第51卷第5期 南京航空航天大学学报 Vol. 51 No. 5
2019年10月 Journal of Nanjing University of Aeronautics & Astronautics Oct. 2019

DOI:10.16356/j.1005-2615.2019.05.008

基于区分矩阵的多粒度属性约简

翁 冉[1] 王俊红[1,2] 魏 巍[1,2] 崔军彪[1] 黄卫华[3]

(1.山西大学计算机与信息技术学院,太原,030006;2.山西大学计算智能与中文信息处理教育部重点实验室,太原,030006;3.文山学院数学与工程学院,文山,663099)

摘要: 多粒度是粒计算领域的重要研究方向之一,它在两个或多个不同的粒度下进行问题求解,已经成为解决复杂问题的一种新的范式。属性约简作为粗糙集理论的核心内容之一,已被成功地应用于粒计算、数据挖掘等领域。将多粒度思想应用于属性约简将是一个有意义的研究方向。为此,本文运用粒计算理论中的粒化思想进行属性粒化,构造多个属性粒;然后基于属性粒上的区分矩阵计算属性粒的重要度和属性粒中属性重要度;最后利用这两种重要度设计了一种多粒度属性约简算法。通过在不同的粒中挑选属性,该算法得到的约简结果更具有代表性和差异性。本文利用6个数据集对提出的多粒度属性约简算法的性能进行测试,实验结果表明了提出算法的有效性。

关键词: 粗糙集;属性约简;多粒度;区分矩阵

中图分类号: TP18 **文献标志码:** A **文章编号:** 1005-2615(2019)05-0636-07

Multi-granulation Attribute Reduction Based on Discernibility Matrix

WENG Ran[1], *WANG Junhong*[1,2], *WEI Wei*[1,2], *CUI Junbiao*[1], *HUANG Weihua*[3]

(1. School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, China; 2. Key Laboratory of Computation Intelligence and Chinese Information Processing, Ministry of Education, Taiyuan, 030006, China; 3. School of Mathematics and Engineering, Wenshan University, Wenshan, 663099, China)

Abstract: Multi-granularity is one of the important research directions in the field of granular computing. It represents the research of problem solving at two or more different granules and already has become a new computing method to solve complex problems. Rough set theory is a kind of computing tool to solve uncertain problems effectively. As one of the core contents of rough set theory, attribute reduction has been widely studied in the fields of data mining, machine learning, and granular computing. It is a meaningful problem to apply the idea of multi-granularity to attribute reduction. The granulation idea in granular computing theory is used to granulate attributes to form multiple granules. Then we evaluate the significance of attribute granules and the significance of attributes in attribute granules based on discernibility matrix. Finally, a multi-granularity attribute reduction algorithm based on these two significance measures is designed. By selecting attributes from different granules, the reduction results obtained by this algorithm are more representative and different. In order to verify the effectiveness of our proposed method, experiments on six data sets show that the effectiveness of the proposed algorithm.

Key words: rough set; attribute reduction; multi-granulation; discernibility matrix

基金项目: 国家自然科学基金(61772323, 61303008)资助项目;山西省自然科学基金(201701D121051)资助项目;云南省教育厅课题(2018JS490)资助项目。

收稿日期: 2019-05-30;**修订日期:** 2019-08-30

通信作者: 魏巍,男,教授,E-mail:weiwei@sxu.edu.cn。

引用格式: 翁冉,王俊红,魏巍,等.基于区分矩阵的多粒度属性约简[J].南京航空航天大学学报,2019,51(5):636-642.
WENG Ran, WANG Junhong, WEI Wei. Multi-granulation Attribute Reduction Based on Discernibility Matrix[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2019, 51(5): 636-642.

粗糙集理论[1]由 Pawlak 于 1982 年提出,它是一种有效处理信息系统中不精确、不确定信息的分析方法。在粗糙集理论中,属性约简[2-6]是一个重要的概念,它通过删除冗余和不相关属性来保持和原始知识库一样的分类能力。粒计算概念则是由 Zadeh[7]于 1979 年在模糊集背景下首先提出,该文献给出了为什么要研究粒度计算和如何进行计算的基本框架,之后粒计算受到了研究者[8-13]的广泛关注。粒计算从多个角度,多个不同粒度层次出发,对不确定、不精确或复杂的问题进行求解,得到了广泛的认可。

粗糙集作为粒计算的代表性模型之一,已成为粒计算研究领域的一个重要方向。近些年来,通过将多粒度思想融入到粗糙集理论研究,已经产生了许多优秀的研究成果。Qian 等[14]提出了一种新的基于多个二元关系的粗糙集模型,称为多粒度粗糙集,其中目标近似集是由多个等价关系重构得到。在此基础上多粒度粗糙集模型进行了扩展:如乐悲观多粒度粗糙集[15-16]、多粒度决策粗糙集[17]、不完备多粒度粗糙集[18-19]以及多粒度模糊粗糙集[20]等。吴志远等[21]提出了程度多粒度,该文章在描述下近似时,对象在粒度下的等价类划分不需要严格的满足是目标概念的子集;张明等[22]的加权多粒度中粒度带有权重,最终挑选的粒度满足对应的粒度权重之和大于某个阈值即可;Xu 等[23]定义了广义的多粒度下近似,避免了原始方法在进行对象选取时要求太严格或太放松。上述文献进行多粒度属性约简时大多是在多粒度粗糙集模型下进行研究,将单个属性看作一个粒度,在不同的模型的下近似分布保持不变的条件下以启发式约简进行计算。本文从属性粒化的角度出发,通过构造属性粒构建多粒度空间,并基于区分矩阵定义了属性粒和属性粒内部属性的重要性度量,进而构造了相应的多粒度属性约简算法。由于构建的属性粒内部的属性具有相似性,不同属性粒中的属性具有差异性,从每个属性粒中先选取一个属性作为代表整个粒,这样得到的属性约简中属性将会具有较好的代表性和多样性。

1　基本概念

本节将简单地回顾粗糙集的基本概念和定义。

定义 1[1]　一个四元组 $S=(U,AT,V,f)$ 表示一个信息系统,其中:U 为对象的非空有限集合,即论域;AT 为属性的非空有限集合;$V=\bigcup_{a\in AT}V_a$,V_a 是 a 的值域;$f:U\times AT\to V$ 表示一个信息函数,为每个对象的每个属性赋予一个值,$\forall a\in AT$,$x\in U$,$f(x,a)\in V_a$. $S=(U,AT,V,f)$ 可以简记为 $S=(U,AT)$。

若 $AT=C\cup D$,C 是条件属性集合,D 是决策属性的集合,则 S 称为决策信息系统。

定义 2[1]　给定信息系统 $S=(U,AT)$,$\forall A\subseteq AT$,那么 A 的不可区分关系能够被定义为

$$\mathrm{IND}(A)=\{(x,y)\in U^2|\forall a\in A,f(x,a)=f(y,a)\} \tag{1}$$

对于 $\forall x\in U$,x 的等价类被定义为

$$[x]_A=\{y\in U|(x,y)\in \mathrm{IND}(A)\} \tag{2}$$

定义 3[1]　设信息系统 $S=(U,AT)$,$A\subseteq AT$,对于 $X\subseteq U$,X 的上下近似集合和边界域定义为

$$\underline{A}(X)=\{x\in U|[x]_A\subseteq X\} \tag{3}$$

$$\overline{A}(X)=\{x\in U|[x]_A\cap X\neq\varnothing\} \tag{4}$$

$$\mathrm{BND}_A(X)=\overline{A}(X)-\underline{A}(X) \tag{5}$$

定义 4[1]　设决策表 $S=(U,C\cup D)$,则 D 的 C 正域(记为 $\mathrm{pos}_C(D)$)定义为

$$\mathrm{pos}_C(D)=\bigcup_{X\in U/D}\underline{C}(X) \tag{6}$$

定义 5[1]　设 $B\subseteq C$,B 为 C 的 Q 约简当且仅当 B 是 C 的子集,且 $\mathrm{pos}_B(Q)=\mathrm{pos}_C(Q)$,$C$ 的 Q 约简简称为相对约简。

2　属性粒及属性粒中属性的重要度

本节将提出基于包含度的属性粒化方法,并在此基础上给出属性粒及属性粒中属性的重要度。

2.1　基于包含度的属性粒化

粒化是粒计算理论中的一个基本问题,文献[13]指出将论域中的具有相似关系、邻近关系、功能关系等聚成一个类,这些类称之为粒。本文采用的方法是将一组相似的属性构成一个属性粒。

定义 6[1]　决策 D 关于属性 C 的依赖度

$$r_C(D)=\frac{|\mathrm{pos}_C(D)|}{|U|} \tag{7}$$

由依赖性的定义可知:若设 $C=\{a_1,a_2,\cdots,a_m\}$,$\forall a_i,a_j\in C$,那么 $r_{a_i}(d)$,$r_{a_j}(d)$ 分别表示决策属性对条件属性 a_i,a_j 的的依赖程度。当 r_{a_i} 与 r_{a_j} 值很接近,表示决策属性对条件属性 a_i,a_j 的依赖程度相似,可以说明属性 a_i 与 a_j 的区分能力相似。因此,可以将这两个属性看作一个粒来进行考虑,基于以上分析可以给出了属性粒的定义。

定义 7　设决策信息系统 $S=(U,AT)$ 中,$C=\{a_1,a_2,\cdots,a_m\}$,$G=\{G_1,G_2,\cdots,G_k\}$ 为属性 C 构成的 k 个粒,$\forall a_i,a_j\in C$,若 $r_{a_i}-r_{a_j}<\mu$,则将属性 a_i 和

a_j 划为一个属性粒 G_m，其中：$G_m \subseteq C$，$G_m \neq \varnothing$，$\forall m=1,2,\cdots,k$；$G_p \cap G_q = \varnothing, \forall p,q=1,2,\cdots,k, p \neq q$。

本文首先计算决策对每个属性的依赖度，然后将所有条件属性按照依赖度进行排序，如果相邻的两个属性依赖度的差值小于某个阈值，那么就将这两个依赖度对应的属性划为同一个属性粒，通过该方法可以将相似的属性构建成属性粒。

2.2 属性粒的重要度

属性粒的重要度用来刻画属性粒整体的重要性.每个属性粒是由多个属性组成，不同属性粒可以区分不同对象，有些属性粒的区分能力较强，在全属性下能够区分的对象，在这些属性集下也能够区分，那么该属性集就比较重要，区分能力较强，由该属性集组成的属性粒也就比较重要。

区分矩阵能够表示出区别一对对象的所有属性集合，而正域意义下的区分矩阵是将协调部分（正域中包含的对象）和非协调部分（非正域包含的对象）考虑在内分别对待。因此可以通过构建区分矩阵来定义属性粒的重要度。

定义 8[1,3] 给定信息表 $S=(U,AT)$，C 是条件属性，则属性集下的区分矩阵为 $M_C^R=[m_C^R]_{ij}$ 其中

$$[m_C^R]_{ij}=\{c\in C: f(x_i,c)\neq f(x_j,c)\} \tag{8}$$

式中：M_C^R 表示通过全体属性集 C 建立的区分矩阵，$x_i,x_j\in U$。

定义 9[1,3] 给定决策表 $S=(U,AT)$，其中 $AT=C\cup D$，则正域意义下的区分矩阵为 $\boldsymbol{M}_C^p=[\boldsymbol{m}_C^p]_{ij}$，其中

$$[m_C^p]_{ij}=\begin{cases}\{c\in C: f(x_i,c)\neq f(x_j,c)\} & f(x_i,d)\neq f(x_j,d)\\ & x_i,x_j\in U_1\\ \{c\in C: f(x_i,c)\neq f(x_j,c)\}, & x_i\in U_1, x_j\in U_2\\ \varnothing & \text{其他}\end{cases} \tag{9}$$

式中：U_1 是决策表 S 的协调部分；U_2 是决策表 S 的不协调部分。M_C^P 表示通过全体属性集 C 建立的正域意义下的区分矩阵。根据上面2个区分矩阵可以定义属性粒的重要度如下：

定义 10 给定一个决策表 $S=(U,AT)$，$G=\{G_1,G_2,\cdots,G_k\}$ 为 k 个属性粒，$\forall G_l\in G$，则属性粒 G_l 的重要度定义为

$$\mathrm{sig}_{G_l}=\sum_{i=1}^{|U|}\sum_{j=1}^{|U|}\delta([\boldsymbol{m}_{G_l}^R]_{ij},[\boldsymbol{m}_C^p]_{ij}) \tag{10}$$

其中

$$\delta([\boldsymbol{m}_{G_l}^R]_{ij},[\boldsymbol{m}_C^p]_{ij})=\begin{cases}1 & [\boldsymbol{m}_{G_l}^R]_{ij}\neq\varnothing,[\boldsymbol{m}_C^p]_{ij}\neq\varnothing\\ -1 & [\boldsymbol{m}_{G_l}^R]_{ij}\neq\varnothing,[\boldsymbol{m}_C^p]_{ij}=\varnothing\\ 0 & [\boldsymbol{m}_{G_l}^R]_{ij}=\varnothing\end{cases} \tag{11}$$

式中：$[\boldsymbol{m}_{G_l}^R]_{ij}$ 表示通过属性粒 G_l 包含的属性建立的区分矩阵；$[\boldsymbol{m}_C^p]_{ij}$ 表示全体属性对应的正域意义下的区分矩阵。

$[\boldsymbol{m}_{G_l}^R]_{ij}$ 与 $[\boldsymbol{m}_C^p]_{ij}$ 的关系可以分为3种情况：

（1）若全属性正域下能够区分某对对象，基于属性粒构建的区分矩阵也能够区分该对对象，则为该属性粒的重要度加1，表示该属性粒对该对对象的区分起到了积极的作用；

（2）若全属性正域下不需要区分某对对象，基于属性粒构建的区分矩阵却区分了该对对象，则为该粒的重要度减1，表示该属性粒对该对对象的区分起到了消极作用；

（3）若基于属性粒构建的区分矩阵没有区分某对对象，那么表示属性粒在区分这对对象没有做影响，设为0。

从上面分析可以看出，若某个属性粒的重要度越大，则表明该属性粒与全体属性区分的对象越接近，因此该属性粒越重要。

2.3 属性粒中属性重要度

一个属性在其所在属性粒对应的区分矩阵中出现的频率越高，该属性区分不同类别对象的能力就越强，因此可以利用属性粒对应区分矩阵中属性出现的频率评价属性的重要度。

定义 11 给定一个信息系统 $S=(U,AT)$，$G=\{G_1,G_2,\cdots,G_k\}$ 为 k 个粒，$\forall G_l\in G$，则 G_l 中的属性 a 重要度的定义为

$$\mathrm{sig}_{G_l}(a)=|\{a|a\in[\boldsymbol{m}_{G_l}^R]_{ij}\}| \tag{12}$$

3 基于区分矩阵的多粒度属性约简

属性约简是保持知识库分类能力不变的条件下，删除其中不相关或不重要的知识。

由于属性粒是由一组相似的属性构成，因此可以选取属性粒中的一个属性代表整个属性粒.我们通过定义属性粒的重要度，按从大到小的顺序对属性粒进行排序，排序靠前的属性粒较重要，优先被考虑。

然后根据排序顺序依次在属性粒中挑选一个

属性直到满足正域条件结束,挑选属性根据属性粒中属性的重要度进行选取.该方法通过借助属性粒的作用将属性挑出来,这样的属性约简集合由于来自不同属性粒,具有较好的代表性和多样性。

例:给定一个决策信息系统 $S=(U,C\cup D)$(如表 1 所示),其中 $U=\{x_1,x_2,\cdots,x_8\}$, $C=\{a_1,a_2,\cdots,a_{10}\}$, $U/d=\{D_1,D_2,D_3\}$, 其中 $D_1=\{x_1,x_3,x_5\}$,$D_2=\{x_2,x_4,x_6\}$,$D_3=\{x_7,x_8\}$。

表 1　决策信息系统

Tab. 1　Decision information system

U	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	d
x_1	1	1	1	1	0	0	0	1	1	0	0
x_2	2	2	0	0	1	2	2	0	0	1	1
x_3	1	1	1	1	0	0	0	1	1	0	0
x_4	2	2	0	0	1	3	2	0	0	1	1
x_5	0	0	1	1	1	3	2	1	3	1	0
x_6	2	2	0	2	1	1	3	0	0	2	1
x_7	2	2	0	2	1	1	3	0	0	2	2
x_8	3	2	2	0	1	2	3	2	0	1	2

下面进行约简的计算:

(1) 首先计算决策对每个属性的依赖度

$r_{a_1}(d)=0.500\,0$, $r_{a_2}(d)=0.375\,0$,

$r_{a_3}(d)=0.500\,0$, $r_{a_4}(d)=0.375\,0$,

$r_{a_5}(d)=0.250\,0$, $r_{a_6}(d)=0.250\,0$,

$r_{a_7}(d)=0.250\,0$, $r_{a_8}(d)=0.500\,0$,

$r_{a_9}(d)=0.375\,0$, $r_{a_{10}}(d)=0.250\,0$

(2) 对依赖度进行排序,阈值选取 0.1,生成 3 个属性粒:$G_1=\{a_1,a_3,a_8\}$, $G_2=\{a_2,a_4,a_9\}$, $G_3=\{a_5,a_6,a_7,a_{10}\}$。

(3)全属性下基于正域建立区分矩阵为(由于区分矩阵是对称矩阵,因此只给出了矩阵的一半元素)

$$\boldsymbol{M}_C^P=\begin{bmatrix}
\varnothing & & & & & & & & & \\
C & \varnothing & & & & & & & & \\
\varnothing & C & \varnothing & & & & & & & \\
C & \varnothing & C & \varnothing & & & & & & \\
\varnothing & \{a_1,a_2,a_3,a_4,a_6,a_8,a_9\} & \varnothing & \{a_1,a_2,a_3,a_4,a_8,a_9\} & \varnothing & & & & & \\
C & \{a_4,a_6,a_7,a_{10}\} & C & \{a_4,a_6,a_7,a_{10}\} & \{a_1,a_2,a_3,a_4,a_6,a_7,a_8,a_9,a_{10}\} & \varnothing & & & & \\
C & \{a_4,a_6,a_7,a_{10}\} & C & \{a_4,a_6,a_7,a_{10}\} & \{a_1,a_2,a_3,a_4,a_6,a_7,a_8,a_9,a_{10}\} & \varnothing & \varnothing & & & \\
C & \{a_1,a_3,a_7,a_8\} & C & \{a_1,a_3,a_6,a_7,a_8\} & \{a_1,a_2,a_3,a_4,a_6,a_7,a_8,a_9\} & \{a_1,a_3,a_4,a_6,a_8,a_{10}\} & \{a_1,a_3,a_4,a_6,a_8,a_{10}\} & \varnothing &
\end{bmatrix}$$

每个粒下建立区分矩阵分别为

$$\boldsymbol{M}_{G_1}^R=\begin{bmatrix}
\varnothing & & & & & & & \\
\{a_1,a_3,a_8\} & \varnothing & & & & & & \\
\varnothing & \{a_1,a_3,a_8\} & \varnothing & & & & & \\
\{a_1,a_3,a_8\} & \varnothing & \{a_1,a_3,a_8\} & \varnothing & & & & \\
\{a_1\} & \{a_1,a_3,a_8\} & \{a_1\} & \{a_1,a_3,a_8\} & \varnothing & & & \\
\{a_1,a_3,a_8\} & \varnothing & \{a_1,a_3,a_8\} & \varnothing & \{a_1,a_3,a_8\} & \varnothing & & \\
\{a_1,a_3,a_8\} & \varnothing & \{a_1,a_3,a_8\} & \varnothing & \{a_1,a_3,a_8\} & \varnothing & \varnothing & \\
\{a_1,a_3,a_8\} & \{a_1,a_3,a_8\} & \{a_1,a_3,a_8\} & \{a_1,a_3,a_8\} & \{a_1,a_3,a_8\} & \{a_1,a_3,a_8\} & \{a_1,a_3,a_8\} & \varnothing
\end{bmatrix}$$

$$\boldsymbol{M}_{G_2}^{R}=\begin{bmatrix}
\varnothing & & & & & & & \\
\{a_2,a_4,a_9\} & \varnothing & & & & & & \\
\varnothing & \{a_2,a_4,a_9\} & \varnothing & & & & & \\
\{a_2,a_4,a_9\} & \varnothing & \{a_2,a_4,a_9\} & \varnothing & & & & \\
\{a_2,a_9\} & \{a_2,a_4,a_9\} & \{a_2,a_9\} & \{a_2,a_4,a_9\} & \varnothing & & & \\
\{a_2,a_4,a_9\} & \{a_4\} & \{a_2,a_4,a_9\} & \{a_4\} & \{a_2,a_4,a_9\} & \varnothing & & \\
\{a_2,a_4,a_9\} & \{a_4\} & \{a_2,a_4,a_9\} & \{a_4\} & \{a_2,a_4,a_9\} & \varnothing & \varnothing & \\
\{a_2,a_4,a_9\} & \varnothing & \{a_2,a_4,a_9\} & \varnothing & \{a_2,a_4,a_9\} & \{a_4\} & \{a_4\} & \varnothing
\end{bmatrix}$$

$$\boldsymbol{M}_{G_3}^{R}=\begin{bmatrix}
\varnothing & & & & & & & \\
\{a_5,a_6,a_7,a_{10}\} & \varnothing & & & & & & \\
\varnothing & \{a_5,a_6,a_7,a_{10}\} & \varnothing & & & & & \\
\{a_5,a_6,a_7,a_{10}\} & \{a_6\} & \{a_5,a_6,a_7,a_{10}\} & \varnothing & & & & \\
\{a_5,a_6,a_7,a_{10}\} & \{a_6\} & \{a_5,a_6,a_7,a_{10}\} & \varnothing & \varnothing & & & \\
\{a_5,a_6,a_7,a_{10}\} & \{a_6,a_7,a_{10}\} & \{a_5,a_6,a_7,a_{10}\} & \{a_6,a_7,a_{10}\} & \{a_6,a_7,a_{10}\} & \varnothing & & \\
\{a_5,a_6,a_7,a_{10}\} & \{a_6,a_7,a_{10}\} & \{a_5,a_6,a_7,a_{10}\} & \{a_6,a_7,a_{10}\} & \{a_6,a_7,a_{10}\} & \varnothing & \varnothing & \\
\{a_5,a_6,a_7,a_{10}\} & \{a_7\} & \{a_5,a_6,a_7,a_{10}\} & \{a_6,a_7\} & \{a_6,a_7\} & \{a_6,a_{10}\} & \{a_6,a_{10}\} & \varnothing
\end{bmatrix}$$

通过定义10可计算出：$\mathrm{sig}_{G_1}=2\times(19-2)=34$，$\mathrm{sig}_{G_2}=2\times(21-2)=38$，$\mathrm{sig}_{G_3}=2\times(22-3)=38$。

由于G_3和G_2的重要度相同，但是G_3中包含属性个数多，因此对粒排序为G_3,G_2,G_1。

(4)考虑G_3的属性粒中属性重要度，其中a_6的重要度最大：$\mathrm{sig}_{G_3}(a_5)=24$，$\mathrm{sig}_{G_3}(a_6)=48$，$\mathrm{sig}_{G_3}(a_7)=42$，$\mathrm{sig}_{G_3}(a_{10})=40$，因此，将$a_6$作为粒$G_3$的代表。

接着考虑G_2，其中a_4的重要度最大$\mathrm{sig}_{G_2}(a_4)=42$，因此，将$a_4$作为粒$G_2$的代表。

由于$\mathrm{pos}_{a_6,a_4}(D)\neq\mathrm{pos}_C(D)$，接着考虑$G_1$，其中$a_1$的重要度最大：$\mathrm{sig}_{G_1}(a_1)=42$，由于$\mathrm{pos}_{a_6,a_4,a_1}(D)=\mathrm{pos}_C(D)$，因此最终的属性约简为$\mathrm{red}=\{a_1,a_4,a_6\}$。

基于上述分析，本文设计了多粒度属性约简算法，为了提高算法的效率，每个属性粒建立区分矩阵以及求属性粒的重要度和属性粒中属性重要度时，由于求解过程是独立的，可以采用并行处理。

算法 基于区分矩阵的多粒度属性约简

输入：决策表$S=(U,C\cup D)$，阈值μ

输出：属性约简red

步骤1 $\forall a_i\in C$，计算依赖度$r(a_i,d)$。

步骤2 将$r(a_i,d)$降序排列，设排名为j的属性依赖度为r_j，若$r_j-r_{j-1}\leqslant\mu$，则属性划为一个属性粒。

步骤3 在每个属性粒下建立保持等价关系的区分矩阵：$\boldsymbol{M}_{G_1}^{R},\boldsymbol{M}_{G_2}^{R},\ldots,\boldsymbol{M}_{G_k}^{R}$。

步骤4 构建全属性下的正域区分矩阵$\boldsymbol{M}_C^P$。

步骤5 计算每个属性粒的重要度，并根据重要度的大小对k个属性粒进行降序排列。设排名为l的属性粒为$G_{[l]}$，其中$\bigcup_{l=1}^{k}G_{[l]}=C$，排序结果为$G=\{G_{[1]},G_{[2]},\cdots,G_{[k]}\}$。

步骤6 $\mathrm{red}=\varnothing$

While $\mathrm{pos}_{\mathrm{red}}(D)\neq\mathrm{pos}_C(D)$

for $l=1,2,\cdots,k$

$a^*=\underset{a\in G_{[l]}-\mathrm{red}}{\arg\max}\,\mathrm{sig}_{G_{[l]}}(a)$

$\mathrm{red}=\mathrm{red}\cup\{a^*\}$

If $\mathrm{pos}_{\mathrm{red}}(D)=\mathrm{pos}_C(D)$

Break;

End

End

End While

步骤7 输出red。

该算法步骤1计算决策关于每个属性的依赖度，其时间复杂度为$O(U)$；步骤2的时间复杂度为$O(|C|\log|C|)$；步骤3的时间复杂度为$O(|G_i||U|^2)$。步骤4的时间复杂度为$O(|C||U|^2)$；步骤5和6中计算属性粒的重要度和属性粒中属性重要度并行时间复杂度为$O(|U|^2)$，对k个粒进行降序排列时间复杂度为$O(k\log(k))$，属性粒内属性排序时间复杂度为$O(|G_i|\log|G_i|)$；步骤6的启发式约简过程时间复

杂度为$O(|C||U|)$。因此，计算多粒度属性约简最终的时间复杂度为$O(|C||U|^2)$。

4　实验分析

为了验证本文提出的基于区分矩阵的多粒度属性约简算法(Disernibility matrix multigranulation attribute reduction, DMR)的性能，把该算法与传统正域意义下的属性约简算法[5](Positive attribute reduction, AR_1)，基于熵的属性约简算法[6](Entropy attribute reduction, AR_2)的性能进行了比较。在5个UCI公开数据集和1个高维数据集(见表2)上对以上算法进行了实验测试。这些数据集中，数据集Zoo是离散的，其余均是连续的。对于连续型数据进行了离散化操作，将属性值离散化为4个取值。实验的硬件环境是3.60 GHz的CPU，8 GB的内存，编程语言为C#。

实验中，本文采用10折交叉验证方法，通过在6个数据集的属性粒化过程中将阈值μ设为0.001，并用KNN分类器(K=5)的分类精度评估属性约简算法的优劣。实验结果如表3所示，DMR与传统正域意义下的AR_1，AR_2相比，DMR算法获得的约简虽然数目较多，但是构造的KNN算法的分类性能比AR_1算法和AR_2算法更优。

为了提高算法的效率，可以将算法中求解独立的过程采用并行处理。本文提出的DMR算法可以将每个属性粒建立区分矩阵以及求属性粒的重要度和属性粒中属性重要度这3个步骤分别采取并行计算。从表4可以清楚地看出，所提出的多粒度属性约简算法采用并行计算所需的时间远远小于串行所用的时间。

表2　实验数据集

Tab. 2　Data sets used in experiments

Data sets	Samples	Attributes	Class
Wine	178	13	3
Ionosphere	351	34	2
Zoo	101	16	7
Dermatology	366	34	6
Auto	205	25	6
Lung	203	3312	5

表3　DMR和AR_1，AR_2约简结果的分类精度(ACC)和所选特征个数(NUM)的比较

Tab. 3　Comparision of DMR , AR_1 and AR_2 on classification accuracy and the number of selected attributes

Data set	DMR		AR_1		AR_2	
	ACC	NUM	ACC	NUM	ACC	NUM
Wine	0.932 4	7	0.870 6	6	0.849 0	6
Ionosphere	0.880 3	21	0.869 1	10	0.849 2	9
Zoo	0.880 9	9	0.852 7	5	0.841 8	5
Dermatology	0.937 2	26	0.846 7	10	0.902 1	11
Auto	0.628 1	20	0.529 0	13	0.529 0	11
Lung	0.826 9	10	0.798 3	5	0.807 1	5

表4　DMR算法采用串行和并行方式所用时间

Tab. 4　Running time of DMR algorithm by serial and parallel way　　s

Data set	serial	parallel
Wine	0.030 1	0.015 9
Ionosphere	1.532 1	0.813 0
Zoo	0.016 9	0.009 9
Dermatology	0.352 1	0.118 7
Auto	0.050 8	0.033 9
Lung	2.750 2	2.083 9

5　结　　论

本文从多粒度的角度开展属性约简算法研究.论文基于属性依赖度构造了属性粒。利用属性粒上区分矩阵与全部属性区分矩阵之间的差异定义了属性粒的重要度，并利用属性粒中的属性在区分矩阵中出现的频率定义了属性的重要度。在此基础上，运用这两种重要度设计了基于区分矩阵的多粒度属性约简算法，该算法可以平衡各个属性粒中属性对约简结果的贡献，得到的约简结果更具有代表性和差异性，实验结果表明了算法的有效性。

参考文献：

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

[2] HU Qinghua, XIE Zongxia, YU D. Hybrid attribute reduction based on a novel fuzzy rough model and information granulation[J]. Pattern Recognition, 2007, 40(12): 3509-3521.

[3] WEI Wei, LIANG Jiye, WANG Junhong, et al. Decision-relative discernibility matrices in the sense of entropies[J]. International Journal of General Systems, 2013, 42(7): 721-738.

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

[5] HU Xiaohua, NICK C. Learning in relational databases: A rough set approach[J]. International Journal of Computational Intelligence, 1995, 11 (2): 323-338.

[6] WANG Guoyin, YU Hong, YANG Dachun. Decision table reduction based on conditional information entropy[J]. Chinese Journal of Computers, 2002, 25 (7): 759-766.

[7] ZADEH L A. Fuzzy sets and information granularity, advances in fuzzy set theory and application[M].North-Holland: Amsterdam Publishing, 1979: 3-18.

[8] CHEN Yuming, MIAO Duoqian, JIAO Na. Attribute reduction based on binary granules and granular computing[J]. Journal of Guangxi Normal University, 2008, 26(2): 81-824.

[9] PAL S K, SHANKAR B U, MITRA P. Granular computing, rough entropy and object extraction[J]. Pattern Recognition Letters, 2005, 26(16): 2509-2517.

[10] ZADEH L A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic [J]. Fuzzy Sets & Systems, 1997, 90 (90): 111-127.

[11] WU Wei, LEUNG Y, MI J S. Granular computing and knowledge reduction in formal contexts[J]. IEEE Transactions on Knowledge & Data Engineering, 2009, 21(10): 1461-1474.

[12] LIANG Jiye, QIAN Yuhua. Information granules and entropy theory in information systems[J]. Science in China Series F (Information Science), 2008, 51 (10): 1427-1444.

[13] YAO Jingtao, VASILAKOS A V, PEDRYCZ W. Granular computing: Perspectives and challenges[J]. IEEE Transactions on Cybernetics, 2013, 43 (6): 1977-1989.

[14] QIAN Yuhua, LIANG Jiye. Rough set method based on multigranulations [C]//Proceedings of the 5th IEEE International Conference on Cognitive Informatics. Beijing, China: IEEE, 2006: 297-304.

[15] QIAN Yuhua, LIANG Jiye, YAO Yiyu, et al. MGRS: A multi-granulation rough set[J]. Information Sciences, 2010, 180(6): 949-970.

[16] 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 Yanli, QIAN Yuhua. A granular space reduction approach to pessimistic multi-granulation rough sets [J]. Pattern Recognition and Artificial Intelligence, 2012, 25(3): 361-366.

[17] QIAN Yuhua, ZHANG Hu, SANG Yanli, et al. Multigranulation decision-theoretic rough sets[J]. International Journal of Approximate Reasoning, 2014, 55 (1): 225-237.

[18] QIAN Yuhua, LIANG Jiye, DANG Chuangyin. Incomplete multigranulation rough set[J]. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 2010,40(2): 420-431.

[19] YANG Xibei, SONG Xiaoning, CHEN Zehua, et al. On multigranulation rough sets in incomplete information system[J]. International Journal of Machine Learning & Cybernetics, 2012,3(3): 223-232.

[20] XU Weihua, WANG Qiaorong, ZHANG Xiantao. Multi-granulation fuzzy rough sets in a fuzzy tolerance approximation space[J]. International Journal of Fuzzy Systems, 2011, 13(4): 246-259.

[21] 吴志远, 钟培华, 胡建根. 程度多粒度粗糙集[J]. 模糊系统与数学, 2014, 28(3): 165-172.
WU Zhiyuan, ZHONG Peihua, HU Jiangen. Graded multi-granulation rough sets [J]. Fuzzy Systems and Mathematics, 2014, 28(3): 165-172.

[22] 张明, 程科, 杨习贝, 等. 基于加权粒度的多粒度粗糙集[J]. 控制与决策, 2015(2): 222-228.
ZHANG Ming, CHENG Ke, YANG Xibei, et al. Multigranulation rough set based on weighted granulations[J]. Control and Decision, 2015(2): 222-228.

[23] XU Weihua, LI Wentao, ZHANG Xiantao. Generalized multigranulation rough sets and optimal granularity selection[J]. Granular Computing, 2017, 2 (4): 271-288.

(编辑:刘彦东)