文章编号: 1001-0920 (2015) 08-1469-10 DOI: 10.13195/j.kzyjc.2014.0981

基于双重粒化准则的邻域多粒度粗糙集模型

徐 怡, 杨宏健, 纪 霞

(安徽大学 a. 计算机科学与技术学院，b. 计算智能与信号处理教育部重点实验室，合肥 230601)

摘 要: 为了从多粒度、多层次的角度有效处理名义型属性和数值型属性并存的混合数据, 首先基于不同的属性集序列和不同的邻域半径构建双重粒化准则, 建立基于双重粒化准则的邻域多粒度粗糙集模型; 然后给出该模型的相关性质, 提出该模型下的属性约简算法, 约简结果可以根据实际问题的需要灵活选择合适的属性集和邻域半径. 实例分析验证了所提出模型和算法的有效性.

关键词: 粗糙集；多粒度；邻域关系；双重粒化准则；属性约简

中图分类号: TP18 **文献标志码:** A

Neighborhood multi-granulation rough set model based on double granulate criterion

XU Yi, *YANG Hong-jian*, *JI Xia*

(a. College of Computer Science and Technology，b. Key Lab of IC & SP of Ministry of Education，Anhui University, Hefei 230601，China. Correspondent：YANG Hong-jian，E-mail：yhjian2012@163.com)

Abstract: In order to deal with the heterogeneous data including categorical attributes and numerical attributes from the multi-granularity and multi-level perspective effectively, double granulate criterion is built on different attribute sets sequence and different neighborhood radii. The neighborhood multi-granulation rough set model based on the double granulate criterion is proposed. Several relevant properties of the model are given, and an attribute reduction algorithm of the proposed model is presented. The reduction result can choose the appropriate attribute set and neighborhood radius flexibly according to the need of practical problems. The effectiveness of the proposed model and algorithm is verified by some examples.

Keywords: rough set；multi-granulation；neighborhood relation；double granulate criterion；attribute reduction

0 引 言

粗糙集理论[1]是 Pawlak 于 1982 年提出的一种能够有效处理不精确和不确定信息的数学工具. 粗糙集理论从新的视角对知识进行定义, 把知识看作是不可分辨关系对论域的分类能力, 并引入上、下近似集的概念来刻画知识的不确定程度. 该理论无需数据集之外的任何先验信息, 对问题的描述和处理比较客观, 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[2-6]. 但是, 经典粗糙集理论是基于等价关系的, 只能处理名义型属性, 无法直接处理数值型属性以及名义型属性和数值型属性并存的混合数据. 而在实际应用中, 数值型属性以及名义型属性和数值型属性并存的混合型数据广泛存在. 为了解决这个问题, Lin 等[7]利用邻域关系替代等价关系, 提出了邻域粗糙集模型, 该模型利用邻域关系刻画对象之间的相似性, 可以直接处理数值型属性; Hu 等[8-9]提出了利用邻域关系处理名义型属性和数值型属性并存的混合型数据的方法. 从粒计算[10]的角度来看, 上述的经典粗糙集模型和邻域粗糙集模型都是基于单个等价关系或单个邻域关系来对论域进行分类, 也就是说是基于单粒度和单层次的, 无法从多粒度、多层次的角度对问题进行分析和处理. Qian 等[11]采用一个属性集序列而非一个属性集来对论域进行分类, 即由多个不可分辨关系确定论域的层次划分, 构造多粒度的论域空间, 进而在多粒度论域空间上进行目标概念的近似逼近, 提出了多粒度粗糙集模型的概念, 定义了两种具体的多粒度模型: 乐观多粒度粗糙集模型和悲观多粒度粗糙集模型[11-13], 并证明了多粒

收稿日期: 2014-06-19；**修回日期:** 2014-12-31.

基金项目: 国家自然科学基金项目(61402005)；安徽省自然科学基金项目(1308085QF114)；安徽省高等学校省级自然科学基金项目(KJ2013A015, KJ2011Z020)；安徽大学计算智能与信号处理教育部重点实验室课题项目.

作者简介: 徐怡(1981−), 女, 副教授, 博士, 从事智能信息处理、粗糙集理论等研究；杨宏健(1990−), 男, 硕士, 从事粗糙集理论的研究.

度粗糙集模型是经典粗糙集模型的泛化. 在此基础上, 学者们提出了各种扩展的多粒度粗糙集模型和算法[14-16]. 为了从多粒度、多层次的角度处理名义型属性和数值型属性并存的混合数据, Lin 等[17]将邻域粗糙集模型扩展到多粒度空间, 提出了邻域多粒度粗糙集模型的概念, 以名义型和数值型属性的不同组合方式构建属性集序列, 以此作为粒化准则, 定义了两种邻域多粒度粗糙集模型. 该模型的粒化准则和经典多粒度粗糙集模型的粒化准则是一样的, 即采用属性集序列构建论域的层次划分, 只是在构建属性集序列时, 考虑了属性集中名义型和数值型属性的不同组合方式. 但是该模型是基于相同的邻域半径建立的, 因此只能处理邻域半径固定不变的问题. 而在实际应用中, 一方面, 当数据是多源的或者是分布式的时候, 数据的采集标准往往不相同, 因此用相同的邻域半径处理多源数据或分布式数据不太合适. 另一方面, 即使数据不是多源数据或分布式数据, 可以基于相同的邻域半径来处理, 也应该注意到, 在名义型属性和数值型属性并存的混合数据中, 除了属性集外, 不同邻域半径的选取对分类也有很大影响. 另外, 属性集和邻域半径对分类的影响具有如下的规律性: 属性集越大, 包含的属性个数越多, 分类粒度越细, 分类精度越高, 计算复杂性增加; 属性集越小, 包含的属性个数越少, 分类粒度越粗, 分类精度越低, 计算复杂性降低; 邻域半径越小, 分类粒度越细, 分类精度越高, 计算复杂性增加; 邻域半径越大, 分类粒度越粗, 分类精度越低, 计算复杂性降低. 因此, 从问题求解的角度来看, 有必要充分考虑属性集和邻域半径对分类共同作用的效果, 希望能够找到一个合适的属性集和邻域半径, 在该粒度层次上求解问题, 既可以达到问题求解的精度, 又可以降低问题求解的复杂性.

本文针对名义型属性和数值型属性并存的混合型数据, 基于不同的属性集序列和不同的邻域半径, 构建双重粒化准则, 建立基于双重粒化准则的邻域多粒度粗糙集模型, 并给出该模型的相关性质, 提出该模型下的属性约简算法, 最终的约简结果可以根据实际问题的需要, 灵活选择合适的属性集和邻域半径. 实例分析验证了所提出模型和算法的有效性.

1 基本概念

下面简单介绍本文将用到的相关概念.

1.1 经典粗糙集模型

粗糙集理论认为知识是一种分类能力, 利用不可分辨关系对论域的划分形成知识库, 利用知识库中的已知概念通过上、下近似集来近似逼近未知概念, 通过知识约简, 导出问题的决策或分类规则, 从而发现隐含的知识, 揭示潜在的规律[1].

定义 1 对于决策信息系统 $\mathrm{DIS}=(U, A=C\bigcup D, V, f)$. 其中: U 是对象的非空有限集合, 称为论域; A 是属性的非空有限集合, $A=C\bigcup D$, $C\bigcap D=\varnothing$, C 称为条件属性集合, D 称为决策属性集合; $V=\bigcup\limits_{a\in A}V_a$ 是属性值的集合, V_a 是属性 $a\in A$ 的值域; f 是信息函数, $f: U\times A\to V$, 即 $f(x, a)\in V_a$, 它指定了 U 中每一对象 x 的属性值[1].

定义 2 设 $\mathrm{DIS}=(U, A=C\bigcup D, V, f)$, 对于 $X\subseteq U$, $B\subseteq C$, X 的下、上近似集和边界域定义为[1]

$$\underline{B}(X)=\{x\in U: [x]_B\subseteq X\}, \tag{1}$$

$$\overline{B}(X)=\{x\in U: [x]_B\bigcap X\neq\varnothing\}, \tag{2}$$

$$\mathrm{BND}_B(X)=\overline{B}(X)-\underline{B}(X), \tag{3}$$

其中 $[x]_B$ 是 x 在 B 上的等价类.

当 $\mathrm{BND}_B(X)=\varnothing$, 即 $\overline{B}(X)=\underline{B}(X)$ 时, 称 X 在决策信息系统 $\mathrm{DIS}=(U, A=C\bigcup D, V, f)$ 中是可以定义的, 否则称 X 是粗糙的.

1.2 邻域粗糙集模型

由于经典粗糙集理论是基于等价关系的, 仅能处理名义型属性, 无法直接处理实际应用中广泛存在的数值型属性以及名义型属性和数值型属性并存的混合型数据. 为了解决这一问题, Lin 等[7]利用邻域关系替代等价关系, 提出了邻域粗糙集模型. 该模型利用邻域关系刻画对象之间的相似性, 可以直接处理数值型属性以及名义型属性和数值型属性并存的混合型数据.

定义 3 设 $\langle U, \Delta\rangle$ 为非空度量空间, 其中: $x\in U$, $\delta\geqslant 0$. 称点集

$$\delta(x)=\{y|\Delta(x, y)\leqslant\delta,\ y\in U\} \tag{4}$$

为以 x 为中心, 以 δ 为半径的闭球, 也称为 x 的 δ 邻域[7]. 其中 Δ 为距离函数, 常见的距离函数有欧氏距离函数、闵科夫斯基距离函数等, 本文将距离函数定义为如下形式.

定义 4 两点 $x_i=\{x_{i1}, x_{i2}, \cdots, x_{iN}\}$ 和 $x_j=\{x_{j1}, x_{j2}, \cdots, x_{jN}\}$ 的距离函数 Δ 定义为

$$\Delta(x_i, x_j)=\sum_{l=1}^{N}|x_{il}-x_{jl}|. \tag{5}$$

由定义 3 可知, 给定一非空度量空间 $\langle U, \Delta\rangle$, 如果有 $\delta_1\leqslant\delta_2$, 则有 $\forall x_i\in U$, $\delta_1(x_i)\subseteq\delta_2(x_i)$. 邻域粒子族 $\{\delta(x_i)|i=1, 2, \cdots, n\}$ 构成论域 U 的一个覆盖, 一般不构成一个划分.

定义 5 设 $\langle U, \Delta\rangle$ 为非空度量空间, 其中: $x\in U$, $\delta\geqslant 0$. 当名义型属性和数值型属性共存时, 设 $B_1\subseteq C$ 和 $B_2\subseteq C$ 分别是名义型属性和数值型属性, 则 x 的邻域定义为[7]

$$n_{B_1}(x)=\{x_i|\Delta_{B_1}(x,x_i)=0,\ x_i\in U\},\tag{6}$$

$$n_{B_2}(x)=\{x_i|\Delta_{B_2}(x,x_i)\leqslant\delta,\ x_i\in U\},\tag{7}$$

$$n_{B_1\bigcup B_2}(x)=$$

$$\{x_i|\Delta_{B_1}(x,x_i)=0\bigwedge\Delta_{B_2}(x,x_i)\leqslant\delta,\ x_i\in U\}.\tag{8}$$

定义 6　定义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: U 是对象的非空有限集合, 称为论域; A 是属性的非空有限集合, $A=C\bigcup D$, $C\bigcap D=\varnothing$, C 称为条件属性集合, D 称为决策属性集合; $V=\bigcup\limits_{a\in A}V_a$ 是属性值的集合, V_a 是属性 $a\in A$ 的值域; f 是信息函数, $f:U\times A\to V$, 即 $f(x,a)\in V_a$, 它指定了 U 中每一对象 x 的属性值; N 是由 C 生成的论域上的邻域关系[7].

定义 7　设邻域决策信息系统 NDIS $=(U,A=C\bigcup D,V,f,N)$, $X\subseteq U$, $B\subseteq C$ 生成 U 上的邻域关系 N_B, X 的下近似集和上近似集分别定义为[7]

$$\underline{N_B}(X)=\{x_i|n_B(x_i)\subseteq X,x_i\in U\},\tag{9}$$

$$\overline{N_B}(X)=\{x_i|n_B(x_i)\bigcap X\neq\varnothing,x_i\in U\}.\tag{10}$$

定义 8　设邻域决策信息系统 NDIS $=(U,A=C\bigcup D,V,f,N)$, $X\subseteq U$, $B\subseteq C$ 生成 U 上的邻域关系 N_B, X 的边界域定义为[7]

$$\mathrm{BN}_B(X)=\overline{N_B}(X)-\underline{N_B}(X).\tag{11}$$

当 $\mathrm{BN}_B(X)=\varnothing$, 即 $\overline{N_B}(X)=\underline{N_B}(X)$ 时, 称在邻域决策信息系统 NDIS $=(U,A=C\bigcup D,V,f,N)$ 中是可定义的, 否则称 X 是粗糙的.

1.3　多粒度粗糙集模型

从粒计算的角度来看, 经典粗糙集模型和邻域粗糙集模型都是基于单个等价关系或单个邻域关系对论域进行分类, 进而逼近未知概念, 对问题的处理是基于单粒度和单层次的, 无法从多粒度或者多层次的角度对问题进行分析和处理. Qian 等[11]提出, 采用一个属性集序列而非一个属性集来对论域进行分类, 构造多粒度的论域空间, 进而在多粒度论域空间上进行目标概念的近似逼近, 提出了多粒度粗糙集模型的概念, 定义了两种具体的多粒度模型: 乐观多粒度粗糙集模型和悲观多粒度粗糙集模型[11-13].

定义 9　设 DIS $=(U,A=C\bigcup D,V,f)$, 令 $B=\{B_1,B_2,\cdots,B_m\}$ 是 C 的 m 个属性子集, 定义 $X\subseteq U$ 关于 B 的乐观多粒度下近似集和上近似集分别记为[11]

$$\underline{\sum_{i=1}^{m}B_i}^{O}(X)=$$

$$\{x\in U:[x]_{B_1}\subseteq X\bigvee[x]_{B_2}\subseteq X\bigvee\cdots\bigvee[x]_{B_m}\subseteq X\},\tag{12}$$

$$\overline{\sum_{i=1}^{m}B_i}^{O}(X)=\sim\underline{\sum_{i=1}^{m}B_i}^{O}(\sim X).\tag{13}$$

其中: $[x]_{B_i}(1\leqslant i\leqslant \mathrm{m})$ 是 x 在属性子集 B_i 上的等价类, $\sim X$ 是 X 的补集. $\left\langle\overline{\sum_{i=1}^{m}B_i}^{O}(X),\underline{\sum_{i=1}^{m}B_i}^{O}(X)\right\rangle$ 称为关于属性集 $B_1,B_2,\cdots,B_m$ 的乐观多粒度粗糙集模型.

定义 10　设 DIS $=(U,A=C\bigcup D,V,f)$, 令 $B=\{B_1,B_2,\cdots,B_m\}$ 是 C 的 m 个属性子集, 定义 $X\subseteq U$ 关于 B 的悲观多粒度下近似集和上近似集分别记为[11]

$$\underline{\sum_{i=1}^{m}B_i}^{P}(X)=$$

$$\{x\in U:[x]_{B_1}\subseteq X\bigwedge[x]_{B_2}\subseteq X\bigwedge\cdots\bigwedge[x]_{B_m}\subseteq X\},\tag{14}$$

$$\overline{\sum_{i=1}^{m}B_i}^{P}(X)=\sim\underline{\sum_{i=1}^{m}B_i}^{P}(\sim X).\tag{15}$$

称 $\left\langle\overline{\sum_{i=1}^{m}B_i}^{P}(X),\underline{\sum_{i=1}^{m}B_i}^{P}(X)\right\rangle$ 为关于属性集 $B_1,B_2,\cdots,B_m$ 的悲观多粒度粗糙集模型.

1.4　邻域多粒度粗糙集模型

为了从多粒度、多层次的角度处理名义型属性和数值型属性并存的混合数据, 林国平将邻域粗糙集模型扩展到多粒度空间, 提出了邻域多粒度粗糙集模型的概念, 以名义型和数值型属性的不同组合方式构建属性集序列, 以此作为粒化准则, 定义了两种邻域多粒度粗糙集模型[17].

1.4.1　一类邻域多粒度粗糙集模型

定义 11　设邻域信息系统 NIS $=(U,\mathrm{AT},N)$, 其中: U 是对象的非空有限集合, AT 是包含名义型属性和数值型属性的属性集, N 表示邻域关系. $A\subseteq$ AT 为一个名义型属性集合, $B\subseteq$ AT 为一个数值型属性集合, $A\bigcup B\subseteq$ AT 为一个混合属性集合. U/A、U/B 和 $U/(A\bigcup B)$ 表示论域 U 上的一个划分和两个覆盖. 对于任意的 $X\subseteq U$, U 上由 A 和 B 导出的一类乐观多粒度下近似集和上近似集为[17]

$$\underline{(A+B)}^{O}(X)=\{x\in U|n_A(x)\subseteq X\bigvee n_B(x)\subseteq X\},\tag{16}$$

$$\overline{(A+B)}^{O}(X)=\sim\underline{(A+B)}^{O}(\sim X),\tag{17}$$

其中 $n_A(x)$ 表示对象 x 基于属性集 A 计算的邻域粒, 具体计算见定义 5.

定义 12　设邻域信息系统 NIS $=(U,\mathrm{AT},N)$, $A\subseteq$ AT 为一个名义型属性集合, $B\subseteq$ AT 为一个数值型属性集合, $A\bigcup B\subseteq$ AT 为一个混合属性集

合. U/A、U/B 和 $U/(A\bigcup B)$ 表示论域 U 上的一个划分和两个覆盖. 对于任意的 $X\subseteq U$, U 上由 A 和 B 导出的一类悲观多粒度下近似集和上近似集为[17]

$$\underline{(A+B)}^{P}(X)=\{x\in U|n_A(x)\subseteq X\bigwedge n_B(x)\subseteq X\},\tag{18}$$

$$\overline{(A+B)}^{P}(X)=\sim\underline{(A+B)}^{P}(\sim X).\tag{19}$$

1.4.2 二类邻域多粒度粗糙集模型

定义 13 设邻域信息系统 $\mathrm{NIS}=(U,\mathrm{AT},N)$, N_1 和 N_2 是论域 U 上的两个邻域关系, N_1 由 A_1 和 B_1 导出, N_2 由 A_2 和 B_2 导出. 其中: A_1 和 A_2 是两个名义型属性集合, B_1 和 B_2 是两个数值型属性集合. 对于任意的 $X\subseteq U$, 二类乐观多粒度下近似集和上近似集为[17]

$$\underline{(N_1+N_2)}^{O}(X)=$$
$$\{x\in U|n_{(A_1+B_1)}(x)\subseteq X\bigvee n_{(A_2+B_2)}(x)\subseteq X\},\tag{20}$$
$$\overline{(N_1+N_2)}^{O}(X)=\sim\underline{(N_1+N_2)}^{O}(\sim X).\tag{21}$$

定义 14 设邻域信息系统 $\mathrm{NIS}=(U,\mathrm{AT},N)$, N_1 和 N_2 是论域 U 上的两个邻域关系, N_1 由 A_1 和 B_1 导出, N_2 由 A_2 和 B_2 导出. 其中: A_1 和 A_2 是两个名义型属性集合, B_1 和 B_2 是两个数值型属性集合. 对于任意的 $X\subseteq U$, 二类悲观多粒度下近似集和上近似集为[17]

$$\underline{(N_1+N_2)}^{P}(X)=$$
$$\{x\in U|n_{(A_1+B_1)}(x)\subseteq X\bigwedge n_{(A_2+B_2)}(x)\subseteq X\},\tag{22}$$
$$\overline{(N_1+N_2)}^{P}(X)=\sim\underline{(N_1+N_2)}^{P}(\sim X).\tag{23}$$

注意, 在二类邻域多粒度粗糙集模型中, 两个邻域关系 N_1 和 N_2 只是由不同的属性集诱导出, 但是邻域半径是相同的[17].

文献 [17] 中的邻域多粒度粗糙集模型, 不论是一类还是二类, 在处理实际问题时, 邻域半径是取一个固定值. 一方面, 没有考虑到在实际应用中, 当数据是多源的或者是分布式的, 数据的采集标准往往不相同, 因此用相同的邻域半径处理多源数据或分布式数据不太合适. 另一方面, 即使数据不是多源数据或分布式数据, 可以基于相同的邻域半径来处理, 也应该注意到, 在名义型属性和数值型属性并存的混合数据中, 除了属性集外, 不同邻域半径的选取对分类也有很大影响, 有必要充分考虑属性集和邻域半径对分类共同作用的效果. 因此, 本文基于不同的属性集序列和不同的邻域半径, 构建双重粒化准则, 建立基于双重粒化准则的邻域多粒度粗糙集模型.

2 基于双重粒化准则的邻域多粒度粗糙集模型

定义 15 设 $\mathrm{NDIS}=(U,A=C\bigcup D,V,f,N)$, $B=\{B_1,B_2,\cdots,B_m\}$ 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 δ 基于双重粒化准则的乐观邻域多粒度粗糙集下近似集和上近似集分别为

$$\underline{\sum_{i}^{j}B_i\delta_j}^{O}(X)=$$
$$\{x\in U:[x]_{B_1\delta_1}\subseteq X\bigvee\cdots\bigvee$$
$$[x]_{B_i\delta_j}\subseteq X\bigvee\cdots\bigvee[x]_{B_m\delta_n}\subseteq X\},\tag{24}$$
$$\overline{\sum_{i}^{j}B_i\delta_j}^{O}(X)=\sim\underline{\sum_{i}^{j}B_i\delta_j}^{O}(\sim X),\tag{25}$$

其中 $[x]_{B_i\delta_j}$ 是对象 x 在属性集 B_i 和邻域半径 δ_j 下的邻域粒 $(1\leqslant i\leqslant m, 1\leqslant j\leqslant n)$.

定义 16 X 基于双重粒化准则的乐观邻域多粒度粗糙集的边界域定义为

$$\mathrm{BND}^{o}_{\sum_{i}^{j}B_i\delta_j}=\overline{\sum_{i}^{j}B_i\delta_j}^{O}(X)-\underline{\sum_{i}^{j}B_i\delta_j}^{O}(X).\tag{26}$$

定理 1 设 $\mathrm{NDIS}=(U,A=C\bigcup D,V,f,N)$, $B=\{B_1,B_2,\cdots,B_m\}$ 是 C 的 m 个属性子集, $\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$ 是 n 个邻域半径, 对任意的 $X\subseteq U$, 可得

$$\overline{\sum_{i}^{j}B_i\delta_j}^{O}(X)=$$
$$\{x\in U:[x]_{B_1\delta_1}\bigcap X\neq\varnothing\bigwedge\cdots\bigwedge$$
$$[x]_{B_i\delta_j}\bigcap X\neq\varnothing\bigwedge\cdots\bigwedge[x]_{B_m\delta_n}\bigcap X\neq\varnothing\}.$$

证明 由定义 15 可知

$$x\in\overline{\sum_{i}^{j}B_i\delta_j}^{O}(X)\Leftrightarrow$$
$$x\notin\underline{\sum_{i}^{j}B_i\delta_j}^{O}(\sim X)\Leftrightarrow$$
$$[x]_{B_1\delta_1}\not\subseteq(\sim X)\neq\varnothing\bigwedge\cdots\bigwedge[x]_{B_i\delta_j}\not\subseteq$$
$$(\sim X)\bigwedge\cdots\bigwedge[x]_{B_m\delta_n}\not\subseteq(\sim X)\Leftrightarrow$$
$$[x]_{B_1\delta_1}\bigcap X\neq\varnothing\bigwedge\cdots\bigwedge[x]_{B_i\delta_j}\bigcap X\neq$$
$$\varnothing\bigwedge\cdots\bigwedge[x]_{B_m\delta_n}\bigcap X\neq\varnothing.\qquad\square$$

定理 2 设 $\mathrm{NDIS}=(U,A=C\bigcup D,V,f,N)$, $B=\{B_1,B_2,\cdots,B_m\}$ 是 C 的 m 个属性子集, $\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$ 是 n 个邻域半径, 对于任意的 $X\subseteq U$, 可得

1) $\underline{\sum_{i}^{j}B_i\delta_j}^{O}(X)=\underline{B_1\delta_1}(X)\bigcup\underline{B_1\delta_2}(X)\bigcup\cdots\bigcup\underline{B_i\delta_j}(X)\bigcup\cdots\bigcup\underline{B_m\delta_n}(X)$;

2) $\overline{\sum_{i}^{j}B_i\delta_j}^{O}(X)=\overline{B_1\delta_1}(X)\bigcap\overline{B_1\delta_2}(X)\bigcap\cdots\bigcap\overline{B_i\delta_j}(X)\bigcap\cdots\bigcap\overline{B_m\delta_n}(X)$.

证明　1) 由定义 15 可知

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

故

$$\underline{\sum_{i}^{j} B_i\delta_j}^{O}(X) = \underline{B_1\delta_1}(X) \bigcup \underline{B_1\delta_2}(X) \bigcup \cdots \bigcup \underline{B_i\delta_j}(X) \bigcup \cdots \bigcup \underline{B_m\delta_n}(X);$$

2) 由定理 1 可知

$$x \in \overline{\sum_{i}^{j} B_i\delta_j}^{O}(X) \Leftrightarrow$$
$$[x]_{B_1\delta_1} \bigcap X \neq \varnothing \bigwedge \cdots \bigwedge [x]_{B_i\delta_j} \bigcap X \neq \varnothing \bigwedge \cdots \bigwedge [x]_{B_m\delta_n} \bigcap X \neq \varnothing \Leftrightarrow$$
$$x \in \overline{B_1\delta_1}(X) \bigwedge \cdots \bigwedge x \in \overline{B_i\delta_j}(X) \bigwedge \cdots \bigwedge x \in \overline{B_m\delta_n}(X),$$

故

$$\overline{\sum_{i}^{j} B_i\delta_j}^{O}(X) = \overline{B_1\delta_1}(X) \bigcap \overline{B_1\delta_2}(X) \bigcap \cdots \bigcap \overline{B_i\delta_j}(X) \bigcap \cdots \bigcap \overline{B_m\delta_n}(X). \quad \square$$

定义 17　设 NDIS $= (U, A = C \bigcup D, V, f, N)$, $B = \{B_1, B_2, \cdots, B_m\}$ 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 δ 基于双重粒化准则的悲观邻域多粒度粗糙集下近似集和上近似集分别为

$$\underline{\sum_{i}^{j} B_i\delta_j}^{P}(X) = \{x \in U : [x]_{B_1\delta_1} \subseteq X \bigwedge \cdots \bigwedge [x]_{B_i\delta_j} \subseteq X \bigwedge \cdots \bigwedge [x]_{B_m\delta_n} \subseteq X\}, \tag{27}$$

$$\overline{\sum_{i}^{j} B_i\delta_j}^{P}(X) = \sim \underline{\sum_{i}^{j} B_i\delta_j}^{P}(\sim X). \tag{28}$$

定义 18　X 基于双重粒化准则的悲观邻域多粒度粗糙集的边界域定义为

$$\mathrm{BND}^{P}_{\sum_{i}^{j} B_i\delta_j} = \overline{\sum_{i}^{j} B_i\delta_j}^{P}(X) - \underline{\sum_{i}^{j} B_i\delta_j}^{P}(X). \tag{29}$$

定理 3　设 NDIS $= (U, A = C \bigcup D, V, f, N)$, $B = \{B_1, B_2, \cdots, B_m\}$ 是 C 的 m 个属性子集, $\delta = \{\delta_1, \delta_2, \cdots, \delta_n\}$ 是 n 个邻域半径, 对于任意的 $X \subseteq U$, 可得

$$\overline{\sum_{i}^{j} B_i\delta_j}^{P}(X) = \{x \in U : [x]_{B_1\delta_1} \bigcap X \neq \varnothing \bigvee \cdots \bigvee [x]_{B_i\delta_j} \bigcap X \neq \varnothing \bigvee \cdots \bigvee [x]_{B_m\delta_n} \bigcap X \neq \varnothing\}.$$

证明　由定义 17 可知

$$x \in \overline{\sum_{i}^{j} B_i\delta_j}^{P}(X) \Leftrightarrow$$
$$x \notin \underline{\sum_{i}^{j} B_i\delta_j}^{P}(\sim X) \Leftrightarrow$$
$$[x]_{B_1\delta_1} \not\subseteq (\sim X) \neq \varnothing \bigvee \cdots \bigvee [x]_{B_i\delta_j} \not\subseteq (\sim X) \bigvee \cdots \bigvee [x]_{B_m\delta_n} \not\subseteq (\sim X) \Leftrightarrow$$
$$[x]_{B_1\delta_1} \bigcap X \neq \varnothing \bigvee \cdots \bigvee [x]_{B_i\delta_j} \bigcap X \neq \varnothing \bigvee \cdots \bigvee [x]_{B_m\delta_n} \bigcap X \neq \varnothing. \quad \square$$

定理 4　设 NDIS $= (U, A = C \bigcup D, V, f, N)$, $B = \{B_1, B_2, \cdots, B_m\}$ 是 C 的 m 个属性子集, $\delta = \{\delta_1, \delta_2, \cdots, \delta_n\}$ 是 n 个邻域半径, 对于任意的 $X \subseteq U$, 可得

1) $$\underline{\sum_{i}^{j} B_i\delta_j}^{P}(X) = \underline{B_1\delta_1}(X) \bigcap \underline{B_1\delta_2}(X) \bigcap \cdots \bigcap \underline{B_i\delta_j}(X) \bigcap \cdots \bigcap \underline{B_m\delta_n}(X);$$

2) $$\overline{\sum_{i}^{j} B_i\delta_j}^{P}(X) = \overline{B_1\delta_1}(X) \bigcup \overline{B_1\delta_2}(X) \bigcup \cdots \bigcup \overline{B_i\delta_j}(X) \bigcup \cdots \bigcup \overline{B_m\delta_n}(X).$$

证明　1) 由定义 17 可知

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

故

$$\underline{\sum_{i}^{j} B_i\delta_j}^{P}(X) = \underline{B_1\delta_1}(X)(X) \bigcap \underline{B_1\delta_2}(X) \bigcap \cdots \bigcap \underline{B_i\delta_j}(X) \bigcap \cdots \bigcap \underline{B_m\delta_n}(X);$$

2) 由定理 3 可知

$$x \in \overline{\sum_{i}^{j} B_i\delta_j}^{P}(X) \Leftrightarrow$$
$$[x]_{B_1\delta_1} \bigcap X \neq \varnothing \bigvee \cdots \bigvee [x]_{B_i\delta_j} \bigcap X \neq \varnothing \bigvee \cdots \bigvee [x]_{B_m\delta_n} \bigcap X \neq \varnothing \Leftrightarrow$$
$$x \in \overline{B_1\delta_1}(X) \bigvee \cdots \bigvee x \in \overline{B_i\delta_j}(X) \bigvee \cdots \bigvee x \in \overline{B_m\delta_n}(X),$$

故

$$\overline{\sum_{i}^{j} B_i\delta_j}^{P}(X)=\overline{B_1\delta_1}(X)(X)\bigcup\overline{B_1\delta_2}(X)\bigcup\cdots\bigcup$$

$$\overline{B_i\delta_j}(X)\bigcup\cdots\bigcup\overline{B_m\delta_n}(X). \quad \square$$

定理5 设NDIS $=(U,A=C\bigcup D,V,f,N)$, $B=\{B_1,B_2,\cdots,B_m\}$是C的m个属性子集, $\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$是$n$个邻域半径, 对于任意的$X\subseteq U$, 可得

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

$\underline{\sum_{i}^{j} B_i\delta_j}^{P}(X)\subseteq X\subseteq\overline{\sum_{i}^{j} B_i\delta_j}^{P}(X)$;

2) $\underline{\sum_{i}^{j} B_i\delta_j}^{O}(\varnothing)=\overline{\sum_{i}^{j} B_i\delta_j}^{O}(\varnothing)=$

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

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

$\underline{\sum_{i}^{j} B_i\delta_j}^{P}(U)=\overline{\sum_{i}^{j} B_i\delta_j}^{P}(U)=U$;

3) $X\subseteq Y\Rightarrow\underline{\sum_{i}^{j} B_i\delta_j}^{O}(X)\subseteq\underline{\sum_{i}^{j} B_i\delta_j}^{O}(Y)$,

$X\subseteq Y\Rightarrow\overline{\sum_{i}^{j} B_i\delta_j}^{O}(X)\subseteq\overline{\sum_{i}^{j} B_i\delta_j}^{O}(Y)$,

$X\subseteq Y\Rightarrow\underline{\sum_{i}^{j} B_i\delta_j}^{P}(X)\subseteq\underline{\sum_{i}^{j} B_i\delta_j}^{P}(Y)$,

$X\subseteq Y\Rightarrow\overline{\sum_{i}^{j} B_i\delta_j}^{P}(X)\subseteq\overline{\sum_{i}^{j} B_i\delta_j}^{P}(Y)$.

定理5的1)、2)、3)由定义15和定义17易知, 故证明在此略去.

定理6 设NDIS $=(U,A=C\bigcup D,V,f,N)$, $B=\{B_1,B_2,\cdots,B_m\}$是C的m个属性子集, $\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$是$n$个邻域半径, 若$\delta_1<\delta_2<\cdots<\delta_n$, 则对于任意的$X\subseteq U$, 可得

1) $\underline{\sum_{i}^{j} B_i\delta_j}^{O}(X)=\underline{\sum_{i}^{1} B_i\delta_1}^{O}(X)$;

2) $\overline{\sum_{i}^{j} B_i\delta_j}^{O}(X)=\overline{\sum_{i}^{1} B_i\delta_1}^{O}(X)$;

3) $\underline{\sum_{i}^{j} B_i\delta_j}^{P}(X)=\underline{\sum_{i}^{n} B_i\delta_n}^{P}(X)$;

4) $\overline{\sum_{i}^{j} B_i\delta_j}^{P}(X)=\overline{\sum_{i}^{n} B_i\delta_n}^{P}(X)$.

证明 1) 如果$\delta_1<\delta_2<\cdots<\delta_n$, $j=2,3,\cdots,n$, 则$[x]_{B_i\delta_j}\subseteq X\Rightarrow[x]_{B_i\delta_1}\subseteq X$, 所以有

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

$$[x]_{B_i\delta_j}\subseteq X\bigvee\cdots\bigvee[x]_{B_m\delta_n}X\}=$$

$$\{x\in U:[x]_{B_1\delta_1}\subseteq X\bigvee\cdots\bigvee$$

$$[x]_{B_j\delta_1}\subseteq X\bigvee\cdots\bigvee[x]_{B_m\delta_1}X\}=$$

$$\underline{\sum_{i}^{1} B_i\delta_1}^{O}(X).$$

2)、3)和4)的证明与1)类似, 在此不再赘述. $\square$

由定理6可知, 如果每个属性子集都在$\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$上构造邻域粒, 且$\delta_1<\delta_2<\cdots<\delta_n$, 则基于双重粒化准则的乐观邻域多粒度上下近似集就收缩为在最小邻域半径上的上下近似集, 基于双重粒化准则的悲观邻域多粒度上下近似集就收缩为在最大邻域半径上的上下近似集.

定理7 设NDIS $=(U,A=C\bigcup D,V,f,N)$, $B=\{B_1,B_2,\cdots,B_m\}$是C的m个属性子集, $\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$是$n$个邻域半径. 若$\delta_1=\delta_2=\cdots=\delta_n=0$, 则对于任意的$X\subseteq U$, 可得

1) $\underline{\sum_{i}^{j} B_i\delta_j}^{O}(X)=\underline{\sum_{i=1}^{m} B_i}^{O}(X)$;

2) $\overline{\sum_{i}^{j} B_i\delta_j}^{O}(X)=\overline{\sum_{i=1}^{m} B_i}^{O}(X)$;

3) $\underline{\sum_{i}^{j} B_i\delta_j}^{P}(X)=\underline{\sum_{i=1}^{m} B_i}^{P}(X)$;

4) $\overline{\sum_{i}^{j} B_i\delta_j}^{P}(X)=\overline{\sum_{i=1}^{m} B_i}^{P}(X)$.

证明 当$\delta_1=\delta_2=\cdots=\delta_n=0$时, 邻域关系就退化为等价关系, 由定义9、定义10、定义15和定义17易知1)、2)、3)和4)成立. $\square$

由定理7可知, 当邻域半径$\delta_1=\delta_2=\cdots=\delta_n=0$时, 基于双重粒化准则的邻域多粒度粗糙集模型就退化为Qian等[11]提出的经典多粒度粗糙集模型. 可见经典多粒度粗糙集模型是本文所提模型的特例.

定理8 设NDIS $=(U,A=C\bigcup D,V,f,N)$, $B=\{B_1,B_2\}$是C的2个属性子集, 且B_1为名义型

属性集, B_2 为数值型属性集, $\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$ 是 n 个邻域半径. 若 $\delta_1=\delta_2=\cdots=\delta_n$, 则对于任意的 $X\subseteq U$, 可得

1) $\sum_{i}^{j}B_i\delta_j\ \ (X)=\underline{(B_1+B_2)}^{O}(X)$;

2) $\overline{\sum_{i}^{j}B_i\delta_j\ \ (X)}=\overline{(B_1+B_2)}^{O}(X)$;

3) $\sum_{\overline{i}}^{\overline{j}}B_i\delta_j\ \ (X)=\underline{(B_1+B_2)}^{P}(X)$;

4) $\overline{\sum_{\overline{i}}^{\overline{j}}B_i\delta_j\ \ (X)}=\overline{(B_1+B_2)}^{P}(X)$.

证明　当 $\delta_1=\delta_2=\cdots=\delta_n$ 时, 多个邻域半径退化为一个邻域半径, 由定义 11、定义 12、定义 15 和定义 17 易知 1)、2)、3) 和 4) 成立. □

由定理 8 可知, 文献 [17] 提出的一类邻域多粒度粗糙集模型是本文模型的特例.

定理 9　设 NDIS $=(U,A=C\bigcup D,V,f,N)$, $B=\{B_1,B_2\}$ 是 C 的 2 个属性子集, 且 B_1 和 B_2 都为包含名义型属性和数值型属性的混合属性集, $\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$ 是 n 个邻域半径. 若 $\delta_1=\delta_2=\cdots=\delta_n$, 则对于任意的 $X\subseteq U$, 可得

1) $\sum_{i}^{j}B_i\delta_j\ \ (X)=\underline{(N_1+N_2)}^{O}(X)$;

2) $\overline{\sum_{i}^{j}B_i\delta_j\ \ (X)}=\overline{(N_1+N_2)}^{O}(X)$;

3) $\sum_{\overline{i}}^{\overline{j}}B_i\delta_j\ \ (X)=\underline{(N_1+N_2)}^{P}(X)$;

4) $\overline{\sum_{\overline{i}}^{\overline{j}}B_i\delta_j\ \ (X)}=\overline{(N_1+N_2)}^{P}(X)$.

其中 N_1 和 N_2 是分别由 B_1 和 B_2 导出的邻域关系.

证明　当 $\delta_1=\delta_2=\cdots=\delta_n$ 时, 多个邻域半径退化为一个邻域半径, 由定义 13、定义 14、定义 15 和定义 17 易知 1)、2)、3) 和 4) 成立. □

由定理 9 可知, 文献 [17] 提出的二类邻域多粒度粗糙集模型是本文模型的特例.

3　基于双重粒化准则的邻域多粒度粗糙集模型的属性约简

定义 19　设 NDIS $=(U,A=C\bigcup D,V,f,N)$, $B\subseteq C$ 是 C 的 1 个属性子集, δ 是 1 个邻域半径, $Y=\{Y_1,Y_2,\cdots,Y_r\}$ 是由决策属性集 D 在论域 U 上导出的划分, 定义决策类 Y 在邻域半径 δ 下对属性子集 B 的属性依赖度为

$$\gamma_{B\delta}(Y)=\Big|\sum_{k=1}^{r}\underline{B\delta(Y_k)}\Big|\Big/|U|. \tag{30}$$

定理 10　设 NDIS $=(U,A=C\bigcup D,V,f,N)$, $B_1\subseteq B_2$ 是 C 的 2 个属性子集, δ 是 1 个邻域半径, 则

$$\gamma_{B_1\delta}\leqslant\gamma_{B_2\delta}.$$

证明　由定义 19 和定理 5 可知, 若 $B_1\subseteq B_2$, 则 $\forall k,1\leqslant k\leqslant r,\underline{B_1\delta}(Y_k)\subseteq\underline{B_2\delta}(Y_k)$, 从而 $\sum_{k=1}^{r}\underline{B_1\delta(Y_k)}\subseteq\sum_{k=1}^{r}\underline{B_2\delta(Y_k)}$, 故有 $\gamma_{B_1\delta}\leqslant\gamma_{B_2\delta}$. □

定理 11　设 NDIS $=(U,A=C\bigcup D,V,f,N)$, B 是 C 的 1 个属性子集, δ_1 和 δ_2 是 2 个邻域半径, 且 $\delta_1\leqslant\delta_2$, 则有

$$\gamma_{B\delta_1}\geqslant\gamma_{B\delta_2}.$$

证明　$\delta_1\leqslant\delta_2\Rightarrow\forall x_i\in U:\delta_1(x_i)\subseteq\delta_2(x_i)$, 则 $\forall k,1\leqslant k\leqslant r,\underline{B\delta_1}(Y_k)\supseteq\underline{B\delta_2}(Y_k)$, 故有 $\gamma_{B\delta_1}\geqslant\gamma_{B\delta_2}$. □

定义 20　设 NDIS $=(U,A=C\bigcup D,V,f,N)$, $B\subseteq C$ 是 C 的 1 个属性子集, δ 是 1 个邻域半径. $\forall a\in B$, 若 $\gamma_{(B-\{a\})\delta}(Y)<\gamma_{B\delta}(Y)$, 则称 a 在 δ 下相对于 B 必不可少; 若 $\gamma_{(B-\{a\})\delta}(Y)=\gamma_{B\delta}(Y)$, 则表示在邻域半径 δ 下从 B 中去掉属性 a, 系统的决策正域没有改变, 称 a 在 δ 下相对于 B 是多余的.

定义 21　设 NDIS $=(U,A=C\bigcup D,V,f,N)$, $R\subseteq B\subseteq C$, δ 是 1 个邻域半径, 若

$$1)\ \forall a\in R,\gamma_{(R-\{a\})\delta}(Y)<\gamma_{R\delta}(Y), \tag{31}$$

$$2)\ \gamma_{R\delta}(Y)=\gamma_{B\delta}(Y), \tag{32}$$

则称 R 是在 δ 下 B 的一个约简.

定义 21 要求约简中在不降低系统区分能力的前提下不存在多余属性, 与经典粗糙集模型中的定义在原理上一致.

定义 22　设 NDIS $=(U,A=C\bigcup D,V,f,N)$, $B\subseteq C$ 是 C 的 1 个属性子集, δ 是 1 个邻域半径, $Y=\{Y_1,Y_2,\cdots,Y_r\}$ 是由决策属性集 D 在论域 U 上导出的划分, $\forall a\in C-B$, 定义属性 a 相对于属性集 B 的重要度为

$$\mathrm{SIG}(a,B,Y)=\gamma_{(B\bigcup\{a\})\delta}-\gamma_{B\delta}. \tag{33}$$

下面给出基于双重粒化准则的邻域多粒度粗糙集模型中的常规约简算法. 算法思想为: 若 $\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$, 针对 n 个邻域半径中的每个邻域半径 δ_i, 分别进行以属性重要度为启发因子的前向搜索, 算出每个 δ_i 上的约简集 red$[i]$. 算法描述如下.

算法 1　基于双重粒化准则的邻域多粒度粗糙集模型常规约简算法.

Input: NDIS $=(U,A=C\bigcup D,V,f,N)$, $\delta=\{\delta_1,\delta_2,\cdots,\delta_n\}$;

Output: 约简 $\mathrm{red}[i], 1 \leqslant i \leqslant n$.

Step 1: 计算 $Y = \{Y_1, Y_2, \cdots, Y_r\}$, 对 $\delta = \{\delta_1, \delta_2, \cdots, \delta_n\}$ 排序, 得到升序序列 $\delta' = \{\delta'_1, \delta'_2, \cdots, \delta'_n\}$, $1 \to i$;

Step 2: $\varnothing \to \mathrm{red}[i]$;

Step 3: 对任意 $a_j \in C - \mathrm{red}[i]$, 计算

$$\mathrm{SIG}(a_j, \mathrm{red}[i], Y) = \gamma_{(\mathrm{red}[i] \bigcup \{a_j\})\delta_i} - \gamma_{\mathrm{red}[i]\delta_i};$$

Step 4: 选择满足下式的 a_k:

$$\mathrm{SIG}(a_k, \mathrm{red}[i], Y) = \max(\mathrm{SIG}(a_j, \mathrm{red}[i], Y));$$

Step 5: If $\mathrm{SIG}(a_k, \mathrm{red}[i], Y) > 0$

$\mathrm{red}[i] \bigcup \{a_k\} \to \mathrm{red}[i]$

go to Step 3

else if $i = n$

break

else

i++

go to Step 2

算法 1 中, 在计算每个邻域半径下的约简有两个重要的步骤: 计算每个对象的邻域并分析每个邻域内的对象是否一致. 通过排序, 计算对象的邻域时间复杂度为 $O(n \log n)$, 判断邻域内的对象是否一致的时间复杂度为 $O(n)$. 所以, 当存在 N 个特征, n 个样本, k 个邻域半径时, 算法1的时间复杂度为 $O(N^2 kn \log n)$. 如果对于邻域半径 δ_i 最终有 l_i 个特征被选中, 则算法 1 实际的计算次数为

$$\sum_{i=1}^{k} N \times n \log n + (N-1)n \log n + \cdots + (N - l_i)n \log n = \sum_{i=1}^{k} (N - l_i/2)(l_i + 1) \times n \log n.$$

由距离的对称性可知, 当对象 x 的邻域包含对象 y 时, 对象 y 也必然属于对象 x 的邻域. 在计算时, 可以将 x 直接标记为 y 的邻域, 无需重新计算.

由定理 11 可知, 随着邻域半径的增加, 分类精度降低. 当邻域半径大到一定程度时, 即使基于属性全集计算分类精度, 可能依然达不到问题求解的精度需要. 所以给出基于双重粒化准则的邻域多粒度粗糙集模型的改进约简算法. 算法思想如下.

首先, 先对 $\delta = \{\delta_1, \delta_2, \cdots, \delta_n\}$ 排序, 得到新的升序邻域半径序列 $\delta' = \{\delta'_1, \delta'_2, \cdots, \delta'_n\}$. 然后, 在 $\delta' = \{\delta'_1, \delta'_2, \cdots, \delta'_n\}$ 上依次计算在条件属性全集 C 上的分类精度. 由定理 10 和定理 11 可知, 基于属性全集 C 和最小邻域半径 δ'_1 计算的分类精度 $\gamma_{C\delta'_1}$ 是最高的. 当计算到某个邻域半径 δ_Δ, 使得 $\gamma_{C\delta_\Delta} < \gamma_{C\delta'_1} \times (1 - \beta)$ (β 可根据实际需要设定), 则终止计算. 因为对于所有大于 δ_Δ 的邻域半径, 分类精度更低, 没有计算的必要. 最后, 针对 $\delta' = \{\delta'_1, \delta'_2, \cdots, \delta'_{\Delta-1}\}$ 的每个邻域半径 δ_j, 进行以属性重要度为启发因子的前向搜索, 算出每个 δ_j 上的约简集 $\mathrm{red}[j]$. 算法描述如下.

算法 2 基于双重粒化准则的邻域多粒度粗糙集模型改进的约简算法.

Input: $\mathrm{NDIS} = (U, A = C \bigcup D, V, f, N)$, $\delta = \{\delta_1, \delta_2, \cdots, \delta_n\}$, β;

Output: 约简 $\mathrm{red}[j], 1 \leqslant j \leqslant \Delta - 1$.

Step 1: 计算 $Y = \{Y_1, Y_2, \cdots, Y_r\}$, 对 $\delta = \{\delta_1, \delta_2, \cdots, \delta_n\}$ 排序, 得到升序序列 $\delta' = \{\delta'_1, \delta'_2, \cdots, \delta'_n\}$, $1 \to i$;

Step 2: 计算 $\gamma_{C\delta'_i}$;

Step 3: If $\gamma_{C\delta'_i} >= \gamma_{C\delta'_1} \times (1 - \beta)$

i++

go to Step 2

else

break// 此时算出的 δ_i 即为 δ_Δ;

Step 4: $1 \to j$;

Step 5: $\varnothing \to \mathrm{red}[j]$;

Step 6: 对任意 $a_k \in C - \mathrm{red}[j]$, 计算

$$\mathrm{SIG}(a_k, \mathrm{red}[j], Y) = \gamma_{(\mathrm{red}[j] \bigcup \{a_k\})\delta'_j} - \gamma_{\mathrm{red}[j]\delta'_j};$$

Step 7: 选择满足下式的 a_t:

$$\mathrm{SIG}(a_t, \mathrm{red}[j], Y) = \max(\mathrm{SIG}(a_k, \mathrm{red}[j], Y));$$

Step 8: If $\mathrm{SIG}(a_t, \mathrm{red}[i], Y) > 0$

$\mathrm{red}[j] \bigcup \{a_t\} \to \mathrm{red}[j]$

go to Step 6

else if $j = i$

break

else

j++

go to Step 5

当存在 N 个特征, n 个样本, k 个邻域半径, 符合条件的邻域半径有 s 个时, 算法 2 的实际计算次数为

$$\sum_{i=1}^{s} (N - l_i/2)(l_i + 1) \times n \log n + (s+1)n \log n,$$

所以算法 2 比算法 1 节省的计算次数为

$$\sum_{i=s+1}^{k} (N - l_i/2)(l_i + 1) \times n \log n - (s+1)n \log n.$$

4 实例分析

为了说明本文提出的模型和算法相对于参考文献 [17] 的优势和意义, 下面用两个例子具体说明.

例 1 利用粗糙集理论, 从多粒度和多层次角度研究使用电脑对人视力的影响, 研究对象为一群人. 假设有 3 个不同的研究小组分别对这群人展开研究,

而且这3个小组研究的侧重点有所不同. 第1个小组主要从每个人的个体特征角度来研究, 条件属性由年龄、性别、是否有眼部疾病等属性构成, 其中年龄是数值型属性, 性别和是否有眼部疾病是名义型属性. 第2个小组主要从使用电脑的个人习惯角度来研究, 条件属性由是否做眼保健操、每天使用电脑时间、座椅距离电脑屏幕的距离等属性构成, 其中每天使用电脑时间和座椅距离电脑屏幕的距离是数值型属性, 是否做眼保健操是名义型属性. 第3个小组主要从使用电脑的环境角度来研究, 条件属性由显示器尺寸、室内光线强度、显示器类型等属性构成, 其中显示器尺寸和室内光线强度是数值型属性, 显示器类型是名义型属性.

这3个研究小组的研究数据可以看作是3份不同来源的数据, 每组数据均是包含名义型属性和数值型属性的混合数据. 不难看出, 这3份数据中数值型属性的来源不同, 采集标准不同, 对视力的影响程度也不同, 所以利用文献[17]的方法, 取相同的邻域半径进行处理显然不合适. 因此需要用本文的方法, 针对不同的属性集, 采用不同的邻域半径进行处理.

例2　表1是一个关于病人疾病诊断的决策信息表, $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\}$表示10个病人; 条件属性集$C=\{c_1,c_2,c_3,c_4,c_5\}$表示5项检查指标, c_1和c_2是名义型属性, c_3、c_4、c_5是数值型属性; 决策属性$D=\{d\}$表示病人患病种类, 患第1种病用$d=1$表示, 患第2种病用$d=2$表示, 没患病用$d=0$表示.

表1　病人疾病诊断的决策信息表

U	c_1	c_2	c_3	c_4	c_5	d
x_1	1	0	0.01	0.02	0.03	1
x_2	1	0	0.12	0.14	0.12	1
x_3	1	0	0.23	0.25	0.14	1
x_4	1	0	0.28	0.36	0.45	2
x_5	1	0	0.41	0.47	0.48	2
x_6	0	1	0.53	0.58	0.52	2
x_7	0	1	0.64	0.61	0.82	0
x_8	0	1	0.75	0.72	0.86	0
x_9	0	1	0.86	0.84	0.88	0
x_{10}	0	1	0.97	0.96	0.91	2

由表1可得, $Y=U/\mathrm{IND}(D)=\{Y_1,Y_2,Y_3\}$, 其中: $Y_1=\{x_1,x_2,x_3\}, Y_2=\{x_4,x_5,x_6,x_{10}\}, Y_3=\{x_7,x_8,x_9\}$. 假设$\delta=\{\delta_1,\delta_2,\delta_3,\delta_4,\delta_5,\delta_6,\delta_7,\delta_8,\delta_9,\delta_{10}\}=\{0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1\}$, $\beta=0.1$, 距离函数采用定义4计算.

例1中给出的δ序列已经排好序, 可以直接由算法2计算各邻域半径下全属性集的分类精度, 结果为

$$\gamma_{C\delta_1'}=1,\ \gamma_{C\delta_2'}=1,$$
$$\gamma_{C\delta_3'}=0.9,\ \gamma_{C\delta_4'}=0.9,\ \gamma_{C\delta_5'}=0.8.$$

由于$\gamma_{C\delta'_5}=0.8$, 满足$\gamma_{C\delta_{s+1}}<\gamma_{C\delta_1}\times(1-\beta)$, 所以只要在$\delta=\{\delta_1',\delta_2',\delta_3',\delta_4'\}$上, 利用算法2继续进行前向搜索, 其中: $\delta_1'=0.1, \delta_2'=0.2, \delta_3'=0.3, \delta_4'=0.4$.

在$\delta_1'=0.1$上, 初始约简集为$\varnothing$, 首先选择一个属性, 对于$\varnothing$的属性重要度最大, 将其加入约简集中, 计算结果如下:

$$\mathrm{SIG}(c_1,\varnothing,Y)=\gamma_{(\{c_1\})\delta_1'}-\gamma_{\varnothing\delta_1'}=0,$$
$$\mathrm{SIG}(c_2,\varnothing,Y)=\gamma_{(\{c_2\})\delta_1'}-\gamma_{\varnothing\delta_1'}=0,$$
$$\mathrm{SIG}(c_3,\varnothing,Y)=\gamma_{(\{c_3\})\delta_1'}-\gamma_{\varnothing\delta_1'}=0.8,$$
$$\mathrm{SIG}(c_4,\varnothing,Y)=\gamma_{(\{c_4\})\delta_1'}-\gamma_{\varnothing\delta_1'}=0.8,$$
$$\mathrm{SIG}(c_5,\varnothing,Y)=\gamma_{(\{c_5\})\delta_1'}-\gamma_{\varnothing\delta_1'}=0.6.$$

由以上结果可以看出, 对当前约简集$\varnothing$的属性重要度最大的是属性c_3和属性c_4, 程序中选择前边一个, 即属性c_3, 加入临时约简集. 重复以上步骤, 选择一个相对于临时约简集$\{c_3\}$, 属性重要度最大的属性, 计算结果如下:

$$\mathrm{SIG}(c_1,\{c_3\},Y)=\gamma_{(\{c_3\}\bigcup\{c_1\})\delta_1'}-\gamma_{\{c_3\}\delta_1'}=0,$$
$$\mathrm{SIG}(c_2,\{c_3\},Y)=\gamma_{(\{c_3\}\bigcup\{c_2\})\delta_1'}-\gamma_{\{c_3\}\delta_1'}=0,$$
$$\mathrm{SIG}(c_4,\{c_3\},Y)=\gamma_{(\{c_3\}\bigcup\{c_4\})\delta_1'}-\gamma_{\{c_3\}\delta_1'}=0.2,$$
$$\mathrm{SIG}(c_5,\{c_3\},Y)=\gamma_{(\{c_3\}\bigcup\{c_5\})\delta_1'}-\gamma_{\{c_3\}\delta_1'}=0.2.$$

对于临时约简集$\{c_3\}$, 属性重要度最大的属性选择c_4. 重复以上步骤, 直到所有剩余属性相对于临时约简集的属性重要度都为0, 即

$$\mathrm{SIG}(c_1,\{c_3,c_4\},Y)=$$
$$\gamma_{(\{c_3,c_4\}\bigcup\{c_1\})\delta_1'}-\gamma_{(\{c_3,c_4\})\delta_1'}=0,$$
$$\mathrm{SIG}(c_2,\{c_3,c_4\},Y)=$$
$$\gamma_{(\{c_3,c_4\}\bigcup\{c_2\})\delta_1'}-\gamma_{(\{c_3,c_4\})\delta_1'}=0,$$
$$\mathrm{SIG}(c_5,\{c_3,c_4\},Y)=$$
$$\gamma_{(\{c_3,c_4\}\bigcup\{c_5\})\delta_1'}-\gamma_{(\{c_3,c_4\})\delta_1'}=0.$$

至此, 在$\delta_1'=0.1$上的计算结束, 得出$\delta_1'=0.1$上的约简集

$$\mathrm{red}[1]=\{c_3,c_4\},\ \gamma_{red[1]\delta_1'}=1.$$

继续计算可得

$$\mathrm{red}[2]=\{c_3,c_4,c_5\},\ \gamma_{red[2]\delta_1'}=1,$$
$$\mathrm{red}[3]=\{c_3,c_4,c_5\},\ \gamma_{red[3]\delta_3'}=0.9,$$
$$\mathrm{red}[4]=\{c_3,c_4,c_5\},\ \gamma_{red[4]\delta_4'}=0.9.$$

故U的约简为

$$R_{\delta=0.1}=\{c_3,c_4\},$$
$$R_{\delta=0.2}=\{c_3,c_4,c_5\},$$
$$R_{\delta=0.3}=\{c_3,c_4,c_5\},$$
$$R_{\delta=0.4}=\{c_3,c_4,c_5\}.$$

从以上结果可以看出, 基于双重粒化准则的邻域多粒度粗糙集模型改进的属性约简算法通过对邻域

半径的排序计算, 有效降低了计算次数, 提高了计算效率. 且本文的属性约简算法可以给出多个邻域半径下的不同属性约简结果, 在实际应用中, 便于用户根据实际问题的需要, 灵活选择合适的属性集和邻域半径.

针对例 2, 如果用文献 [17] 的方法来处理, 只能给定一个固定的邻域半径, 求得一个固定的属性约简结果. 但是, 采用本文的方法可以计算出, 当$\delta = 0.1$时, 得到的约简属性集为$R_{\delta=0.1} = \{c_3,c_4\}$; 当$\delta = 0.4$时, 得到的约简属性集为$R_{\delta=0.4} = \{c_3,c_4,c_5\}$. 假设$c_5$属性在实际应用中为某项血液化验指标, 获取$c_5$的时间成本和经济成本都较高, 为了避免获取属性$c_5$, 降低问题求解的代价, 可以取邻域半径$\delta = 0.1$, 约简结果为$R_{\delta=0.1} = \{c_3,c_4\}$, 即此时问题求解的粒度层次为: 属性集取$\{c_3,c_4\}$, 邻域半径$\delta = 0.1$. 反过来, 当取邻域半径$\delta = 0.1$时, 若数值型属性取值较密集, 此时邻域半径较小, 划分的邻域粒较多, 势必会增加计算量, 当计算量增加的代价超过获取属性c_5的代价时, 更合适的做法是取$\delta = 0.4$, 用约简结果$R_{\delta=0.4} = \{c_3,c_4,c_5\}$代替$R_{\delta=0.1} = \{c_3,c_4\}$. 可见, 本文的方法可以根据实际问题的求解需要, 灵活选择属性集和邻域半径, 从而在更好的粒度层次上求解问题.

5 结 论

本文针对名义型属性和数值型属性并存的混合型数据, 基于不同的属性集序列和不同的邻域半径, 构建双重粒化准则, 建立基于双重粒化准则的邻域多粒度粗糙集模型, 并给出该模型的相关性质, 提出了基于双重粒化准则的邻域多粒度粗糙集模型的属性约简算法. 实例分析验证了所提模型和算法的有效性. 需要注意的是, 在文献 [17] 中, “单个邻域半径”的选取是事先设定的, 在本文的方法中, “一组邻域半径参数”也是事先设定的, 可根据实际问题的需要, 由先验知识确定或由领域专家指定. 在以后的工作中, 还需进一步研究如何利用机器学习等方法, 以及自动确定一组合适的邻域半径参数的方法.

参考文献(References)

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

[2] Sun B Z, Ma W M. Soft fuzzy rough sets and its application in decision making[J]. Artificial Intelligence Review, 2014, 41(1): 67-80.

[3] Chai J Y, 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 H, 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] Pai P F, Li L L, Hung W Z, et al. Using ADABOOST and rough set theory for predicting debris flow disaster[J]. Water Resources Management, 2014, 28(4): 1143-1155.

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

[7] Lin T Y, Liu Q, Huang K J, et al. Rough sets, neighborhood systems and approximation[C]. Proc of the 5th Int Symposium on Methodologies of Intelligent Systems. Knoxville: 1990: 130-141.

[8] Hu Q H, Yu D R, Liu J F, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences, 2008, 178(18): 3577-3594.

[9] Hu Q H, Yu D R, Xie Z X. Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34(2): 866-876.

[10] Lin T Y. Granular computing[C]. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. Berlin Heidelberg: Springer, 2003: 16-24.

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

[12] Qian Y H, Liang J Y, Dang C Y. Incomplete multigranulation rough set[J]. IEEE Trans on Systems, Man and Cybernetics, 2010, 40(2): 420-431.

[13] Qian Y H, Zhang H, Sang Y L, et al. Multigranulation decision-theoretic rough sets[J]. Int J of Approximate Reasoning, 2014, 55(1): 225-237.

[14] 张明, 唐振民, 徐维艳, 等. 可变粒度粗糙集模型[J]. 模式识别与人工智能, 2012, 25(4): 709-720.
(Zhang M, Tang Z M, Xu W Y, et al. Variable multigranulation rough set model[J]. Pattern Recognition and Artificial Intelligence, 2012, 25(4): 709-720.)

[15] 桑妍丽, 钱宇华. 一种悲观多粒度粗糙集中的粒度约简算法[J]. 模式识别与人工智能, 2012, 25(3): 361-366.
(Sang Y L, Qian Y H. A granular space reduction approach to pessimistic multi-granulation rough sets[J]. Pattern Recognition and Artificial Intelligence, 2012, 25(3): 361-366.)

[16] 马睿, 刘文奇. 基于集值信息系统的多粒度粗糙集[J]. 系统工程与电子技术, 2014, 36(5): 920-925.
(Ma R, Liu W Q. Multi-granulation rough set model based on set-valued information system[J]. Systems Engineering and Electronics, 2014, 36(5): 920-925.)

[17] Lin G P, Qian Y H, Li J J. NMGRS: Neighborhood-based multigranulation rough sets[J]. Int J of Approximate Reasoning, 2012, 53(1): 1080-1093.

(责任编辑: 齐 霁)