文章编号: 1001-0920(2017)06-1121-07 DOI: 10.13195/j.kzyjc.2016.0665

基于双重粒化准则的邻域多粒度粗集快速约简算法

马福民[1†], 陈静雯[1], 张腾飞[2]

(1. 南京财经大学 信息工程学院，南京 210023；2. 南京邮电大学 自动化学院，南京 210023)

摘 要: 由于可以从多粒度、多层次的角度对名词型和数值型属性并存的混合数据进行有效处理,邻域多粒度粗糙集模型受到了广泛关注. 为了有效降低属性约简计算过程中的迭代次数,实现邻域多粒度粗糙集模型的快速属性约简算法,基于双重粒化准则,深入分析不同属性子集序列和邻域半径对正域的影响,结合正域在属性子集和邻域半径共同作用下的单调性,提出一种基于双重粒化准则的邻域多粒度粗集快速约简算法,并通过理论分析与实例对比验证了算法的有效性和优越性.

关键词: 粗糙集；邻域关系；双重粒化准则；属性约简；快速算法

中图分类号: TP18 **文献标志码:** A

Quick attribute reduction algorithm for neighborhood multi-granulation rough set based on double granulate criterion

MA Fu-min[1†], *CHEN Jing-wen*[1], *ZHANG Teng-fei*[2]

(1. College of Information Engineering，Nanjing University of Finance and Economics，Nanjing 210023，China; 2. College of Automation，Nanjing University of Posts and Telecommunications，Nanjing 210023，China)

Abstract: Neighborhood multi-granulation rough set model has captured more and more attention, due to its superior performance on dealing with heterogeneous data, including categorical attributes and numerical attributes, from the perspective of multi-granularity and multi-level. To effectively reduce the iterations in computing attribute reduction and realize the quick attribute reduction algorithm, the effect on positive region, caused by different attribute subsets and different neighborhood radiuses, is deeply analyzed based on the double granulate criterion. Considering the monotonicity of positive region with the joint function of attribute subset and neighborhood radius, a quick attribute reduction algorithm of neighborhood multi-granulation rough set model based on the double granulate criterion is developed. The theoretical analysis and comparable experiments, verify the effectiveness and superiority of the proposed algorithm.

Keywords: rough sets；neighborhood relation；double granulate criterion；attribute reduction；quick algorithm

0 引 言

粗糙集理论[1]作为一种处理不确定性知识的数学工具,目前已广泛应用于数据挖掘、决策分析、机器学习和知识发现等领域[2-5]. 基于粗糙集理论的属性约简是粗糙集理论的核心研究内容之一,近年来得到了快速发展,取得了大量研究成果. 然而,随着各领域数字化程度的不断提高,各个领域的数据均呈现出多维、海量的特征,给传统的属性约简算法带来了巨大的挑战. 如何进一步降低属性约简算法的复杂度,即结合具体的粗糙集模型实现快速的属性约简算法已成为当前新的研究热点.

在过去的30多年,诸多学者提出了若干个针对粗糙集及其扩展模型的属性约简算法. 由于经典粗糙集模型定义的等价关系只适合处理名词型数据,无法直接处理现实应用中广泛存在的数值型数据,Lin等[6]于1998年提出了邻域模型的概念,采用邻域关系代替等价关系,可直接处理数值型数据. Hu等[7]进一步证明了邻域粗糙集模型下正域与属性集的单调关系,设计了一种适用于处理名词型与数值型并存的混合型数据的快速约简算法.

文献[8-10]从粒计算的角度出发,详细分析了仅采用单一的不可分辨关系(单粒度、单层次)对问题进

收稿日期: 2016-05-26；**修回日期:** 2016-09-24.

基金项目: 国家自然科学基金项目(61403184, 61105082)；南京邮电大学1311人才计划基金项目(NY2013)；江苏高校优势学科建设工程项目；国家电子商务信息处理国际联合研究中心项目(2013B01035).

作者简介: 马福民(1979–), 女, 副教授, 博士, 从事智能信息处理、智能生产系统等研究；陈静雯(1993–), 女, 硕士生, 从事智能信息处理的研究.

†通讯作者. E-mail: fmmatj@126.com

行分析和求解的缺陷. 为了从多粒度、多层次的粒计算角度解决相关问题,Qian等[8]采用多个不可分辨等价关系确定论域的层次划分,提出了多粒度粗糙集模型的概念,并定义了乐观多粒度和悲观多粒度两种具体的多粒度粗糙集模型. 随后,Lin等[11]基于邻域关系将邻域粗糙集模型扩展到多粒度空间,采用属性集序列构建论域的层次划分,进一步提出了邻域多粒度粗糙集模型的概念. 徐怡等[12]指出,Lin的多粒度模型仍然是基于相同的邻域半径而建立的,只适用于处理邻域半径固定不变的问题,进而详细分析了属性集和邻域半径对分类精度的影响规律,基于不同的属性集序列和邻域半径,构建了双重粒化准则,提出了基于双重粒化准则的邻域多粒度粗糙集模型,并设计了该模型下的属性约简算法,较为充分地考虑了属性集序列和邻域半径对分类精度共同作用的效果. 然而,如何降低该模型下属性约简算法的复杂度仍然是需要进一步考虑的问题.

在基于正域的属性约简算法中,正域的计算是关键步骤,正域计算的复杂度也较大程度地影响了属性约简算法的复杂度. Liu等[13]根据数据之间的距离度量,将数据集分成一系列的Hash散列桶,采用Hash函数对数据进行散列,大幅减少了属性约简时正域的计算次数,将正域计算的时间复杂度降低为$O(m|U|)$,并给出一种复杂度为$O(m^2|U|)$的快速、高效的属性约简算法. 该算法仍然是基于传统的邻域粗糙集模型,采用的是单一固定的邻域半径,但其算法为双重粒化准则下基于不同属性集序列和不同邻域半径的快速属性约简算法的实现提供了一种新的思路.

本文以双重粒化准则的邻域多粒度粗糙集模型为理论基础,进一步深入探索该模型在属性子集和邻域半径共同作用下的正域单调性,通过减少属性约简时正域的计算次数,设计基于双重粒化准则的邻域多粒度粗糙集快速约简算法. 该算法可以灵活选择合适的属性集和邻域半径,便于结合实际的应用需要,给出不同邻域半径下的属性约简结果. 多种数据集下的实验对比分析验证了所设计算法的优越性.

1 相关知识

1.1 邻域粗糙集模型

邻域粗糙集模型采用邻域关系代替经典粗糙集模型中的等价关系,以刻画对象之间的相似性.

定义1 设$\langle U,\Delta\rangle$为一非空度量空间,其中U为对象的非空有限集,称为论域. 以$x\in U$为中心、以δ为半径的闭球,称为x的δ邻域,定义[6]为

$$n(x)=\{y\in U|\Delta(x,y)\leqslant\delta\}.$$

其中:$\delta\geqslant 0$,Δ为距离函数. 对于论域中的两点$x_i=\{x_{i1},x_{i2},\cdots,x_{in}\}$和$x_j=\{x_{j1},x_{j2},\cdots,x_{jn}\}$,通常可采用如下所示的曼哈顿距离公式:

$$\Delta(x_i,x_j)=\sum_{l=1}^{n}|x_{il}-x_{jl}|.$$

定义2 设$\langle U,\Delta\rangle$为一非空度量空间,当名词型属性和数值型属性共存时,设$B_1\subseteq C$和$B_2\subseteq C$分别是名词型属性和数值型属性,则x的邻域定义[6]为

$$\begin{aligned}&n_{B_1}(x)=\{x_i\in U|\Delta_{B_1}(x,x_i)=0\},\\&n_{B_2}(x)=\{x_i\in U|\Delta_{B_2}(x,x_i)\leqslant\delta\},\\&n_{B_1\cup B_2}(x)=\{x_i\in U|(\Delta_{B_1}(x,x_i)=0)\bigwedge\\&\qquad(\Delta_{B_2}(x,x_i)\leqslant\delta)\}.\end{aligned}$$

定义3 设邻域决策信息系统是一个五元组,即NDIS$=(U,A=C\bigcup D,V,f,N)$. 其中:C为条件属性集,D为决策属性集,$C\bigcap D=\varnothing$;$V=\bigcup\limits_{a\in A}V_a$,$V_a$为属性$a$的值域;$f:U\times A\to V$为信息函数,对于$\forall a\in A,x\in U,f(x,a)\in V_a$;$N$为由条件属性$C$生成的邻域关系. 若由$B\subseteq C$生成$U$上的邻域关系为$N_B$,则对于集合$X\subseteq U$,$X$的下、上近似集合和边界区域可以定义[7]为

$$\begin{aligned}&\underline{N_B}(X)=\{x_i\in U|n_B(x_i)\subseteq X\},\\&\overline{N_B}(X)=\{x_i\in U|n_B(x_i)\bigcap X\neq\varnothing\},\\&BN_B(X)=\overline{N_B}(X)-\underline{N_B}(X).\end{aligned}$$

下近似集$\underline{N_B}(X)$也称为X的B正区域,记为POS$_B(X)$. 由上述定义可以得到如下性质.

性质1 给定一邻域决策信息系统NDIS,如果$B_1\subseteq B_2\subseteq C$,$X\subseteq U$,则有[7]$\underline{N_{B_1}}(X)\subseteq\underline{N_{B_2}}(X)$.

进一步可以得到如下性质.

性质2 给定一邻域决策信息系统NDIS,如果$\delta_1\geqslant\delta_2$,$B\subseteq C$,$X\subseteq U$,设$X$在$\delta_1$下关于属性$B$的下近似集合为$\underline{N_B^{\delta_1}}(X)$,在$\delta_2$下关于属性$B$的下近似集合为$\underline{N_B^{\delta_2}}(X)$,则有如下关系存在:

$$\underline{N_B^{\delta_1}}(X)\subseteq\underline{N_B^{\delta_2}}(X).$$

证明 由于$\delta_1\geqslant\delta_2$,由定义1和定义2可知,对于$\forall x\in U$,有$n_B^{\delta_2}(x)\subseteq n_B^{\delta_1}(x)$. 根据定义3中下近似的定义,若$x\in\underline{N_B^{\delta_1}}(X)$,则$n_B^{\delta_1}(x)\subseteq X$,从而有$n_B^{\delta_2}(x)\subseteq n_B^{\delta_1}(x)\subseteq X$. 即对于$\forall x\in\underline{N_B^{\delta_1}}(X)$,$x\in\underline{N_B^{\delta_2}}(X)$必然成立,由此可得$\underline{N_B^{\delta_1}}(X)\subseteq\underline{N_B^{\delta_2}}(X)$. □

定义4 给定一邻域决策信息系统NDIS,δ是邻域半径,x_0是论域U的一个特殊样本,满足$f(x_0,a)=\min\{\forall_{x_i\in U,a\in C}f(x_i,a)\}$,$U$中所有的样本可以被划分进一系列的桶$B_{t_0},\cdots,B_{t_k}$[13]内,有

$$B_{t_k}=\{x_i|x_i\in U\bigwedge\lceil\Delta(x_0,x_i)/\delta\rceil=k\}.$$

这些桶的划分规则可以看作是一种hash函数.

性质3　给定一邻域决策信息系统NDIS和已划分的一系列桶$B_{t_0},\cdots,B_{t_k}$,对于$\forall x_i \in B_{t_q}(q=1,2,\cdots,k-1)$,$x_i$的邻域只可能存在于$B_{t_{q-1}}$、$B_{t_q}$或$B_{t_{q+1}}$中;如果$x_i \in B_{t_0}$,则$x_i$的邻域只可能存在于$B_{t_0}$或$B_{t_1}$中; 如果$x_i \in B_{t_k}$,则$x_i$邻域只可能存在于$B_{t_{k-1}}$或$B_{t_k}$中[13].

1.2　基于双重粒化准则的邻域粗糙集模型

多粒度粗糙集模型采用多个属性子集序列而非单个属性集对论域进行划分,从而构造多粒度的论域空间. 文献[11]采用邻域关系将多粒度粗糙集模型扩展为邻域多粒度粗糙集模型. 文献[12]考虑了不同邻域半径的选取对分类精度的影响, 进一步提出了基于不同的属性子集序列和不同邻域半径双重粒化准则的邻域多粒度粗糙集模型.

定义5　给定一邻域决策信息系统NDIS,$B=\{B_1,B_2,\cdots,B_m\}$是C的m个属性子集,$\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$是$n$个邻域半径. 定义$X\subseteq U$关于$B$和$\delta$基于双重粒化准则的乐观邻域多粒度粗糙集下近似、上近似和边界域分别[12]为

$$\underline{\sum_{i}^{j}B_i^{\delta_j}}^{O}(X)=\{x\in U|([x]_{B_1}^{\delta_1}\subseteq X)\bigvee\cdots\bigvee([x]_{B_i}^{\delta_j}\subseteq X)\bigvee\cdots\bigvee([x]_{B_m}^{\delta_n}\subseteq X)\},$$

$$\overline{\sum_{i}^{j}B_i^{\delta_j}}^{O}(X)=\sim\underline{\sum_{i}^{j}B_i^{\delta_j}}^{O}(\sim X),$$

$$\mathrm{BND}^{o}_{\sum_i^j B_i^{\delta_j}}=\overline{\sum_{i}^{j}B_i^{\delta_j}}^{O}(X)-\underline{\sum_{i}^{j}B_i^{\delta_j}}^{O}(X).$$

其中:$[x]_{B_i}^{\delta_j}$为x在属性子集B_i和邻域δ_j下的邻域粒,$\sim X$为X的补集.

定义6　给定一邻域决策信息系统NDIS,$B=\{B_1,B_2,\cdots,B_m\}$是C的m个属性子集,$\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$是$n$个邻域半径. 定义$X\subseteq U$关于$B$和$\delta$基于双重粒化准则的悲观邻域多粒度粗糙集下近似、上近似和边界域分别[12]为

$$\underline{\sum_{i}^{j}B_i^{\delta_j}}^{P}(X)=\{x\in U|([x]_{B_1}^{\delta_1}\subseteq X)\bigwedge\cdots\bigwedge([x]_{B_i}^{\delta_j}\subseteq X)\bigwedge\cdots\bigwedge([x]_{B_m}^{\delta_n}\subseteq X)\},$$

$$\overline{\sum_{i}^{j}B_i^{\delta_j}}^{P}(X)=\sim\underline{\sum_{i}^{j}B_i^{\delta_j}}^{P}(\sim X),$$

$$\mathrm{BND}^{P}_{\sum_i^j B_i^{\delta_j}}=\overline{\sum_{i}^{j}B_i^{\delta_j}}^{P}(X)-\underline{\sum_{i}^{j}B_i^{\delta_j}}^{P}(X).$$

在基于双重粒化准则的邻域多粒度粗糙集模型中,若邻域半径δ_j保持不变,则根据性质1可以得到如下推论.

推论1　给定一邻域决策信息系统NDIS,B_1和B_2是C的两个属性子集,且$B_1\subseteq B_2$,δ是邻域半径,则对于$\forall X\subseteq U$,$\underline{N_{B_1}^{\delta}}(X)\subseteq\underline{N_{B_2}^{\delta}}(X)$成立.

同样,若属性子集保持不变,则根据性质2可以得到如下推论.

推论2　给定一邻域决策信息系统NDIS,B是C的一个属性子集,δ_1和δ_2是两个邻域半径,且$\delta_1\geqslant\delta_2$,则对于$\forall X\subseteq U$,$\underline{N_B^{\delta_1}}(X)\subseteq\underline{N_B^{\delta_2}}(X)$成立.

由上述两个推论可以给出如下定理.

定理1　给定一邻域决策信息系统NDIS,δ_1和δ_2是两个邻域半径,且$\delta_1\geqslant\delta_2$,$B_1$和$B_2$是$C$的两个属性子集,且$B_1\subseteq B_2$,则对于$\forall X\subseteq U$,$\underline{N_{B_1}^{\delta_1}}(X)\subseteq\underline{N_{B_2}^{\delta_2}}(X)$成立.

证明　由推论1可知$\underline{N_{B_1}^{\delta_1}}(X)\subseteq\underline{N_{B_2}^{\delta_1}}(X)$,由推论2可以得到$\underline{N_{B_2}^{\delta_1}}(X)\subseteq\underline{N_{B_2}^{\delta_2}}(X)$,由此可知

$$\underline{N_{B_1}^{\delta_1}}(X)\subseteq\underline{N_{B_2}^{\delta_1}}(X)\subseteq\underline{N_{B_2}^{\delta_2}}(X),$$

因此$\underline{N_{B_1}^{\delta_1}}(X)\subseteq\underline{N_{B_2}^{\delta_2}}(X)$成立. □

定义7　给定一邻域决策信息系统NDIS,$B=\{B_1,B_2,\cdots,B_m\}$是C的m个属性子集,$\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$是$n$个邻域半径,$Y=\{Y_1,Y_2,\cdots,Y_r\}$是由决策属性集合$D$在论域$U$上导出的划分,定义决策类$Y$在邻域半径$\delta_j$下对属性子集$B_i$的属性依赖度[12]为

$$\gamma_{B_i}^{\delta_i}(Y)=|\mathrm{POS}_{B_i}^{\delta_i}(Y)|/|U|=|\bigcup\underline{N_{B_i}^{\delta_i}}(Y_k)|/|U|.$$

定义8　给定一邻域决策信息系统NDIS,$B=\{B_1,B_2,\cdots,B_m\}$是C的m个属性子集,$\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$是$n$个邻域半径,对于$\forall a\in B_i$,若$\gamma_{(B_i-\{a\})}^{\delta_j}(Y)<\gamma_{B_i}^{\delta_j}(Y)$,则称$a$在$\delta_j$下相对于$B_i$是必不可少的;若$\gamma_{(B_i-\{a\})}^{\delta_j}(Y)=\gamma_{B_i}^{\delta_j}(Y)$,则表示在邻域半径$\delta_j$下从$B_i$中去掉属性$a$,系统的决策正域没有改变,称$a$在$\delta_j$下相对于$B_i$是冗余的[12].

设属性子集$R\subseteq B_i$,若同时满足: 1) 对于$\forall a\in R$,$\gamma_{(R-\{a\})}^{\delta_j}(Y)<\gamma_R^{\delta_j}(Y)$; 2) $\gamma_R^{\delta_j}(Y)=\gamma_{B_i}^{\delta_j}(Y)$. 则称$R$是在$\delta_j$下$B_i$的一个约简.

定理2　给定一邻域决策信息系统NDIS,B_1和B_2是C的两个属性子集,且$B_1\subseteq B_2$,δ_j是一个邻域半径,$Y=\{Y_1,Y_2,\cdots,Y_r\}$是由决策属性集合D在论域U上导出的划分,则有[12]$\gamma_{B_1}^{\delta_j}(Y)\leqslant\gamma_{B_2}^{\delta_j}(Y)$.

定理3　给定一邻域决策信息系统NDIS,B_i是C的一个属性子集,δ_1和δ_2是两个邻域半径,且$\delta_1\geqslant\delta_2$,$Y=\{Y_1,Y_2,\cdots,Y_r\}$是由决策属性集合$D$在论域$U$上导出的划分,则有[12]$\gamma_{B_i}^{\delta_1}(Y)\leqslant\gamma_{B_i}^{\delta_2}(Y)$.

定理4　给定一邻域决策信息系统NDIS,δ_1和

δ_2是两个邻域半径,且$\delta_1 \geqslant \delta_2$, B_1和B_2是C的两个属性子集,且$B_1 \subseteq B_2, Y = \{Y_1, Y_2, \cdots, Y_r\}$是由决策属性集合$D$在论域$U$上导出的划分,则对于$\forall X \subseteq U, \gamma_{B_1}^{\delta_1}(Y) \leqslant \gamma_{B_2}^{\delta_2}(Y)$成立.

证明 由定理2可知$\gamma_{B_1}^{\delta_1}(Y) \leqslant \gamma_{B_2}^{\delta_1}(Y)$,由定理3可以得到$\gamma_{B_2}^{\delta_1}(Y) \leqslant \gamma_{B_2}^{\delta_2}(Y)$,由此可知$\gamma_{B_1}^{\delta_1}(Y) \leqslant \gamma_{B_2}^{\delta_1}(Y) \leqslant \gamma_{B_2}^{\delta_2}(Y)$,因此$\gamma_{B_1}^{\delta_1}(Y) \leqslant \gamma_{B_2}^{\delta_2}(Y)$成立. □

2 基于双重粒化准则的快速约简算法

2.1 快速约简算法思路

文献[12]所设计的属性约简算法考虑了不同属性子集序列和不同邻域半径对结果的影响,但没有进一步考虑采用一组邻域半径后属性子集序列和邻域半径的共同作用对正域计算的影响.

由定理1和定理4可知,正域在属性集增大和邻域半径减小的共同作用下具有单调性,即若样本x为已选条件属性集B在邻域半径δ下的正域样本,则在逐个向B中添加任一新属性r或将δ减小至δ'时,x仍然是正域样本. 属性的增加或邻域半径的减小,不会改变原有下近似区域中数据对象的归属关系,有可能会改变边界区域中数据对象的归属关系. 根据这一特性,在计算决策属性D对$(B + r)$在δ下的属性依赖度或计算D对B在δ'下的属性依赖度时,不需要对所有数据对象的邻域和正域重新进行计算,只需对属性子集B在δ下的边界区域中的数据对象进行计算即可.

例1 以表1所示的决策表为例进一步加以说明,表中最后一列为决策属性.

表1 决策表

U	c_1	c_2	c_3	c_4	D
x_1	0.10	0.20	0.61	0.20	Yes
x_2	0.13	0.22	0.56	0.10	Yes
x_3	0.14	0.23	0.40	0.31	No
x_4	0.16	0.41	0.30	0.16	No

由表1可得,$Y = \{Y_1, Y_2\}$. 其中:$Y_1 = \{x_1, x_2\}$, $Y_2 = \{x_3, x_4\}$. 假设$\delta = \{\delta_1, \delta_2\} = \{0.2, 0.18\}$,距离度量采用曼哈顿距离公式进行计算.

选取$\delta_1 = 0.2$,根据文献[12]提出的约简算法,将约简属性集B初始化为$\varnothing$,然后选择一个使得决策类Y对当前属性依赖度最大的属性加入初始约简集合,待选择属性子集、对应的正域和属性依赖度计算为

$$\mathrm{POS}_{c_1}^{\delta_1}(Y) = \varnothing,\ \gamma_{c_1}^{\delta_1} = 0;$$
$$\mathrm{POS}_{c_2}^{\delta_1}(Y) = \varnothing,\ \gamma_{c_2}^{\delta_1} = 0;$$
$$\mathrm{POS}_{c_3}^{\delta_1}(Y) = \{x_1, x_4\},\ \gamma_{c_3}^{\delta_1} = 0.5;$$
$$\mathrm{POS}_{c_4}^{\delta_1}(Y) = \varnothing,\ \gamma_{c_4}^{\delta_1} = 0.$$

由此可知,决策类Y对属性c_3的依赖度最大,将c_3加入约简属性集中,此时约简集为$B = \{c_3\}$. 将剩余的属性选取依赖度最大的加入约简集合,计算过程为

$$B' = \{c_1, c_3\},\ \mathrm{POS}_{B'}^{\delta_1}(Y) = \{x_1, x_4\},\ \gamma_{B'}^{\delta_1} = 0.5;$$
$$B' = \{c_2, c_3\},\ \mathrm{POS}_{B'}^{\delta_1}(Y) = \{x_1, x_4\},\ \gamma_{B'}^{\delta_1} = 0.5;$$
$$B' = \{c_3, c_4\},\ \mathrm{POS}_{B'}^{\delta_1}(Y) = \{x_1, x_2, x_3, x_4\},\ \gamma_{B'}^{\delta_1} = 1.$$

可知,决策类Y对$B' = \{c_3, c_4\}$的属性依赖度最大,将c_4加入约简属性集中. 此时,决策类Y对当前约简属性集的依赖度为1,由定义8可知,$B = \{c_3, c_4\}$是在$\delta_1 = 0.2$下的一个约简. 同样可以得到表1在$\delta_2 = 0.18$时的属性约简为$B = \{c_3, c_4\}$.

上述属性约简的整个计算过程没有考虑属性子集和邻域半径的共同作用对正域的影响,每一次计算正域都必须考虑所有的样本,而且在计算每一个样本的邻域时都必须与其他所有的数据对象进行比较. 另外,在获得决策信息系统在$\delta_1 = 0.2$下的属性约简结果后,若将邻域半径改变为$\delta_2 = 0.18$,则需要完全重复上述的计算过程. 在上述整个约简过程中,仅仅距离度量就需要计算168次.

为了减少计算复杂度,现考虑属性子集序列和邻域半径的共同作用对正域的影响(同一情况只列举一例进行说明).

在邻域半径$\delta_1 = 0.2$的情况下获得重要属性$\{c_3\}$后,此时$\mathrm{POS}_{c_3}^{\delta_1}(Y) = \{x_1, x_4\}$,在后续计算属性集合$B' = \{c_3, c_i\}(i = 1, 2, 4)$下的正域时,只需考虑剩余的数据对象$x_2$和$x_3$是否属于正域即可. 另外,当改变为$\delta_2 = 0.18$时,由于$\delta_2 < \delta_1$,在第1步计算属性$c_3$下的正域时,仍然只需判断剩余的$x_2$和$x_3$是否属于正域;且在第2步计算属性集合$B' = \{c_3, c_i\}(i = 1, 2, 4)$下的正域时,在$\delta_1 = 0.2$的情况下已属于正域的数据对象无需再考虑. 如在$B' = \{c_3, c_4\}$时,$\mathrm{POS}_{B'}^{\delta_1}(Y) = \{x_1, x_2, x_3, x_4\}$,直接可以得到$\mathrm{POS}_{B'}^{\delta_2}(Y) = \{x_1, x_2, x_3, x_4\}$,无需重新计算. 整个约简过程只需进行96次数据对象的距离度量计算即可. 在本实例中,决策表的条件属性较少,随着条件属性的增多,这种方法的优势会更加明显.

由上述分析可知,利用正域在属性集和邻域半径共同作用下的单调性可以在很多情况下简化冗余计算、提高运算效率.

2.2 快速约简算法步骤

基于双重粒化准则的邻域多粒度粗糙集模型,充分考虑属性子集序列和邻域半径共同作用对正域计算的影响,给出一种基于双重粒化准则的邻域多

粒度粗糙集快速属性约简算法(FARA_NRSDG). 考虑到过大的邻域半径无法保障最后计算结果的分类精度,借鉴文献[12]的改进算法,对分类精度小于$\gamma_C^{\delta_n} \times (1-\beta)$的邻域半径不再进行计算($\delta_n$为最小邻域半径);在第$j$个邻域半径$\delta_j$下第$s$次迭代得到的约简属性子集序列的正域记为$\mathrm{POS}_j^s$,比较得到当前最大的正域集$\max_\mathrm{POS} = \max\{\mathrm{POS}_j^s\}$,进而得到当前的约简集$\mathrm{RED}_j^s$,将正域之外的剩余数据样本集记为边界样本集$Q$,以缩小参与后续计算的数据样本规模. 当算法遍历所有的条件属性,得到邻域半径δ_j下最终的属性约简RED_j.

算法1　快速属性约简算法FARA_NRSDG.

输入: NDIS$= (U, A, V, f, N), \delta = \{\delta_1, \delta_2, \cdots, \delta_n\}, \beta$;

输出:约简集合$\mathrm{RED}_j, 1 \leqslant i \leqslant j \leqslant n$.

Step 1:计算决策类$Y = \{Y_1, Y_2, \cdots, Y_r\}$,按照从大到小对邻域半径$\delta = \{\delta_1, \delta_2, \cdots, \delta_n\}$进行排序,得到降序的$\delta' = \{\delta'_1, \delta'_2, \cdots, \delta'_n\}$.

Step 2:计算$\gamma_c^{\delta'_i}$,将符合条件$\gamma_c^{\delta'_i} \geqslant \gamma_c^{\delta'_n} \times (1-\beta)$的邻域半径$\delta_j (1 \leqslant i \leqslant j \leqslant n)$记为有效邻域半径集$\delta'' = \{\delta'_i, \delta'_{i+1}, \cdots, \delta'_n\}$,初始化$j = i$.

Step 3:对集合δ''中的每个有效邻域半径执行以下操作:

Step 3.1:初始化$s = 0, \mathrm{RED}_j^s \leftarrow \varnothing, Q \leftarrow U$, $\max\_\mathrm{POS}_j^s \leftarrow \varnothing$, max$= \varnothing$.

Step 3.2:若$(C - \mathrm{RED}_j^s) = \varnothing$,则转至Step 4,当$(C - \mathrm{RED}_j^s) \neq \varnothing$时,执行以下操作:

Step 3.2.1:如果$j \neq i$,则转至Step 3.2.2,否则,对第1个有效邻域半径下的约简进行计算,对于每个条件属性$c_k \in (C - \mathrm{RED}_j^s)$,计算决策系统关于属性集$\mathrm{RED}_j^s \bigcup \{c_k\}$的正域$\mathrm{POS}_j^s, s = s+1$,转至Step 3.3;

Step 3.2.2:如果$\mathrm{RED}_j^s \neq \varnothing$,则转至Step 3.2.3,否则,更新边界样本集Q,使$Q = Q - \mathrm{POS}_{(j-1)}^s$,在$\mathrm{POS}_{(j-1)}^s$的基础上对每个条件属性$c_k \in (C - \mathrm{RED}_j^s)$更新决策系统关于属性集$\mathrm{RED}_j^s \bigcup \{c_k\}$的正域$\mathrm{POS}_j^s, s = s+1$,转至Step 3.3;

Step 3.2.3:比较RED_j^s与$\mathrm{RED}_{(j-1)}^s$,如果$\mathrm{RED}_j^s = \mathrm{RED}_{(j-1)}^s$,则更新边界集$Q, Q = Q - \mathrm{POS}_{(j-1)}^s \bigcup \mathrm{POS}_j^{(s-1)}$,在$\mathrm{POS}_{(j-1)}^s \bigcup \mathrm{POS}_j^{(s-1)}$的基础上对每个条件属性$c_k \in (C - \mathrm{RED}_j^s)$更新决策系统关于$\mathrm{RED}_j^s \bigcup \{c_k\}$的正域$\mathrm{POS}_j^s, s = s+1$,转至Step 3.3,否则, 直接计算关于$\mathrm{RED}_j^s \bigcup \{c_k\}$的正域$\mathrm{POS}_j^s$.

Step 3.3:比较每个条件属性c_k加入后决策系统关于属性集$\mathrm{RED}_j^s \bigcup \{c_k\}$的正域$\mathrm{POS}_j^s$,求出当前最大正域$\max_\mathrm{POS}_j^s$,并与上一次的结果进行比较,若$|\max_\mathrm{POS}_j^s| = |\max_\mathrm{POS}_j^{(s-1)}|$,则属性约简保持不变,转至Step 4;若$|\max_\mathrm{POS}_j^s| > |\max_\mathrm{POS}_j^{(s-1)}|$,则将当前的属性$c_k$加入约简集$\mathrm{RED}_j^s$, $\mathrm{RED}_j^s = \mathrm{RED}_j^s \bigcup \{c_k\}$.更新边界样本集$Q, Q = Q - \max_\mathrm{POS}_j^s$,返回Step 3.2.

Step 4:令$\mathrm{RE}D_j = \mathrm{RED}_j^s, j = j+1$.若$j \leqslant n$,则返回Step 3;若$j > n$,则算法结束.

例2　为了进一步说明上述算法的计算过程,仍以表1所示的决策表为例进行说明.

假设$\beta = 0.1, \delta = \{\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \delta_7, \delta_8, \delta_9, \delta_{10}\} = \{0.5, 0.45, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05\}$. 邻域半径集$\delta$已是降序排列,直接按照Step 2计算各邻域半径下全属性集的分类精度,选取满足$\gamma_C^{\delta_i} \geqslant \gamma_C^{\delta_{10}} \times (1-\beta)$的有效邻域半径序列,最终得到$\delta'' = \{\delta_7, \delta_8, \delta_9, \delta_{10}\} = \{0.2, 0.15, 0.1, 0.05\}$.

按照Step 3,首先计算在邻域半径取值为0.2时决策表的属性约简. 根据算法步骤,将约简属性集B初始化为$\varnothing$,选择一个使得决策类Y对其依赖度最大的属性加入初始约简集合. 正如例1的计算结果,决策类Y对属性c_3的依赖度最大,将c_3加入约简属性集中,此时约简集$B = \{c_3\}$,正域$\mathrm{POS}_{c_3}^{\delta_7}(Y) = \{x_1, x_4\}$. 将$x_1$和$x_4$从初始的$Q$中去除, 得到边界样本集$Q = \{x_2, x_3\}$. 针对剩余的属性,继续选取属性依赖度最大的$c_4$加入约简集合,此时约简集$B = \{c_3, c_4\}$, $\mathrm{POS}_B^{\delta_7}(Y) = \{x_1, x_2, x_3, x_4\}$,属性依赖度$\gamma_B^{\delta_7} = 1$,得到$\delta_7 = 0.2$时的约简集$\mathrm{RED}_{\delta_7} = \{c_3, c_4\}$.

当下一个循环邻域半径取值为$\delta_8 = 0.15$时,从Step 3.2.2开始执行. 由于$\mathrm{POS}_{c_3}^{\delta_7}(Y) = \{x_1, x_4\}$,对于$\mathrm{POS}_{c_3}^{\delta_8}(Y)$,可直接从$\mathrm{POS}_{c_3}^{\delta_7}(Y)$的基础上进行更新,只需判断边界样本集$Q$中的数据对象$x_2$和$x_3$是否属于正域即可. 更新后$\mathrm{POS}_{c_3}^{\delta_8}(Y) = \{x_1, x_2, x_3, x_4\}$,根据Step 3.3,容易得到$\mathrm{RED}_{\delta_8} = \{c_3\}$.

以同样步骤得到$\mathrm{RED}_{\delta_9} = \{c_3\}, \mathrm{RED}_{\delta_{10}} = \{c_3\}$.

2.3　算法时间复杂度以及计算量分析

假设邻域决策信息系统有$|U|$个样本、m个条件属性特征. 在上述属性约简算法中,正域的计算是其关键步骤,而在Step 3中,针对每一个条件属性$c_k \in (C - \mathrm{RED}_j^s)$,由文献[13]的分析可知,计算或更新正域$\mathrm{POS}_j^s$的时间复杂度为$O(m|U|)$. 因此,当有$s$个有效邻域半径时,算法1的时间复杂度为$O(sm^2|U|)$.

相比较于文献[12]中属性约简算法的时间复杂度$O(sm^2|U|\log|U|)$,当数据样本集U较大时,本文算法具有明显的优势. 下面进一步对算法的计算量进行分析.

在Step 2中, 针对邻域半径序列, 计算有效半径集的计算量为$(s+1)m|U|$. 如果对于一个邻域半径δ_i,最终有l_i个特征被选中,则在Step 3中,每增加一个属性,$|U|/l_i$个样本由边界样本转化为正域(概率). 因此,循环计算s个邻域半径下属性约简的计算量为

$$\sum_{i=1}^{s} m|U| + (m-1)|U| \times \frac{l_i - 1}{l_i} + \cdots +$$

$$(m-l_i) \times \frac{1}{l_i} <$$

$$\sum_{i=1}^{s} \frac{m|U|}{l_i}(l_i + l_i - 1 + \cdots + 1) = \sum_{i=1}^{s} \frac{m|U|(l_i+1)}{2}.$$

考虑到定理4,在条件$\gamma_c^{\delta' i} \geqslant \gamma_c^{\delta' n} \times (1-\beta)$的限制下,邻域半径每减小一次,$|U|\Big(\sqrt[s-1]{\frac{1}{1-\beta}} - 1\Big)$个样本转化为正域(概率)只需对边界集的$|U|\Big(2 - \sqrt[s-1]{\frac{1}{1-\beta}}\Big)$个样本进行计算. 算法1总的计算量为

$$\sum_{i=1}^{s}[m|U|(1+l_i)/2 + (s+1)m|U|]\Big(2 - \sqrt[s-1]{\frac{1}{1-\beta}}\Big).$$

文献[12]约简算法所需计算次数为

$$\sum_{i=1}^{s}(m - l_i/2)(l_i+1) \times |U|\log|U| + (s+1)|U|\log|U|.$$

因此本文FARA_NRSDG算法所节省的计算次数可进一步整理为

$$\sum_{i=1}^{s}(l_i+1)|U|\Big[(m-l_i/2)\log|U| -$$

$$m/2\Big(2 - \sqrt[s-1]{\frac{1}{1-\beta}}\Big)\Big] +$$

$$(s+1)|U|\Big(\log|U| - m\Big(2 - \sqrt[s-1]{\frac{1}{1-\beta}}\Big)\Big) >$$

$$\sum_{i=1}^{s}[(l_i+1)|U|m/2 + (s+1)|U|]\Big[\log|U| -$$

$$m\Big(2 - \sqrt[s-1]{\frac{1}{1-\beta}}\Big)\Big].$$

3 实例分析

为了验证本文FARA_NRSDG算法的有效性,选取UCI数据库中Vowel、Libras movement、Australian、Cmc、 Wdbc和Abalone 6个不同规模的数据集进行测试计算,并与文献[12]所提出的改进算法(为了方便,记为MARA_NRSDG算法)进行对比分析. 6个数据集的特征如表2所示.

表 2 数据集

	数据集	记录个数	属性个数	分类数
1	Vowel	990	13	10
2	Libras	360	90	15
3	Australian	690	14	2
4	Cmc	1 473	9	3
5	Wdbc	569	31	2
6	Abalone	4 177	7	29

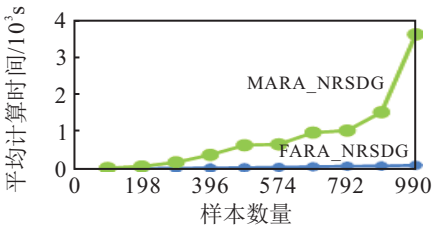
根据数据集的数据特征,初始的邻域半径序列均选用$\{0.05, 0.1, 0.15, 0.2, 0.15, 0.3\}$. 为保证最后的分类精度,定义分类精度参数$\beta = 0.1$. 在相同的邻域半径下,对比两种算法的约简结果如表3所示.

表 3 两种算法约简结果对比

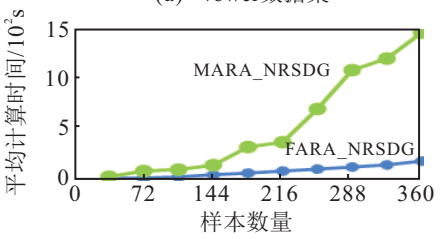
数据集	有效邻域半径	约简是否一致
Vowel	0.05, 0.1, 0.15	是
Libras	0.05, 0.1, 0.15	是
Australian	0.05, 0.1, 0.15, 0.2, 0.25, 0.3	是
Cmc	0.05, 0.1, 0.15	是
Wdbc	0.05, 0.1, 0.15, 0.2, 0.25, 0.3	是
Abalone	0.05, 0.1	是

数据集Vowel、Liras、Cmc在邻域半径大于0.15,数据集Abalone在邻域半径大于0.1时,已无法保障决策信息系统的分类精度,也无法对条件属性进行约简,再次说明了基于双重粒化准则的邻域多粒度粗糙集模型的优势: 在实际应用中仅将邻域半径设置为一个较大的固定单一数值,将无法获取某些数据集的约简结果;在双重粒化准则中,由于设置了一组可供选择的邻域半径,有更大的几率可以在多个邻域半径的对比分析中获得较为理想的约简结果.

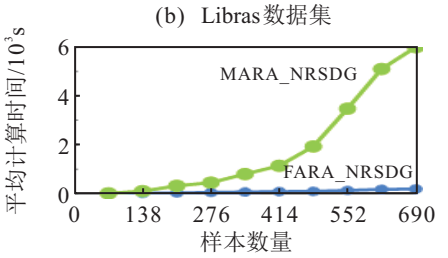
由表3可见,本文所提出的快速算法取得了与文献[12]中所设计算法一致的属性约简结果. 为了进一步测试本文算法的快速性和数据集的样本规模对两种算法计算时间的影响,将6个数据集中的样本随机排序,并大致分成10等份,按照样本规模从小到大,逐渐改变上述6个数据集的样本数量,分别采用这两种算法进行计算,并记录在不同的有效邻域半径下属性约简的平均计算时间. 为了直观地对比两种算法的计算速度,将每一个数据集不同样本数量的平均计算时间表示为图1的形式. 由图1可见,本文所提出的快速约简算法明显快于文献 [12]的算法. 如Abalone数据集,当样本数量为4 417时,算法MARA_NRSDG的平均计算时间是本文算法FARA_NRSDG平均计算时间的51倍. 对于Cmc数据集,样本数量为1 473时,MARA_NRSDG算法的平均计算时间接近本文FARA_NRSDG 算法的71 倍. 由于MARA_NRSDG算法每一步对正域的重新计算都需要考虑全部数据样本,随着数据集样本数量的增加,MARA_NRSDG算法的计算时间迅速大幅度上升,而本文的FARA_NRSDG算法只需考虑上一步正域之外的边界样本集合,在算法的快速性方面具有明显的优势.



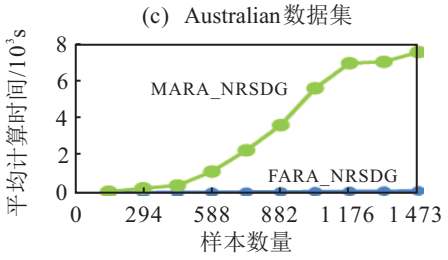
(a) Vowel数据集



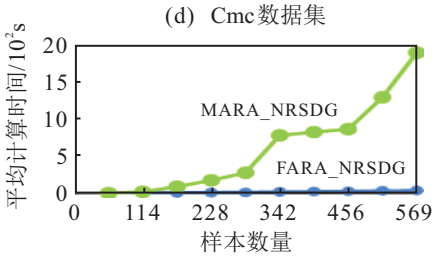
(b) Libras数据集



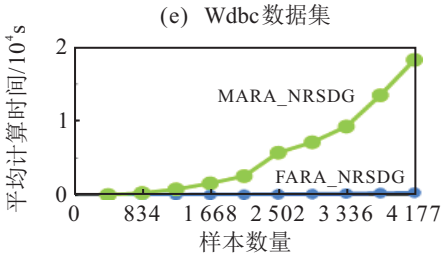
(c) Australian数据集



(d) Cmc数据集



(e) Wdbc数据集



(f) Abalone数据集

图 1　各数据集不同样本数量两种算法的平均约简时间

4　结　论

随着各应用领域数据量的不断增加,信息处理算法的快速性成了一个非常重要的指标. 本文基于双重粒化准则的邻域多粒度粗糙集模型,深入分析了不同属性子集序列和邻域半径共同作用对正域计算的影响,结合正域的单调性分析,提出了一种基于双重粒化准则的邻域多粒度粗集快速约简算法,并通过理论与实例对比分析验证了算法的优越性. 本文仅考虑了静态数据的属性约简,如何结合所提出算法的思路并基于双重粒化准则的邻域多粒度粗糙集模型,进一步探索动态数据的增量式属性约简将是下一步研究工作的重点.

参考文献(References)

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

[2] Shu B, Yang Z, Lee H, et al. Soft fuzzy rough sets and its application in decision making[J]. Artificial Intelligence Review, 2014, 41(1): 67-80.

[3] Chai J, Liu J N K. A novel believable rough set approach for supplier selection[J]. Expert Systems with Applications, 2014, 41(1): 92-104.

[4] Dai J, Xu Q. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification[J]. Applied Soft Computing, 2013, 13(1): 211-221.

[5] Peng L, Niu R, Huang B, et al. Landslide susceptibility mapping based on rough set theory and support vector machines: A case of the Three Gorges area[J]. Geomorphology, 2014, 204(1): 287-301.

[6] Lin T Y. Granular computing on binary relations I: Data mining and neighborhood systems[J]. Rough Sets in Knowledge Discovery, 1998(1): 107-121.

[7] 胡清华, 赵辉, 于达仁. 基于邻域粗糙集的符号与数值属性快速约简算法[J]. 模式识别与人工智能, 2008, 21(6): 732-738.
(Hu Q H, Zhao H, Yu D R. Efficient symbolic and numerical attribute reduction with neighborhood rough sets[J]. Pattern Recognition and Artificial Intelligence, 2008, 21(6): 732-738.)

[8] Qian Y, Liang J, Yao Y, et al. MGRS: A multi-granulation rough set[J]. Information Sciences, 2010, 180(6): 949-970.

[9] Qian Y, Liang J, Dang C. Incomplete multi-granulation rough set[J]. IEEE Trans on Systems, Man, and Cybernetics — Part A Systems and Humans, 2010, 40(2): 420-431.

[10] Qian Y, Liang J, Wei W. Pessimistic rough decision[J]. J of Zhejiang Ocean University：Natural Science, 2010, 29(5): 440-449.

[11] Lin G, Qian Y, Li J. NMGRS: Neighborhood-based multi-granulation rough sets[J]. Int J of Approximate Reasoning, 2012, 53(7): 1080-1093.

[12] 徐怡, 杨宏健, 纪霞. 基于双重粒化准则的邻域多粒度粗糙集模型[J]. 控制与决策, 2015, 30(8): 1469-1478.
(Xu Y, Yang H J, Ji X. Neighborhood multi-granulation rough set model based on double granulate criterion[J]. Control and Decision, 2015, 30(8): 1469-1478.)

[13] Liu Y, Huang W, Jiang Y, et al. Quick attribute reduct algorithm for neighborhood rough set model[J]. Information Sciences, 2014, 271(7): 65-81.

（责任编辑：郑晓蕾）