-mail: wanggy@ieee.org