ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 3,May 2025,pp. 807－820
DOI:10. 16337/j. 1004-9037. 2025. 03. 019

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于决策代价融合度量的不完备邻域决策粗糙集属性约简

张万祥[1,2]，张贤勇[1,2]，杨霁琳[2,3]，陈本卫[1,4]

(1. 四川师范大学数学科学学院，成都 610066;2. 四川师范大学智能信息与量子信息研究所，成都 610066;
3. 四川师范大学计算机科学学院，成都 610101;4. 西华师范大学数学与信息学院，南充 637009)

摘　要：属性约简依赖于知识粒化和不确定性度量，有助于智能识别。针对不完备连续型数据，邻域决策粗糙集诱导了属性约简，但相关的邻域关系需要优化改进，同时存在的决策代价值需要集成强化。本文提出一种新的邻域关系并组建3种决策代价融合度量，构造不完备邻域决策粗糙集并系统研究属性约简。首先，通过改进的距离函数引入不完备邻域关系，提出一种改进的不完备邻域决策粗糙集模型。然后，基于决策代价引入依赖度和邻域熵，采用乘法融合得到3种决策代价融合度量，研究粒化非单调性。进而，基于2种邻域关系和4种决策代价相关度量，采用属性重要度设计8种启发式约简算法。数据实验表明，本文所提的7种新算法中有5种算法具有较好的分类学习性能，改进了基础约简算法。

关键词：属性约简；粗糙集；不完备邻域关系；不确定性度量；决策代价

中图分类号：TP18　　**文献标志码**：A

Attribute Reduction of Incomplete Neighborhood Decision Rough Sets Based on Decision-Cost Fusion Measures

ZHANG Wanxiang[1,2], ZHANG Xianyong[1,2], YANG Jilin[2,3], CHEN Benwei[1,4]

(1. School of Mathematical Sciences, Sichuan Normal University, Chengdu 610066, China; 2. Institute of Intelligent Information and Quantum Information, Sichuan Normal University, Chengdu 610066, China; 3. College of Computer Science, Sichuan Normal University, Chengdu 610101, China; 4. School of Mathematics and Information, China West Normal University, Nanchong 637009, China)

Abstract: Attribute reduction relies on knowledge granulation and uncertainty measurement, thus facilitating intelligent recognition. For incomplete continuous data, neighborhood decision rough sets induce attribute reduction. However, the related neighborhood relation deserves optimal improvements, while the existing decision cost deserves integrated reinforcements. In this paper, a new neighborhood relation is proposed, and three decision-cost fusion measures are constructed, so new incomplete neighborhood decision rough sets are established and the attribute reduction is systematically researched. At first, an improved distance is introduced to produce an incomplete neighborhood relation, so improved rough sets on incomplete neighborhood are proposed. Then, the dependence degree and neighborhood entropy are introduced based on decision costs, so three fusion measures on decision costs are obtained by multiplication fusion, thus acquiring granulation non-monotonicity. Furthermore, eight heuristic reduction

基金项目：国家自然科学基金(61976158)；四川省自然科学基金(2024NSFSC0486，2024NSFSC0443)；四川省科技计划(2022ZYD0001)；教育部人文社科规划基金(23YJA630114)。
收稿日期：2024-04-10；**修订日期**：2024-08-17

algorithms based on attribute importances are designed from two neighborhood relations and four relevant measures of decision costs. As finally verified by data experiments, the five algorithms out of the seven new algorithms have good performance of classification learning, thus improving the basic reduction algorithm.

Key words: attribute reduction; rough set; incomplete neighborhood relation; uncertainty measure; decision cost

引　言

粗糙集理论(尤其是属性约简)在人工智能与模式识别等领域具有广泛研究与应用。Pawlak粗糙集模型[1]主要基于等价关系,难以处理具有噪声的实际数据,需要拓展。通过引入概率信息,Ziarko[2]提出几种广义粗糙集模型;Yao等[3]采用贝叶斯风险决策,建立具有最小决策代价的决策(理论)粗糙集模型。当前,决策粗糙集已经广泛应用于各种信息系统的不确定建模与属性约简。张静等[4]在多粒度信息系统中建立一种柔性多粒化的决策粗糙集;Zhao等[5]在集值信息系统中建立一种新的多集值决策粗糙集;Li等[6]在邻域信息系统中引入决策粗糙集来处理数值数据;针对连续信息系统,Gao等[7]提出一种决策粗糙集从而建立基于最大决策熵的属性约简,张敏等[8]提出一种基于类别可区分度的属性约简算法;Song等[9]定义模糊决策粗糙集进而构建基于最小决策代价的属性约简算法。

在实际应用中,存在许多不完备连续信息系统[10-11],值得采用粗糙集扩张模型进行数据处理,但相关的决策粗糙集及其属性约简还需要深入。Liu等[12]提出不完备信息系统下的决策粗糙集;姚晟等[13]提出不完备邻域粗糙集和邻域混合熵,处理混合属性和不完备数据的属性约简;蔡艳婧等[14]在不完备混合信息系统中提出一种不完备混合决策粗糙集,构造特定类多目标代价敏感属性约简算法;姚晟等[15]在不完备连续型信息系统中引入不完备邻域关系,提出不完备邻域决策粗糙集,构建基于最小化决策代价的属性约简算法。不完备数据系统的属性约简值得综合考虑决策粗糙集度量与不确定性熵度量,从而进行算法发展与业绩提升。

聚焦文献[15],不完备邻域决策粗糙集依赖于邻域关系,而属性约简还依赖于决策代价度量,使得所建立约简算法能够有效处理不完备连续型数据。但是,存在的邻域关系比较严格,使用的决策代价比较单一,因此相关约简算法还具有提升空间,本文主要对此实施改进。针对不完备连续信息系统,本文提出改进的邻域关系与构建决策代价融合度量,推进不确定性建模与约简算法优化,研究框架如图1所示。

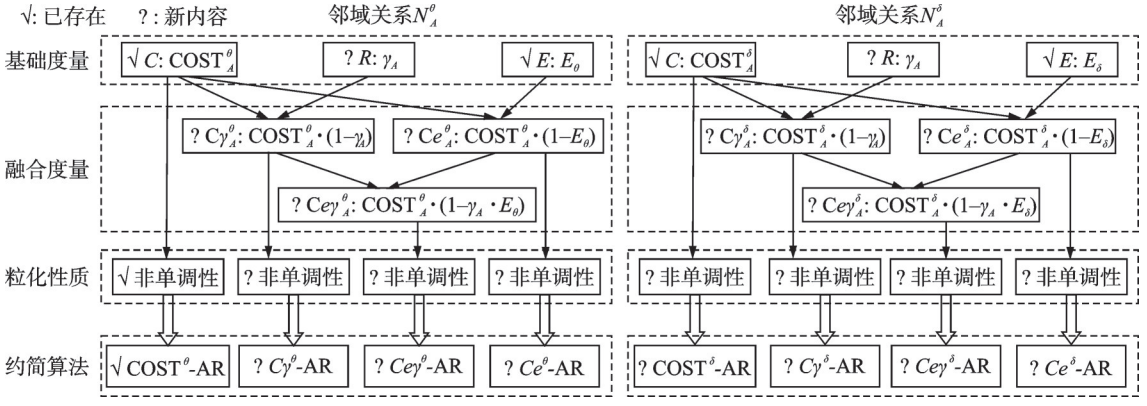

图1　不完备连续信息系统上基于邻域与度量的属性约简

Fig.1　Neighborhood and measure based attribute reduction of incomplete continuous information system

图1中的变量含义见下文相关章节。本文的内容创新如下:(1)优化邻域关系,基于旧新邻域关系来获取2种知识粒化;(2)引入依赖度和邻域熵,提出基于决策代价的融合构建方法,得到3种具体融合度量;(3)基于组合度量与粒化性质,构建$2\times4=8$种非单调的属性约简算法(包含当前算法COST^{θ}-AR[15]与7种新算法),系统获取不完备邻域决策粗糙集的优化学习业绩。

1 不完备邻域决策粗糙集

传统决策粗糙集主要建立在完备的离散信息系统之上,然而实际应用中存在许多不完备连续信息系统,因此不完备邻域决策粗糙集可以通过不完备邻域关系处理不完备连续型数据,得到邻域类来代替决策粗糙集中的等价类,进一步通过决策区域划分得到正区域、边界域和负区域,从而完成三支决策制定。本节主要介绍不完备邻域决策粗糙集的内容,为此首先介绍相关的决策信息系统。

给定决策信息系统$S=(U,AT=C\cup\{d\})$,其中论域$U=\{x_1,x_2,\cdots,x_{|U|}\}$是一组非空有限对象集,$AT$为信息系统的属性集($C$为连续数据构成的条件属性集,$d$为决策属性)。论域中的对象$x\in U$在属性$a\in C$下的属性值表示为$a(x)$。属性集$A\subseteq C$在论域$U$下的等价关系定义为$R=\{(x,y)\in U|\forall a\in A,a(x)=a(y)\}$,对应的等价类为$[x]_R=\{y|a(y)=a(x)\}$。

定义1[16-17] 针对完备决策信息系统$S=(U,AT=C\cup\{d\})$,$x\in U$在条件属性子集$A=\{a_1,a_2,\cdots,a_n\}\subseteq C$上的邻域为

$$\delta_A(x)=\{y\in U|\Delta_A(x,y)\leqslant\delta\} \tag{1}$$

式中:δ为邻域半径;Δ主要采用欧式距离,有

$$\Delta_A(x,y)=\sqrt{\sum_{k=1}^{n}(a_k(x)-a_k(y))^2} \tag{2}$$

作为完备决策信息系统的拓展,下面聚焦不完备决策信息系统,其具有空属性值$*$(即$\exists x\in U,\exists a\in C$,有$a(x)=*$)。对此,下文涉及的$S=(U,AT=C\cup\{d\})$均指不完备决策信息系统,并主要考虑属性子集$A\subseteq C$进行条件粒化,而决策粒化涉及决策分类$U/d=\{D_1,D_2,\cdots,D_m\}$(具有$m$个决策类)。

定义2 针对$S=(U,AT=C\cup\{d\})$及$A=\{a_1,a_2,\cdots,a_n\}\subseteq C$,$x$、$y\in U$关联的距离函数为

$$N_A(x,y)=\sqrt{\sum_{i=1}^{n}N_i^2(x,y)} \tag{3}$$

$$N_i(x,y)=\begin{cases}a_i(x)-a_i(y) & a_i(x)\neq *\wedge a_i(y)\neq *\\ 0 & a_i(x)=*\vee a_i(y)=*\end{cases} \tag{4}$$

进而,不完备邻域关系及对应邻域分别为

$$N_A^{\delta}=\{(x,y)\in U\times U|N_A(x,y)\leqslant\delta\},n_A^{\delta}(x)=\{y\in U|(x,y)\in N_A^{\delta}\} \tag{5}$$

式中$\delta\geqslant0$称为不完备邻域半径。

针对不完备决策信息系统,定义2提出一种邻域粒化,主要拓展完备情况下的欧式距离粒化,并参考与修正了文献[13]的对应公式。此外,文献[15]提供了一种不完备邻域关系,即

$$N_A^{\theta}=\left\{(x,y)\in U\times U|\forall a\in A,(a(x)=*)\vee(a(y)=*)\vee(|a(x)-a(y)|\leqslant\theta)\right\} \tag{6}$$

其粒化涉及邻域$n_A^{\theta}(x)=\{y\in U|(x,y)\in N_A^{\theta}\}$。这两种邻域结构可以采用参数$\delta$、$\theta$来区分。本文主要采用前者(即式(5)或定义2)叙述,最后比较两种邻域结构得到的分类效果。下面模拟文献[13]的邻域熵自然建立关于N_A^{δ}的邻域熵。

定义3 针对$S=(U,AT=C\cup\{d\})$及A,邻域信息熵为

$$E_\delta(A)=\frac{1}{|U|}\cdot\left(1-\frac{|n_A^\delta(x)|}{|U|}\right) \tag{7}$$

加拿大学者Yao采用最小化决策代价原则,提出依托等价类$[x]$的传统决策(理论)粗糙集[3]。对于决策粗糙集,X和$\neg X$分别表示元素在X中或者不在X中的两种状态,a_P、a_B、a_N分别表示1个对象x在状态X中对应的接受、推迟、拒绝3种动作,相关的决策代价3×2矩阵如表1所示。对象x在两种互补状态集下的条件概率为$P(X|[x])=\dfrac{|X\cap[x]|}{|[x]|}$和$P(\neg X|[x])=1-\dfrac{|X\cap[x]|}{|[x]|}$。根据合理损失条件:$\lambda_{PP}\leqslant\lambda_{BP}<\lambda_{NP}$,$\lambda_{NN}\leqslant\lambda_{BN}<\lambda_{PN}$,可得最小风险决策规则:(1)若$P(X|[x])\geqslant\alpha$且$P(X|[x])\geqslant\gamma$,则$x\in\mathrm{POS}(X)$;(2)若$P(X|[x])\leqslant\alpha$且$P(X|[x])\geqslant\beta$,则$x\in\mathrm{BND}(X)$;(3)若$P(X|[x])\leqslant\beta$且$P(X|[x])\leqslant\gamma$,则$x\in\mathrm{NEG}(X)$。其中,阈值参数$\alpha=\dfrac{\lambda_{PN}-\lambda_{BN}}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})}$,$\beta=\dfrac{\lambda_{BN}-\lambda_{NN}}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{BP})}$,$\gamma=\dfrac{\lambda_{PN}-\lambda_{NN}}{(\lambda_{PN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})}$。

表1 决策代价矩阵

Table 1 Decision cost matrix

动作集	X	$\neg X$
a_P	λ_{PP}	λ_{PN}
a_B	λ_{BP}	λ_{BN}
a_N	λ_{NP}	λ_{NN}

姚晟等[15]模拟上述经典决策粗糙集建模,提出不完备邻域信息系统下的决策粗糙集,主要将等价类$[x]$替换与拓展为邻域$n_A^\theta(x)$。下面考虑新建邻域$n_A^\delta(x)$,得到对应的邻域决策粗糙集及相关区域与代数依赖度。

定义4 针对$S=(U,AT=C\cup\{d\})$并关于参数分布$0\leqslant\beta\leqslant\alpha\leqslant1$,决策类$D_j\in U/\{d\}$的决策粗糙集正区域、边界域、负区域分别为

$$\begin{cases}\mathrm{POS}_A^{\alpha,\beta}(D_j)=\left\{x|P(D_j|n_A^\delta(x))\geqslant\alpha\right\}\\ \mathrm{BND}_A^{\alpha,\beta}(D_j)=\left\{x|\beta<P(D_j|n_A^\delta(x))<\alpha\right\}\\ \mathrm{NEG}_A^{\alpha,\beta}(D_j)=\left\{x|P(D_j|n_A^\delta(x))\leqslant\beta\right\}\end{cases} \tag{8}$$

式中$P(D_j|n_A^\delta(x))=\dfrac{|D_j\cap n_A^\delta(x)|}{|n_A^\delta(x)|}$为邻域$n_A^\delta(x)$对于决策类$D_j$的条件概率。进而关于决策分类$U/\{d\}$可以确定正区域与依赖度,有

$$\begin{cases}\mathrm{POS}_A^\delta(d)=\bigcup\limits_{j=1}^{m}\mathrm{POS}_A^{\alpha,\beta}(D_j)\\ \gamma_A(d)=\dfrac{|\mathrm{POS}_A^\delta(d)|}{|U|}\end{cases} \tag{9}$$

2 决策代价融合度量

文献[15]关联于不完备邻域决策粗糙集,提出决策代价,并由此采用最小化决策代价来确立属性约简算法。本节首先引入$n_A^\delta(x)$邻域表示的决策代价,再考虑加入依赖度与邻域熵从而建立3种决策代价融合度量,为后续属性约简奠定基础。

定义5 针对$S=(U,AT=C\cup\{d\})$及A,决策类D_j的决策代价为

$$\mathrm{COST}_A^{\delta}(D_j)=\sum_{x\in \mathrm{POS}_A^{\alpha,\beta}(D_j)}(p\cdot\lambda_{PP}+(1-p)\cdot\lambda_{PN})+\sum_{x\in \mathrm{BND}_A^{\alpha,\beta}(D_j)}(p\cdot\lambda_{BP}+(1-p)\cdot\lambda_{BN})+\sum_{x\in \mathrm{NEG}_A^{\alpha,\beta}(D_j)}(\lambda_{NP}+(1-p)\cdot\lambda_{NN}) \tag{10}$$

式中$p=P(D_j|n_A^{\delta}(x))$。进而决策分类单参数决策代价为

$$\mathrm{COST}_A^{\delta}(d)=\sum_{j=1}^{m}\mathrm{COST}_A^{\delta}(D_j) \tag{11}$$

这里,决策代价$\mathrm{COST}_A^{\delta}(d)$同构采用文献[15]的定义公式,只是依托邻域$n_A^{\delta}(x)$。换句话讲,文献[15]采用的决策代价是依托邻域$n_A^{\theta}(x)$的$\mathrm{COST}_A^{\theta}(d)$。下面主要采用$\mathrm{COST}_A^{\delta}(d)$来叙述。根据定义5,可以解释相关的成本函数选择。关联于贝叶斯优化与三支决策,首先得到决策类层的$\mathrm{COST}_A^{\delta}(D_j)$,最后采用层次集成求和自然得到决策分类层的$\mathrm{COST}_A^{\delta}(d)$,从而支撑后续属性约简与分类学习。

定义6 针对$S=(U,AT=C\cup\{d\})$,定义3种决策代价融合度量

$$\begin{cases}C\gamma_A^{\delta}(d)=\mathrm{COST}_A^{\delta}(d)\cdot(1-\gamma_A(d))\\ Ce_A^{\delta}(d)=\mathrm{COST}_A^{\delta}(d)\cdot(1-E_{\delta}(A))\\ Ce\gamma_A^{\delta}(d)=\mathrm{COST}_A^{\delta}(d)\cdot(1-\gamma_A(d)\cdot E_{\delta}(A))\end{cases} \tag{12}$$

关于度量机制,$\mathrm{COST}_A^{\delta}(d)$来源于三支决策及成本统计,$\gamma_A(d)$来源于决策粗糙集分类正域,而$E_{\delta}(A)$关联于知识粒化信息,三者具有异质性与系统性,相关的融合能够产生更全面与更强大的新度量。在融合构建中,定义6围绕决策代价$\mathrm{COST}_A^{\delta}(d)$,引入依赖度$\gamma_A(d)$和邻域熵$E_{\delta}(A)$建立3种决策代价融合度量,即$C\gamma_A^{\delta}(d)$、$Ce_A^{\delta}(d)$、$Ce\gamma_A^{\delta}(d)$,其中最后的$Ce\gamma_A^{\delta}(d)$完全融合了3个基础度量。

性质1 4个度量$\mathrm{COST}_A^{\delta}(d)$、$C\gamma_A^{\delta}(d)$、$Ce_A^{\delta}(d)$、$Ce\gamma_A^{\delta}(d)$呈现的大小关系为

$$\mathrm{COST}_A^{\delta}(d)\geqslant Ce\gamma_A^{\delta}(d),Ce\gamma_A^{\delta}(d)\geqslant Ce_A^{\delta}(d),Ce\gamma_A^{\delta}(d)\geqslant C\gamma_A^{\delta}(d) \tag{13}$$

证明 根据定义3、4、6,可得$\gamma_A(d)$、$E_{\delta}(A)\in[0,1]$,由此可证。

定理1 (1)关于属性子集A的变化,$\mathrm{COST}_A^{\delta}(d)$具有粒化非单调性;即若$A\subseteq A'\subseteq C$,$\mathrm{COST}_A^{\delta}(d)\leqslant\mathrm{COST}_{A'}^{\delta}(d)$不恒成立而$\mathrm{COST}_A^{\delta}(d)\geqslant\mathrm{COST}_{A'}^{\delta}(d)$也不恒成立。(2)进而类似地,$C\gamma_A^{\delta}(d)$、$Ce_A^{\delta}(d)$和$Ce\gamma_A^{\delta}(d)$均具有粒化非单调性。

证明 (1)对于决策粗糙集,三支判定区域(POS、BND、NEG)对于属性集合包含关系$A\subseteq A'\subseteq C$不具有单调性(即具有粒化非单调性)。同时,虽然$n_A^{\delta}(x)\supseteq n_{A'}^{\delta}(x)$必然诱导$|n_A^{\delta}(x)|\geqslant|n_B^{\delta}(x)|$与$|D_j\cap n_A^{\delta}(x)|\geqslant|D_j\cap n_B^{\delta}(x)|$,但$P(D_j|n_A^{\delta}(x))=\dfrac{|D_j\cap n_A^{\delta}(x)|}{|n_A^{\delta}(x)|}$与$P(D_j|n_B^{\delta}(x))=\dfrac{|D_j\cap n_B^{\delta}(x)|}{|n_B^{\delta}(x)|}$不能确定大小关系,故$\mathrm{COST}_A^{\delta}(d)$具有粒化非单调性。实例佐证参见式(17)。(2)虽然$\gamma_A(d)$具有粒化单调性(即$n_A^{\delta}(x_i)\supseteq n_{A'}^{\delta}(x_i)\Rightarrow\bar{N}(D_j)\subseteq\bar{N}(D_j)\Rightarrow\gamma_A(d)\leqslant\gamma_{A'}(d)$,但$\mathrm{COST}_A^{\delta}(d)$的粒化非单调性进而诱导$C\gamma_A^{\delta}(d)$的粒化非单调性,故不能确定$\mathrm{COST}_A^{\delta}(d)\cdot(1-\gamma_A(d))$与$\mathrm{COST}_{A'}^{\delta}(d)\cdot(1-\gamma_{A'}(d))$的必然大小关系。同理可得$Ce_A^{\delta}(d)$和$Ce\gamma_A^{\delta}(d)$的粒化非单调性。针对这3种融合度量,相关的粒化非单调性佐证参见式(17)。

针对$\mathrm{COST}_A^{\delta}(d)$、$C\gamma_A^{\delta}(d)$、$Ce_A^{\delta}(d)$、$Ce\gamma_A^{\delta}(d)$,性质1自然给出4种度量的大小比较,而定理1挖掘4种度量的粒化非单调性,该深入性质可被后面的数据实验(如式(17))所验证。

根据定义3~6,算法1系统地计算出4种不确定性度量值。算法1有12个步骤,涉及3个模块。模块1对应步骤(1~5)计算$E_{\delta}(A)$,模块2对应步骤(6~10)计算$\gamma_A(d)$和$\mathrm{COST}_A^{\delta}(d)$,模块3对应步骤

(11)合成融合度量$C\gamma_A^\delta(d)$,$Ce_A^\delta(d)$,$Ce\gamma_A^\delta(d)$,最后步骤(12)输出度量计算结果。

算法1 计算4种度量$\mathrm{COST}_A^\delta(d)$,$C\gamma_A^\delta(d)$,$Ce_A^\delta(d)$,$Ce\gamma_A^\delta(d)$。

输入:不完备决策信息系统$S=(U,AT=C\cup\{d\})$,条件属性子集$A\subseteq C$。

输出:$\mathrm{COST}_A^\delta(d)$,$C\gamma_A^\delta(d)$,$Ce_A^\delta(d)$,$Ce\gamma_A^\delta(d)$。

(1) 初始化$E_\delta(A)=0$

(2) for $x\in U$

(3) $n_A^\delta(x)=\{y\in U|(x,y)\in N_A^\delta\}$

(4) $E_\delta(A)\leftarrow E_\delta(A)+\dfrac{1}{|U|}\cdot(1-\dfrac{|n_A^\delta(x)|}{|U|})$

(5) end for

(6) for $j=1:m$

(7) 通过式(8),计算$\mathrm{POS}_A^{\alpha,\beta}(D_j)$、$\mathrm{BND}_A^{\alpha,\beta}(D_j)$、$\mathrm{NEG}_A^{\alpha,\beta}(D_j)$

(8) 通过式(10),计算$\mathrm{COST}_A^\delta(D_j)$

(9) end for

(10) 通过式(9),计算$\gamma_A(d)$;通过式(11),计算$\mathrm{COST}_A^\delta(d)$

(11) 通过式(12),计算$C\gamma_A^\delta(d)$,$Ce_A^\delta(d)$,$Ce\gamma_A^\delta(d)$

(12) 返回$\mathrm{COST}_A^\delta(d)$,$C\gamma_A^\delta(d)$,$Ce_A^\delta(d)$,$Ce\gamma_A^\delta(d)$

例1 不完备决策信息系统$S=(U,AT=C\cup\{d\})$如表2所示,其中*表示缺失值。这里,$U=\{x_1,x_2,\cdots,x_{12}\}$,$C=\{c_1,c_2,c_3,c_4\}$,$d$具有属性值1、2、3分别诱导3分决策类:$D_1=\{x_3,x_4,x_8,x_9\}$,$D_2=\{x_6,x_{10},x_{12}\}$,$D_3=\{x_1,x_2,x_5,x_7,x_{11}\}$。

取邻域参数$\delta=0.1$,设置$\lambda_{PP}=\lambda_{NN}=0$,$\lambda_{BP}=\lambda_{BN}=0.2$,$\lambda_{NP}=\lambda_{PN}=1$可得决策粗糙集双阈值$\alpha=\dfrac{1-0.2}{(1-0.2)+(0.2-0)}=0.8$,$\beta=\dfrac{0.2-0}{(0.2-0)+(1-0.2)}=0.2$。条件属性子集$A\subseteq C$关注属性增链$A_1=\{c_1\}\subset A_2=\{c_1,c_2\}\subset A_3=\{c_1,c_2,c_3\}\subset A_4=\{c_1,c_2,c_3,c_4\}$。

表2 不完备决策信息系统实例

Table 2 Example of incomplete decision information system

U	c_1	c_2	c_3	c_4	d	U	c_1	c_2	c_3	c_4	d
x_1	250	0.180	0.30	1.10	3	x_7	270	0.165	0.30	0.83	3
x_2	256	0.162	0.36	0.63	3	x_8	280	*	0.30	3.40	1
x_3	286	0.183	0.84	2.98	1	x_9	*	0.100	0.52	3.14	1
x_4	288	0.250	0.78	3.32	1	x_{10}	*	*	0.80	1.30	2
x_5	234	0.167	0.30	0.78	3	x_{11}	*	0.164	0.33	0.76	3
x_6	259	0.235	0.75	1.22	2	x_{12}	226	*	0.73	1.10	2

首先,通过最大最小标准化将表2数据进行归一化处理.聚焦属性子集代表$A=A_3=\{c_1,c_2,c_3\}$,由定义2可得12个邻域:$n_A^\delta(x_1)=\{x_1\}$,$n_A^\delta(x_2)=\{x_2,x_{11}\}$,$n_A^\delta(x_3)=\{x_3,x_{10}\}$,$n_A^\delta(x_4)=\{x_4,x_{10}\}$,$n_A^\delta(x_5)=\{x_5,x_{11}\}$,$n_A^\delta(x_6)=\{x_6,x_{10}\}$,$n_A^\delta(x_7)=\{x_7,x_{11}\}$,$n_A^\delta(x_8)=\{x_8,x_{11}\}$,$n_A^\delta(x_9)=\{x_9\}$,$n_A^\delta(x_{10})=\{x_3,x_4,x_6,x_{10}\}$,$n_A^\delta(x_{11})=\{x_2,x_5,x_7,x_8,x_{11}\}$,$n_A^\delta(x_{12})=\{x_{12}\}$。

(1) 由定义3可得邻域熵$E_\delta(A)=0.819$。(2) 由定义4,12个邻域具有12个条件概率:

$P(D_3|n_A^\delta(x_1))=1, P(D_3|n_A^\delta(x_2))=1, P(D_3|n_A^\delta(x_3))=0, P(D_3|n_A^\delta(x_4))=0, P(D_3|n_A^\delta(x_5))=1, P(D_3|n_A^\delta(x_6))=0, P(D_3|n_A^\delta(x_7))=1, P(D_3|n_A^\delta(x_8))=0.5, P(D_3|n_A^\delta(x_9))=0, P(D_3|n_A^\delta(x_{10}))=0, P(D_3|n_A^\delta(x_{11}))=0.8, P(D_3|n_A^\delta(x_{12}))=0$。故可得$\mathrm{POS}_A^\delta(D_3)=\{x_1,x_2,x_5,x_7,x_{11}\}$。同理可以计算得到$\mathrm{POS}_A^\delta(D_1)=\{x_9\}, \mathrm{POS}_A^\delta(D_2)=\{x_6,x_{12}\}$。因此,依赖度$\gamma_A(d)=\dfrac{|\mathrm{POS}_A^\delta(d)|}{|U|}=\dfrac{8}{12}=0.667$。(3)由定义5,$\mathrm{COST}_A^\delta(D_1)=1$。同理可以计算得到$\mathrm{COST}_A^\delta(D_2)=0.6, \mathrm{COST}_A^\delta(D_3)=0.4$,因此总的决策代价$\mathrm{COST}_A^\delta(d)=2$。(4)由定义6,采用乘积集成可得:$C\gamma_A^\delta(d)=2\times(1-0.667)=0.667, Ce_A^\delta(d)=2\times(1-0.819)=0.361, Ce\gamma_A^\delta(d)=2\times(1-0.667\times0.819)=0.907\,4$。

如上4步计算完全吻合于算法1,即相关度量结果可以由算法1得到。类似地,可以计算关于A_1、A_2、A_4的4种度量,所有结果记录于表3。

表3 实例关于条件属性增链的多种度量值

Table 3 Example's multiple measure values based on addition chain of conditional attributes

度量	$\{c_1\}$	$\{c_1,c_2\}$	$\{c_1,c_2,c_3\}$	$\{c_1,c_2,c_3,c_4\}$
$(\mathrm{COST}_A^\delta(d),\gamma_A(d),E_\delta(A))$	(7.1,0,0.44)	(5.143,0.833,0.653)	(2,0.667,0.819)	(0,1,0.861)
$\mathrm{COST}_A^\delta(d)$	7.100	5.143	2.000	0.000
$C\gamma_A^\delta(d)$	7.100	4.714	0.667	0.000
$Ce_A^\delta(d)$	3.944	1.786	0.361	0.000
$Ce\gamma_A^\delta(d)$	7.100	4.863	0.907	0.000

基于表3可以自然检验相关的度量性质。性质1的大小关系明显。关于定理1,这里的$\mathrm{COST}_A^\delta(d)$、$C\gamma_A^\delta(d)$、$Ce_A^\delta(d)$、$Ce\gamma_A^\delta(d)$都是随着属性增加(邻域覆盖粗化)而减少,故只展现一种单调性,而非单调性将在后面的实验中观察得到。

3 基于决策代价融合度量的属性约简算法

本节讨论度量诱导的属性约简算法。特别地,文献[15]采用依托邻域$n_A^\theta(x)$的决策代价$\mathrm{COST}_A^\theta(d)$,采用添加-删除策略进行属性约简。现在,邻域关系$N_A^\theta$推进到$N_A^\delta$,决策代价被改进到3种融合度量,由此得到$2\times4=8$种度量,它们都可以采用文献[15]约简框架进行算法构建。8种度量统一设置为

$$\mathrm{Mea}_A^\heartsuit(d), \mathrm{Mea}\in\{\mathrm{COST}, C\gamma, Ce, Ce\gamma\}, \heartsuit\in\{\theta,\delta\} \tag{14}$$

相关的度量结构与算法符号参见表4(其中AR表示属性约简算法),其一致于图1的相关描述。下面,陈述基于代表度量$\mathrm{Mea}_A^\heartsuit(d)$的代表算法$\mathrm{Mea}^\heartsuit$-AR,其中采用了属性重要度。

表4 2种邻域关系和4种基础度量诱导的8个约简算法(约简度量)

Table 4 Eight reduction algorithms (reduction measures) based on two neighborhood relations and four basis measures

度量邻域关系	$\mathrm{COST}_A^\heartsuit(d)$	$C\gamma_A^\heartsuit(d)$	$Ce_A^\heartsuit(d)$	$Ce\gamma_A^\heartsuit(d)$
N_A^θ	COST^θ-AR ($\mathrm{COST}_A^\theta(d)$)	$C\gamma^\theta$-AR ($C\gamma_A^\theta(d)$)	Ce^θ-AR ($Ce_A^\theta(d)$)	$Ce\gamma^\theta$-AR ($Ce\gamma_A^\theta(d)$)
N_A^δ	COST^δ-AR ($\mathrm{COST}_A^\delta(d)$)	$C\gamma^\delta$-AR ($C\gamma_A^\delta(d)$)	Ce^δ-AR ($Ce_A^\delta(d)$)	$Ce\gamma^\delta$-AR ($Ce\gamma_A^\delta(d)$)

定义7 针对$S=(U,AT=C\cup\{d\})$与度量$\mathrm{Mea}_A^{\heartsuit}(d)$，$a\in C-A$相对于$A\subseteq C$的属性外重要度为

$$\mathrm{SIG}_{\mathrm{out}}\mathrm{Mea}^{\heartsuit}(a,A,d)=\mathrm{Mea}_A^{\heartsuit}(d)-\mathrm{Mea}_{A\cup\{a\}}^{\heartsuit}(d) \tag{15}$$

$a\in A$相对于$A\subseteq C$的属性内重要度为

$$\mathrm{SIG}_{\mathrm{in}}\mathrm{Mea}^{\heartsuit}(a,A,d)=\mathrm{Mea}_A^{\heartsuit}(d)-\mathrm{Mea}_{A-\{a\}}^{\heartsuit}(d) \tag{16}$$

由定义7，针对现有条件A，可计算属性a在增加或者删除时对分类性能影响的重要度。若$\mathrm{SIG}_{\mathrm{out}}\mathrm{Mea}^{\heartsuit}(a,A,d)$大于0且越大，则$a$关于$A$对于$d$的属性重要度越大，可进行添加。若$\mathrm{SIG}_{\mathrm{in}}\mathrm{Mea}^{\heartsuit}(a,A,d)$小于0且越小，则$a$关于$A$对于$d$的属性重要度越大，不需要删除。反之$\mathrm{SIG}_{\mathrm{in}}\mathrm{Mea}^{\heartsuit}(a,A,d)$大于等于0，则$a$不重要，需要删除。由此，可以构造属性约简的启示式算法，如算法2所示。

算法2 基于$\mathrm{Mea}_A^{\heartsuit}(d)$的启发式属性约简算法($\mathrm{Mea}^{\heartsuit}$-AR)

输入：不完备决策信息系统$S=(U,AT=C\cup\{d\})$。

输出：属性约简red。

(1)初始化red$=\varnothing$，记$\mathrm{Mea}_{\varnothing}^{\heartsuit}(d)$为一个很大的值

(2)while $C-\mathrm{red}\neq\varnothing$

(3) for $a\in C-\mathrm{red}$

(4) 根据算法1计算$\mathrm{SIG}_{\mathrm{out}}\mathrm{Mea}^{\heartsuit}(a,\mathrm{red},d)=\mathrm{Mea}_{\mathrm{red}}^{\heartsuit}(d)-\mathrm{Mea}_{\mathrm{red}\cup\{a\}}^{\heartsuit}(d)$

(5) end for

(6) 选择属性a_k，使得$\mathrm{SIG}_{\mathrm{out}}\mathrm{Mea}^{\heartsuit}(a_k,\mathrm{red},d)=max_{a\in C-\mathrm{red}}(\mathrm{SIG}_{\mathrm{out}}\mathrm{Mea}_A^{\heartsuit}(a,\mathrm{red},d))$

(7) 如果$\mathrm{SIG}_{\mathrm{out}}\mathrm{Mea}^{\heartsuit}(a_k,\mathrm{red},d)>0$，则red$\leftarrow$red$\cup\{a_k\}$，否则，break

(8)end while

(9)while red$\neq\varnothing$

(10) for $a\in$red

(11) 根据算法1计算$\mathrm{SIG}_{\mathrm{in}}\mathrm{Mea}^{\heartsuit}(a,\mathrm{red},d)=\mathrm{Mea}_{\mathrm{red}}^{\heartsuit}(d)-\mathrm{Mea}_{\mathrm{red}-\{a\}}^{\heartsuit}(d)$

(12) end for

(13) 选择属性a_h，使得$\mathrm{SIG}_{\mathrm{in}}\mathrm{Mea}^{\heartsuit}(a_h,\mathrm{red},d)=\max_{a\in\mathrm{red}}(\mathrm{SIG}_{\mathrm{in}}\mathrm{Mea}_A^{\heartsuit}(a,\mathrm{red},d))$

(14) 如果$\mathrm{SIG}_{\mathrm{in}}\mathrm{Mea}^{\heartsuit}(a_h,\mathrm{red},d)\geqslant 0$，则red$\leftarrow$red$-\{a_h\}$，否则break

(15)end while

(16)返回约简red

算法2描述了代表算法$\mathrm{Mea}^{\heartsuit}$-AR，故而统一给出8种算法(如表4或图1)。算法2分为两个部分：第一部分(步骤(1～8))由属性外重要度对约简备集red添加重要属性，第二部分(步骤(9～15))由属性内重要度对所得red进一步反向删除相对冗余属性，最后步骤(16)输出约简结果。这里步骤(2～8)的最大计算次数为$m|C|^2|U|^2$，步骤(9～15)的最大计算次数为$m|C|^2|U|^2$，故算法2的时间复杂度为$O(m|C|^2|U|^2)$。

例2 继续以例1相关数据为例，表5给出8种算法的属性约简结果。表6展示了$Ce\gamma^{\delta}$-AR算法的过程和细节。在表6中，上半部分对应算法2的步骤(1～8)，通过最大属性外重要度先后添加属性c_4、c_3，得到中间值red$=\{c_3,c_4\}$；下面部分对应算法2的步骤(9～16)，实施冗余属性的反向删除，但属性内重要度均小于0故而c_3、c_4均不能删除，最终得到属性约简red$=\{c_3,c_4\}$。观察表5可知，融合度量和邻域关系对属性约简都有一定的影响，约简结果具有一些差异性。

表5 8种算法的属性约简结果

Table 5 Attribute reduction results for eight algorithms

度量邻域关系	$\mathrm{COST}_A^{\heartsuit}(d)$	$C\gamma_A^{\heartsuit}(d)$	$Ce_A^{\heartsuit}(d)$	$Ce\gamma_A^{\heartsuit}(d)$
$N_A^{\theta}(\theta=0.12)$	$\{c_2,c_3,c_4\}$	$\{c_2,c_3,c_4\}$	$\{c_3,c_4\}$	$\{c_3,c_4\}$
$N_A^{\delta}(\delta=0.1)$	$\{c_3,c_4\}$	$\{c_2,c_3,c_4\}$	$\{c_3,c_4\}$	$\{c_3,c_4\}$

表6 $Ce\gamma^{\delta}$-AR算法过程

Table 6 $Ce\gamma^{\delta}$-AR algorithmic process

添加步骤	red	$Ce\gamma^{\delta}_{\mathrm{red}}$	$C-\mathrm{red}$	$Ce\gamma^{\delta}_{\mathrm{red}\cup\{a\}}/\mathrm{SIG}_{\mathrm{out}}\mathrm{Mea}_A^{\heartsuit}(a,\mathrm{red},d)$	属性a_k
(1)	$\varnothing$	10.000	$\{c_1,c_2,c_3,c_4\}$	(7.100,6.336,2.372,1.074)/(2.900,3.664,7.628,8.926)	c_4
(2)	$\{c_4\}$	1.074	$\{c_1,c_2,c_3\}$	(0.272,0.733,0)/(0.802,0.341,1.074)	c_3
(3)	$\{c_3,c_4\}$	0.000	$\{c_1,c_2\}$	(0.000,0.000)/(0.000,0.000)	—
删除步骤	red	$Ce\gamma^{\delta}_{\mathrm{red}}$	$\mathrm{red}-\{a\}(a\in\mathrm{red})$	$Ce\gamma^{\delta}_{\mathrm{red}-\{a\}}/\mathrm{SIG}_{\mathrm{in}}\mathrm{Mea}_A^{\heartsuit}(a,\mathrm{red},d)$	属性a_h
(1)	$\{c_3,c_4\}$	0.000	$\{c_3\},\{c_4\}$	(1.074,2.372)/(−1.074,−2.372)	—
约简 $\mathrm{red}\{c_3,c_4\}$					

4 数据实验

基于文献[15]的邻域关系与决策代价,上述研究改进构建了1种邻域关系和3种融合度量,从而系统构建$2\times4=8$种约简算法(如表4或图1)。本节主要实施数据集实验,首先基本验证不确定性度量,进而重点比较约简算法的分类效果。为此,从UCI机器学习数据库(http://archive.ics.uci.edu)选取6个数据集进行实验,如表7所示。这6个数据集均涉及连续型数据,并采用常用的最大-最小归一化进行预处理。数据集需要不完备性设置,在实验中随机选取3%条件属性值处理为缺失值。此外,实验的决策代价将按以下关系随机选取:$\lambda_{PP}=\lambda_{NN}=0,0<\lambda_{BP}<\lambda_{NP}\leqslant1,0<\lambda_{BN}<\lambda_{PN}\leqslant1$。数据实验主要采用具有i5 2.30 GHz处理器、8 GB内存、Windows 10操作系统的计算机,并采用Matlab语言实现算法编程。

表7 实验数据集

Table 7 Experimental datasets 个

序号	数据集名称	样本数量	条件属性数量	决策类数量
(a)	Wine	178	13	3
(b)	Sonar	208	60	2
(c)	Ionosphere	351	34	2
(d)	Heart failure	299	12	2
(e)	Seeds	210	7	3
(f)	Wdbc	569	31	2

4.1 融合度量的计算验证

融合度量验证主要聚焦基于$n_A^{\delta}(x)$邻域表示的4种度量,包括决策代价$\mathrm{COST}_A^{\delta}(d)$与3种融合度量$C\gamma_A^{\delta}(d)$、$Ce_A^{\delta}(d)$、$Ce\gamma_A^{\delta}(d)$。为了多样观察,选取条件属性增链$A_1=\{c_1\}\subset A_2=\{c_1,c_2\}\subset\cdots\subset A_n=C$和$\delta$半径增链$\delta_1=0.02<\delta_2=0.04<\cdots<\delta_{10}=0.2$。主要采用算法1进行计

算,4种度量的相关计算结果描绘于图2的三维曲面。在图2中,x轴对应条件属性,y轴对应δ,z轴对应不确定性度量值。

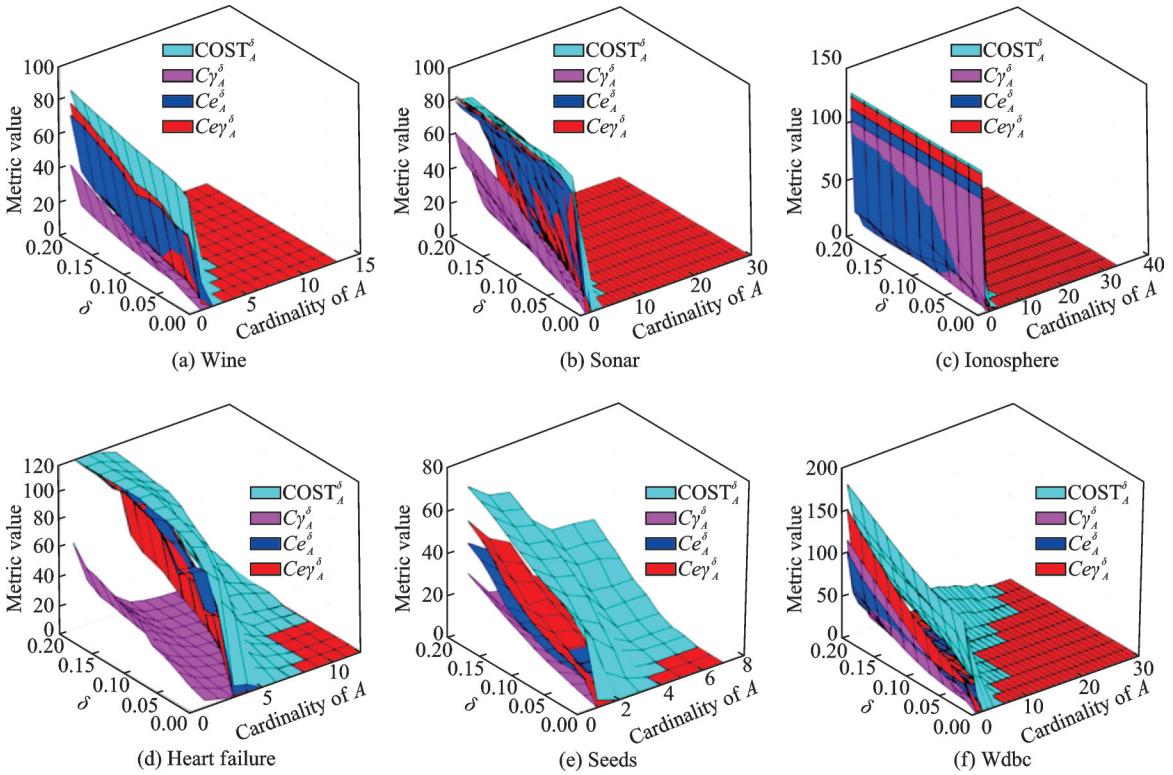

图2 基于属性增链和半径增链的4种度量曲面

Fig.2 Four measure surfaces on attribute addition chain and radius addition chain

图3的相关结果可以验证4个度量的粒化非单调性和大小关系。例如数据集(e)Seeds可以提供如下非单调反例($\delta=0.1$)

$$\begin{cases}\mathrm{COST}_{A_4}^{\delta}(d)=23.251>\mathrm{COST}_{A_5}^{\delta}(d)=11.116<\mathrm{COST}_{A_6}^{\delta}(d)=11.146\\ C\gamma_{A_4}^{\delta}(d)=5.093>C\gamma_{A_5}^{\delta}(d)=1.323<C\gamma_{A_6}^{\delta}(d)=1.327\\ Ce_{A_5}^{\delta}(d)=0.318>Ce_{A_6}^{\delta}(d)=0.310<Ce_{A_7}^{\delta}(d)=0.312\\ Ce\gamma_{A_5}^{\delta}(d)=1.603>Ce\gamma_{A_6}^{\delta}(d)=1.600<Ce\gamma_{A_7}^{\delta}(d)=1.657\end{cases}\tag{17}$$

这些结果验证了定理1。此外,图2中4个度量的大小关系与性质1的结论一致。

4.2 约简算法的分类比较

针对8种约简算法(如表4或图1),第1个COST$^\theta$-AR[15]是已有的最小决策代价属性约简算法,其余7个是新算法。下面通过约简长度和K-近邻分类精度来系统比较相关算法,其中θ和δ分别设置为0.12和0.1。为了更好地对比与改进分析,特别增加不完备信息系统的2个相关算法,对比算法1与2分别为基于邻域混合熵的属性约简算法[13]与特定类多目标代价敏感属性约简算法[14];在上述3%基础上引入8%与15%数据缺失率,最终形成低、中、高3种缺失率供深入实验与综合比较。

针对缺失率3元组(3%,8%,15%),10个算法的约简长度如表8所示,其中黑体标识最小约简长度和最佳实现次数。由表8可知,后4种算法关于数据集与缺失率的总平均约简长度分别为6.13、5.92、6.25、6.00,优于前6种算法的总平均约简长度为10.04、7.00、9.50、9.53、12.44、11.67。因此,所有10种算法都能有效约简属性,但来源于$n_A^\delta(x)$的4种算法具有约简长度优势,表明新建邻域关系的改进性。基于所建算法,$C\gamma^\delta$-AR与$Ce\gamma^\delta$-AR分别取得最优与次优的平均约简长度,接下来是算法COST$^\delta$-AR和Ce^δ-AR。

表8 10种算法的约简长度(缺失率组(3%,8%,15%))

Table 8 Reduction lengths of ten algorithms (miss-rate group (3%,8%,15%))

数据集	对比算法1	对比算法2	COST$^\theta$-AR	$C\gamma^\theta$-AR	Ce^θ-AR	Ce^θ-AR	COST$^\delta$-AR	$C\gamma^\theta$-AR	Ce^δ-AR	$Ce\gamma^\delta$-AR
Wine	(7,8,9)	(4,4,5)	(6,6,6)	(6,7,6)	(6,6,7)	(6,6,6)	(4,5,5)	(4,4,5)	(5,5,5)	(4,4,5)
Sonar	(8,8,8)	(5,5,6)	(7,7,5)	(7,6,6)	(6,6,7)	(6,6,7)	(5,5,5)	(5,5,5)	(5,5,5)	(5,5,5)
Ionasphere	(9,11,31)	(6,6,8)	(9,11,9)	(9.5,11,8)	(10,12,31)	(10,12,30)	(6,7,6)	(4,6,6)	(6,7,6)	(6,6,6)
Heart failure	(10,10,12)	(7,8,9)	(11,12,11)	(11,12,12)	(12,12,12)	(11,11,12)	(7,7.5,8)	(7.5,7.5,7)	(7,7.5,7)	(7,7,8)
Seeds	(5,5,7)	(7,6,6)	(5,5,5)	(5,5,5)	(7,7,7)	(5,6,6)	(6,6,6)	(6,6,6)	(7,7,7)	(6,6,6)
Wdbc	(8,11,15)	(8,11,15)	(18,18,18)	(18,18,18)	(25,26,25)	(24,23,23)	(7,7,8)	(7.5,8,7)	(7,7,7)	(7,8,7)
平均长度	(7.83,8.63,13.67)	(6.17,6.67,8.17)	(9.33,9.83,9.33)	(9.42,10,9.17)	(11,11.5,14.83)	(10.33,10.67,14)	(5.83,6.25,6.33)	(**5.67**,6.08,**6**)	(6.17,6.42,6.17)	(5.83,**6**,6.17)
总平均长度	10.04	7.00	9.50	9.53	12.44	11.67	6.13	**5.92**	6.25	6.00

本文将10种算法的约简结果应用于KNN分类器(K=9)。表9给出了基于十折交叉平均的分类精度,其中黑体标识最优分类精度及最优获取次数。表9实际涉及关于3种缺失率(3%,8%,15%)的3种试验精度模块,每个模块最后有关于6个数据集的统计,包括平均精度与最佳次数;进而,表9的最后1行提供了关于3种缺失率数据集的统计,由此可以提供相关的总平均精度与算法排序:[84.729 3, 86.287 2, 85.756 6, 85.231 8, 83.717 1, 86.193 2, 87.604 6, 87.569 2, 86.179, 88.138 8],$Ce\gamma^\delta$-AR ≻ COST$^\delta$-AR≻$C\gamma^\delta$-AR≻ 对比算法2≻$Ce\gamma^\theta$-AR≻Ce^δ-AR≻COST$^\theta$-AR≻$C\gamma^\theta$-AR≻ 对比算法1≻Ce^θ-AR。可见,3种新算法$Ce\gamma^\delta$-AR、COST$^\delta$-AR、$C\gamma^\delta$-AR分别取得最优、次优、第3的总平均精度,它们优于对比算法1、2与存在算法COST$^\theta$-AR[15],因此具有改进性。此外,新算法Ce^δ-AR也优于存在算法COST$^\theta$-AR及对比算法1,因此基于N_A^δ的4种新算法都具有改进优势,这也表明新建邻域关系N_A^δ的改进性。事实上,邻域关系N_A^θ[15]需要所有属性的距离函数满足同一条件,过多冗余信息导致邻域等价类较多,因此变得相对严格;基于欧式距离的邻域关系N_A^δ充分利用所有属性的值,因此获得更合理的邻域等价类和知识粒化,相关约简算法则自然具有较好的分类效果。

表9 10种算法的KNN分类精度(3种缺失率3%,8%,15%)

Table 9 KNN classification accuracies of ten algorithms (three miss rates 3%, 8%, 15%) %

(3%缺失率)数据集	对比算法1	对比算法2	COST$^{\theta}$-AR	$C\gamma^{\theta}$-AR	Ce^{θ}-AR	$Ce\gamma^{\theta}$-AR	COST$^{\delta}$-AR	$C\gamma^{\delta}$-AR	Ce^{δ}-AR	$Ce\gamma^{\delta}$-AR
Wine	97.153	**97.554 2**	96.909 9	97.153	96.042	97.187 5	94.444	96.319 5	94.444 0	96.519 5
Sonar	78.847	77.546 8	80.902 5	77.366	74.837	80.942 0	81.704	81.727 0	75.752 0	**81.756 5**
Ionasphere	83.148	84.486 8	83.657 5	83.148	83.796	83.842 5	87.222	88.287 0	83.472 0	**88.564 5**
Heart failure	70.552	75.058 4	70.0690	70.230	70.253	70.943 0	78.943	79.598 0	80.247 0	**80.297 0**
Seeds	92.857	91.568 1	**93.571 5**	92.857	91.905	**93.571 5**	92.857	92.381 0	91.429 0	92.619 0
Wdbc	96.901	96.684 5	96.3890	96.491	96.836	97.011 5	95.614	**97.137 0**	96.045 0	96.854 0
平均精度(最佳次数)	86.576 3 (0)	87.149 8 (1)	86.916 6 (1)	86.207 5 (0)	85.611 5 (0)	87.249 7 (1)	88.464 (0)	89.241 6 (1)	86.898 2 (0)	**89.435 1 (3)**
(8%缺失率)数据集	对比算法1	对比算法2	COST$^{\theta}$-AR	$C\gamma^{\theta}$-AR	Ce^{θ}-AR	$Ce\gamma^{\theta}$-AR	COST$^{\delta}$-AR	$C\gamma^{\delta}$-AR	Ce^{δ}-AR	$Ce\gamma^{\delta}$-AR
Wine	95.564	96.665	96.778	**97.513**	96.667	97.222	95.124	93.819	94.444	93.708
Sonar	76.604	77.528	80.852	77.995	76.892	78.945	80.802	81.228	76.416	**81.752**
Ionasphere	81.111	83.736	83.611	86.926	80.926	84.167	87.037	**88.333**	85.556	87.576
Heart failure	69.586	74.451	69.172	68.253	68.897	69.264	78.931	78.241	78.92	**79.287**
Seeds	89.095	90.567	**91.905**	90.952	86.19	91.182	90.952	88.571	86.19	90.571
Wdbc	94.665	95.687	94.727	93.85	93.496	**96.839**	95.786	96.303	95.608	95.959
平均精度(最佳次数)	84.438 7 (0)	86.439 2 (0)	86.174 2 (1)	85.854 8 (1)	83.844 7 (0)	86.269 8 (1)	88.105 3 (0)	87.749 2 (1)	86.789 0 (0)	**88.142 1 (2)**
(15%缺失率)数据集	对比算法1	对比算法2	COST$^{\theta}$-AR	$C\gamma^{\theta}$-AR	Ce^{θ}-AR	$Ce\gamma^{\theta}$-AR	COST$^{\delta}$-AR	$C\gamma^{\delta}$-AR	Ce^{δ}-AR	$Ce\gamma^{\delta}$-AR
Wine	93.597	94.638	**94.972**	92.153	89.931	94.708	93.22	92.639	91.042	92.778
Sonar	75.89	78.675	77.995	76.366	75.564	80.125	81.707	78.268	78.271	**83.183**
Ionasphere	80.093	83.689	81.389	82.685	81.389	81.593	85.185	85.134	81.389	**86.759**
Heart failure	69.897	74.857	70.931	70.575	68.92	71.874	77.966	77.621	**78.310**	77.529
Seeds	85.762	86.159	88.571	88.095	83.333	**88.714**	85.714	87.143	85.714	86.238
Wdbc	93.798	93.618	91.216	91.924	91.034	93.347	93.675	93.496	94.373	**94.549**
平均精度(最佳次数)	83.172 8 (0)	85.272 6 (0)	84.179 0 (1)	83.633 0 (0)	81.695 2 (0)	85.060 2 (1)	86.244 5 (0)	85.716 8 (0)	84.849 8 (1)	**86.839 3 (3)**
总平均精度	84.729 3	86.287 2	85.756 6	85.231 8	83.717 1	86.193 2	87.604 6	87.569 2	86.179	**88.138 8**

约简算法的改进主要来源于邻域关系与不确定性度量这两个维度的改进。上面已经验证了邻域粒化改进对算法的促进,下面主要分析基于度量的算法情况。综合N_A^{θ}与N_A^{δ}两种情形,基于COST分析融合度量,$C\gamma$与Ce不一定促进约简算法,但最终的$Ce\gamma$具有较大的算法改进效果。事实上,基于决

策代价COST,$Ce\gamma^{\delta}$利用依赖度和邻域熵实施了双重融合,因此它的改进效果是明显且合理的。同时,对比2种改进,邻域粒化比度量融合能更好地带来约简算法的分类效益。最终,$Ce\gamma^{\delta}$-AR充分得益于邻域关系N_A^{δ}和融合度量$Ce\gamma^{\delta}$的双重改进,成为最优算法,有效改进当前算法COST$^{\theta}$-AR[15]。

除了上述分类精度,下面补充说明算法运行时间。为此,以15%缺失率为例,图3记录了10种算法的时间,其中,图3(a,b)分别呈现小数据集与大数据集的运算时间。在数据集Ionosphere与Wdbc中,基于N_A^{θ}的4种算法的运算时间偏高但几乎在一个级别,而基于N_A^{δ}的4种算法的运算时间偏低;在其余数据集中,所有10种算法的运算时间差距并不大。因此关于3种对比算法,7种新算法从计算时间来讲是可行、有效的。

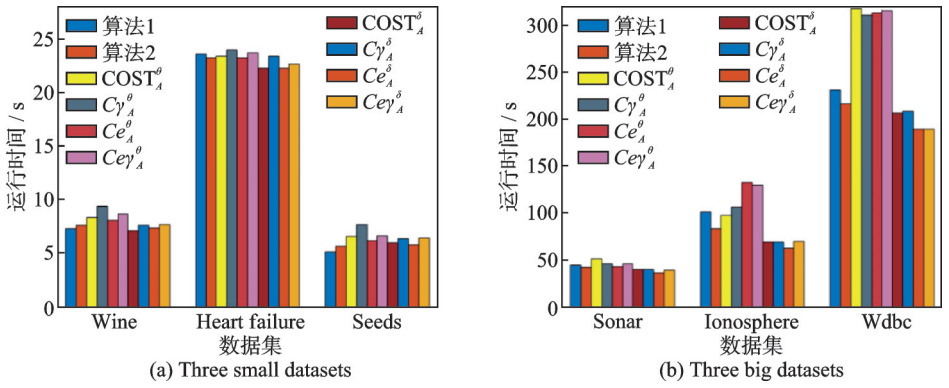

图3 10种算法运行时间比较

Fig.3 Running time comparison of ten algorithms

5 结束语

决策粗糙集及其属性约简有利于不完备连续型信息系统的不确定性处理和数据分析,但目前的相关研究还不够深入[15],具体的不足之处主要体现在:邻域关系具有较多邻域类冗余信息、决策代价度量过于单一、属性约简还具有关于分类学习的提升空间。针对这些问题,本文首先引入新的基于距离改进的邻域关系,建立不完备邻域决策粗糙集,该模型可以有效处理不完备连续数据;然后通过融合决策代价、依赖度和邻域熵得到3种新的不确定性度量;进而基于最小化决策代价,构建$2\times4=8$种启发式属性约简算法。最后数据集实验验证了不确定性度量和约简算法的有效性,提出的新算法大多(尤其是$Ce\gamma^{\delta}$-AR)具有更好的分类效果。对于实际应用提出的邻域粒化与融合度量有利于粒计算效率与测量优化,而建立的属性约简及其算法系统有利于模式识别与分类学习。相关粒化、度量、约简的深化与泛化能够诱导研究发展与应用拓展,成为可能面临的挑战。关于未来改进方向,可以增加不完备缺失值的填充,考虑三支决策等技术来改进与优化现有研究结果,并推广应用于多种信息系统(包括多粒度信息系统与集值信息系统等)。

参考文献:

[1] PAWLAK Z. Rough sets[J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356.

[2] ZIARKO W. Probabilistic approach to rough sets[J]. International Journal of Approximate Reasoning, 2007, 49(2): 272-284.

[3] YAO Y Y, ZHAO Y. Attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2008, 178(17): 3356-3373.

[4] 张静, 鞠恒荣, 杨习贝, 等. 柔性多粒化决策理论粗糙集模型[J]. 南京师大学报: 自然科学版, 2017, 40(1): 48-54.

ZHANG Jing, JU Hengrong, YANG Xibei, et al. Rough set model of flexible multi-grained decision theory[J]. Journal of Nanjing Normal University: Natural Science Edition, 2017, 40(1): 48-54.

[5] ZHAO X R, HU B Q. Three-way decisions with decision-theoretic rough sets in multiset-valued information tables[J]. Information Sciences, 2020, 507: 684-699.

[6] LI W W, HUANG Z Q, JIA X Y, et al. Neighborhood based decision-theoretic rough set models[J]. International Journal of Approximate Reasoning, 2016, 69: 1-17.

[7] GAO C, LAI Z, ZHOU J, et al. Maximum decision entropy-based attribute reduction in decision-theoretic rough set model[J]. Knowledge-Based Systems, 2018, 143(1): 179-191.

[8] 张敏, 朱启兵, 黄敏. 基于可区分度的连续空间属性约简算法研究[J]. 计算机应用研究, 2022, 39(4): 1013-1018.
ZHANG Min, ZHU Qibing, HUANG Min. Research on continuous space attribute reduction algorithm based on distinguishability[J]. Journal of Computer Application Research, 2022, 39(4): 1013-1018.

[9] SONG J, TSANG E C C, CHEN D, et al. Minimal decision cost reduct in fuzzy decision-theoretic rough set model[J]. Knowledge-Based Systems, 2017, 126(15): 104-112.

[10] CHEN B W, ZHANG X Y, YUAN Z. Two-dimensional improved attribute reductions based on distance granulation and condition entropy in incomplete interval-valued decision systems[J]. Information Sciences, 2024, 657: 119910.

[11] WU S Z, WANG L T, GE S Y, et al. Feature selection algorithm using neighborhood equivalence tolerance relation for incomplete decision systems[J]. Applied Soft Computing, 2024, 157: 111463.

[12] LIU D, LIANG D, WANG C. A novel three-way decision model based on incomplete information system[J]. Knowledge-Based Systems, 2016, 91: 32-45.

[13] 姚晟, 汪杰, 徐风, 等. 不完备邻域粗糙集的不确定性度量和属性约简[J]. 计算机应用, 2018, 38(1): 97-103.
YAO Sheng, WANG Jie, XU Feng, et al. Uncertainty measurement and attribute reduction of rough sets in incomplete neighborhood[J]. Journal of Computer Applications, 2018, 38(1): 97-103.

[14] 蔡艳婧, 程实, 王强. 不完备混合决策粗糙集特定类多目标属性约简[J]. 计算机工程与设计, 2020, 41(11): 3063-3071.
CAI Yanjing, CHENG Shi, WANG Qiang. Specific class multi-objective attribute reduction of incomplete mixed decision rough sets[J]. Computer Engineering and Design, 2020, 41(11): 3063-3071.

[15] 姚晟, 李初宴, 吴照玉. 不完备邻域决策粗糙集的最小化代价属性约简算法[J]. 计算机应用研究, 2021, 38(1): 65-68.
YAO Sheng, LI Chuyan, WU Zhaoyu. Minimizing cost attribute reduction algorithm for rough sets of incomplete neighborhood decision[J]. Journal of Computer Application Research, 2021, 38(1): 65-68.

[16] 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[J]. 软件学报, 2008, 19(3): 640-649.
HU Qinghua, YU Daren, XIE Zongxia. Numerical attribute reduction based on neighborhood granulation and rough approximation[J]. Journal of Software, 2008, 19(3): 640-649.

[17] SHU W H, XIA Q, QIAN W B. Neighborhood multigranulation rough sets for cost-sensitive feature selection on hybrid data [J]. Neurocomputing, 2024, 565: 126990.

作者简介:

张万祥(2000-),男,硕士研究生,研究方向:粗糙集与粒计算, E-mail: 1958230400@qq.com。

张贤勇(1978-),**通信作者**,男,教授,研究方向:智能计算与机器学习,E-mail: xianyongzh@sina.com。

杨霁琳(1981-),女,副教授,研究方向:数据分析与三支决策。

陈本卫(1987-),男,讲师,研究方向:粗糙集与特征选择。

(编辑:刘彦东)