

# 相容关系下的分割知识约简算法的研究

焦娜

(华东政法大学信息科学与技术系 上海 201620)

**摘要** 粗糙集理论是一个能有效地删除冗余特征的工具。由于实际应用的数据往往是连续的,并且结构复杂、特征多,现有的粗糙集知识约简方法对真实复杂的数据计算效率较低。为此,首先将相容关系应用于粗糙集的知识约简,再将复杂的信息表纵向分割成简单的缩减表和小规模信息表,然后把缩减表和小规模信息表连接起来进行知识约简。实例表明,提出的方法能够有效提高粗糙集对复杂数据的计算效率。

**关键词** 粗糙集理论,相容关系,复杂数据,分割

**中图分类号** TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.1.011

## Research on Vertical Segmentation Knowledge Reduction Algorithm Based on Tolerance Rough Set Theory

JIAO Na

(Department of Information Science and Technology, East China University of Political Science and Law, Shanghai 201620, China)

**Abstract** Rough set theory is an efficient mathematical tool for further reducing redundancy. However, practical data sets are always continuous and the structure is complex. The efficiency of Many existing knowledge reduction methods based on rough set theory are low. This paper combined tolerance relation together with rough set theory. Then, we put forward a new knowledge reduction method based on vertical segmentation. The large information system was divided into one reduced form and other small scale information forms, then they were joined together in order to solve the original information system. An example demonstrates that the proposed algorithm is effective.

**Keywords** Rough set theory, Tolerance relation, Complex data, Vertical segmentation

## 1 引言

自20世纪80年代Pawlak提出粗糙集理论(Rough Set Theory)以来,它已经被广泛应用在特征选择等领域中<sup>[1-3]</sup>。知识约简一直是粗糙集理论的重要研究内容之一<sup>[4,5]</sup>。由于实际应用的数据往往是连续的,并且结构复杂、特征多<sup>[6,7]</sup>,现有的粗糙集特征选择方法对真实复杂的数据实用性较差、计算效率较低。

实用性差主要是因为传统的粗糙集理论模型主要通过等价关系得到知识约简。传统粗糙集模型运用等价关系计算知识约简时,对处理的数据要求较为严格,即要求数据是离散的数值(如整数、字符串型、枚举型)。由于真正获取的数据往往存在连续值,必须对其进行离散化<sup>[8,9]</sup>,但离散化过程必定会造成某种程度的信息损失。而相容粗糙集则放宽了等价类的严格划分条件,从而避免了一定程度的信息丢失。利用相容粗糙集方法对特征值为连续值的数据进行选取,能最大限度地保持数据集的分类能力<sup>[10]</sup>。计算效率低主要是因为在大数据时代信息表的结构复杂、特征多。因此,本文提出一种分割的方法。分割的思想被用到很多研究领域中<sup>[11-16]</sup>。本文的分割方法主要是把一个大规模的复杂的数据表分割成若干个小的更容易处理的小规模数据表,这些小规模数据表能够用

现有的方法来进行处理,将它们的结果合并到一起能够解决原来的大规模数据表。分割的方法能够使原来的大规模数据表变得更容易解决并且消耗的时间更少。

由于传统粗糙集特征选择算法对真实复杂数据的计算效率较低,本文把分割的思想运用到特征选择算法中,提出相容关系下的分割知识约简算法。实例表明,本文提出的算法是有效可行的。

## 2 粗糙集理论的基本概念

下面简要介绍经典粗糙集的基本概念。

**定义1**  $T=(U, A, V, f)$  是一个信息表,  $U$  为非空有限论域;  $A=C \cup D$  是特征全集,  $C, D$  分别为条件特征集和决策特征集;  $V=\bigcup_{a \in A} V_a$  为特征值域;  $f: U \times A \rightarrow V$  为信息函数。

特征子集  $R \subseteq A$  决定了一个不可分辨关系  $IND(R)$ :

$$IND(R) = \{(x, y) \in U \times U \mid \forall a \in R, f(x, a) = f(y, a)\}$$

显然,  $IND(R)$  是一个等价关系。论域  $U$  关于等价关系  $R$  的划分  $U/IND(R)$  定义为:

$$U/IND(R) = \{[x]_R : x \in U\}$$

$[x]_R$  表示  $U$  中所有与  $x$  在关系  $IND(R)$  下是等价的元素构成的集合,  $[x]_R = \{y \mid \forall a \in R, f(x, a) = f(y, a)\}$ , 即等价类。

**定义2** 在信息表  $T=(U, A, V, f)$  中,  $X \subseteq U$  关于  $R \subseteq A$

的下近似集和上近似集分别记为:  $\underline{RX}, \overline{RX}$ 。

$$\underline{RX} = \bigcup \{[x]_R \in U/IND(R) \mid [x]_R \subseteq X\}$$

$$\overline{RX} = \bigcup \{[x]_R \in U/IND(R) \mid [x]_R \cap X \neq \emptyset\}$$

**定义 3** 在信息表  $T=(U, A, V, f)$  中,  $X \subseteq U$ , 决策特征集  $D$  相对条件特征子集  $R \subseteq C$  的正区域、负区域和边界域分别定义为:

$$PSO_R(D) = \bigcup_{X \in U/IND(D)} \underline{RX}$$

$$NEG_R(D) = U - \bigcup_{X \in U/IND(D)} \overline{RX}$$

$$BND_R(D) = \bigcup_{X \in U/IND(D)} \overline{RX} - \bigcup_{X \in U/IND(D)} \underline{RX}$$

### 3 相容关系的粗糙集

在传统粗糙集中, 要求数据是离散的, 连续值的数据需要进行离散化, 离散化过程会带来信息丢失。为此, 本文定义了相容度、相容关系及相对应的上、下近似及正区域等概念, 用相容关系来代替等价关系, 从而减少了信息损失。

**定义 4** 在信息表  $T=(U, A, V, f)$  中,  $a \in A, x, y \in U$ , 特征  $a$  对于任意两个对象  $x, y$  间的相容度定义为:

$$B_a(x, y) = \frac{|a(x) - a(y)|}{|a_{\max} - a_{\min}|}$$

$a_{\max}, a_{\min}$  分别表示特征  $a$  的最大值和最小值。对于特征子集, 对象间综合相容度有如下定义。

**定义 5** 在信息表  $T=(U, A, V, f)$  中,  $R \subseteq A, x, y \in U$ , 任意两个对象  $x, y$  关于特征子集  $R$  的综合相容度定义为:

$$B_{R, \tau}(x, y) = \left\{ (x, y) \mid \frac{\sum_{a \in R} B_a(x, y)}{|R|} \leq \tau \right\}$$

其中,  $\tau \in (0, 0.5]$  是一个阈值。

**定义 6** 在信息表  $T=(U, A, V, f)$  中,  $R \subseteq A, \tau \in (0, 0.5]$ , 论域  $U$  关于特征子集  $R$  的  $\tau$  相容关系定义为:

$$B_{R, \tau} = \{(x, y) \in U \times U \mid \forall a \in R, (x, y) \in B_{R, \tau}(x, y)\}$$

论域  $U$  关于相容关系  $B_{R, \tau}$  的划分用  $U/B_{R, \tau}$  表示:

$$U/B_{R, \tau} = \{B_{R, \tau}(x) \mid x \in U\}$$

**定义 7** 在信息表  $T=(U, A, V, f)$  中,  $X \subseteq U, R \subseteq C, \tau \in (0, 0.5]$ ,  $X$  关于  $R$  的  $\tau$  下近似集和上近似集分别记为:  $\underline{R}_\tau X, \overline{R}_\tau X$ 。

$$\underline{R}_\tau X = \{x \mid B_{R, \tau}(x) \subseteq X\}$$

$$\overline{R}_\tau X = \{x \mid B_{R, \tau}(x) \cap X \neq \emptyset\}$$

$\langle \underline{R}_\tau X, \overline{R}_\tau X \rangle$  为相容粗糙集。

**定义 8** 在信息表  $T=(U, A, V, f)$  中,  $X \subseteq U, R \subseteq C, \tau \in (0, 0.5]$ , 决策特征集  $D$  相对条件特征子集  $R$  的  $\tau$  正区域、负区域和边界域分别定义为:

$$PSO_{R, \tau}(D) = \bigcup_{X \in U/B_{D, \tau}} \underline{R}_\tau X$$

$$NEG_{R, \tau}(D) = U - \bigcup_{X \in U/B_{D, \tau}} \overline{R}_\tau X$$

$$BND_{R, \tau}(D) = \bigcup_{X \in U/B_{D, \tau}} \overline{R}_\tau X - \bigcup_{X \in U/B_{D, \tau}} \underline{R}_\tau X$$

### 4 相容关系下的分割知识约简算法

在大数据时代, 信息表越来越庞大, 结构越来越复杂, 特征也越来越多。对于一个庞大的信息表, 用传统的粗糙集方法进行计算时效率较低。为了提高结构复杂、特征多的大规

模信息表的计算效率, 本文提出了分割约简算法。首先把一个大规模信息表转化成一个缩减表和几个小规模信息表, 缩减表由决策特征集和中间特征集组成, 每个中间特征对应表示各个小规模信息表的关键字, 小规模信息表由条件特征子集和中间特征组成; 然后把每个小规模信息表代入缩减表进行特征约简, 最终得到原来大规模信息表的约简结果。该方法使原来的大规模信息表变得更容易解决并且用时较少。下面先介绍缩减表和小规模信息表的若干定义, 然后提出相容关系下的分割知识约简方法。

#### 4.1 基本定义和性质

把一个信息表分割成一个缩减表和几个小规模信息表, 主表由决策特征集和几个中间特征组成, 中间特征是小规模信息表中的关键字。小规模信息表由条件特征的子集及中间特征构成。

**定义 9** 给定信息表  $T=(U, A, V, f)$ ,  $A=C \cup D$ 。

(1) 小规模信息表  $T^{R_i} = \langle U^{R_i}, A^{R_i}, V^{R_i}, f^{R_i} \rangle (i=1, 2, \dots, k)$ , 其中  $U^{R_i}$  为非空有限论域;  $A^{R_i} = R_i \cup \{r_i\}, R_i \subseteq C (i=1, 2, \dots, k), C = \bigcup_{i=1}^k R_i, R_i \cap R_j = \emptyset (i \neq j), r_i$  是小规模信息表与缩减表的中间特征, 它是  $T^{R_i}$  的关键字;  $V^{R_i} = \bigcup_{a \in R_i} V_a^{R_i}$ , 其中  $V_a^{R_i}$  为特征  $a$  的值域;  $f^{R_i}: U^{R_i} \times R_i \rightarrow V^{R_i}$  为信息函数。

(2) 缩减表  $\Gamma = \langle U^\Gamma, A^\Gamma, V^\Gamma, f^\Gamma \rangle$ , 其中  $U^\Gamma$  为非空有限论域;  $A^\Gamma = \bigcup_{i=1}^k \{r_i\} \cup D, \bigcup_{i=1}^k \{r_i\}$  是中间特征集,  $D$  为决策特征集;  $V^\Gamma = \bigcup_{a \in A^\Gamma} V_a^\Gamma$ , 其中  $V_a^\Gamma$  为特征  $a$  的值域;  $f^\Gamma: U^\Gamma \times A^\Gamma \rightarrow V^\Gamma$  为信息函数。

(3) 过渡表  $T^{M_i} = \langle U^{M_i}, A^{M_i}, V^{M_i}, f^{M_i} \rangle (i=1, 2, \dots, k)$ , 其中  $U^{M_i}$  为非空有限论域;  $A^{M_i} = (\bigcup_{i=1}^k \{r_i\} - \{r_i\}) \cup R_i \cup D (i=1, 2, \dots, k), D$  为决策特征集;  $V^{M_i} = \bigcup_{a \in A^{M_i}} V_a^{M_i}$ , 其中  $V_a^{M_i}$  为特征  $a$  的值域;  $f^{M_i}: U^{M_i} \times A^{M_i} \rightarrow V^{M_i}$  为信息函数。

根据上面的定义, 有下列性质。

**性质 1** 缩减表  $\Gamma$  的中间特征  $r_i$  是不必要的, 即  $POS_{(\bigcup_{i=1}^k \{r_i\} - \{r_i\})} (D) = POS_{(\bigcup_{i=1}^k \{r_i\})} (D)$ , 当且仅当信息表  $T$  的与中间特征  $r_i$  相对应的条件特征集  $R_i$  是不必要的, 即  $POS_{(C-R_i)} (D) = POS_{(C)} (D)$ 。

**性质 2** 过渡表  $T^{M_i}$  的条件特征  $a$  是不必要的, 即  $PSO_{(A^{M_i}-D-\{a\})} (D) = PSO_{(A^{M_i}-D)} (D)$ , 当且仅当信息表  $T$  的条件特征  $a$  是不必要的, 即  $POS_{(C-\{a\})} (D) = POS_{(C)} (D)$ 。

**性质 3** 若主表  $\Gamma$  的中间特征  $r_i$  是必要的, 即  $POS_{(\bigcup_{i=1}^k \{r_i\} - \{r_i\})} (D) \neq POS_{(\bigcup_{i=1}^k \{r_i\})} (D)$ , 那么在特征表  $T$  中一定存在与中间特征  $r_i$  相对应的条件特征集  $R_i$  的子集  $A$  是必要的, 即  $A \subseteq R_i, POS_{(C-A)} (D) \neq POS_{(C)} (D)$ 。

**性质 4** 若过渡表  $T^{M_i}$  的条件特征  $a$  是必要的, 即  $PSO_{(A^{M_i}-D-\{a\})} (D) \neq PSO_{(A^{M_i}-D)} (D)$ , 那么信息表  $T$  的条件特征  $a$  是必要的, 即  $POS_{(C-\{a\})} (D) \neq POS_{(C)} (D)$ 。

#### 4.2 相容关系下的分割知识约简算法

相容关系下的分割知识约简算法的具体步骤描述如下。

步骤 1 设小规模信息表个数为  $k$ 。

步骤 2 将信息表  $T$  分割成一个缩减表和  $k$  个小规模信息表。

步骤3 设 $\tau$ 为初值,根据定义4—定义6,计算 $k$ 个小规模信息表的相容关系及相容关系下的划分。

步骤4 确定 $k$ 个小规模信息表中的中间特征 $r_i$ ,并确定缩减表 $\Gamma=\langle U^\Gamma, A^\Gamma, V^\Gamma, f^\Gamma \rangle$ 和 $k$ 个小规模信息表 $T^{R_i}=\langle U^{R_i}, A^{R_i}, V^{R_i}, f^{R_i} \rangle (i=1,2,\dots,k)$ 。

步骤5 令 $Red=C$ 。

步骤6 对缩减表 $\Gamma=\langle U^\Gamma, A^\Gamma, V^\Gamma, f^\Gamma \rangle$ 中的每个中间特征 $r_i (i=1,2,\dots,k)$ 进行判断,计算去掉该中间特征 $r_i$ 的 $\tau$ 的正区域和决策特征集 $D$ 相对条件特征集 $C$ 的 $\tau$ 的正区域是否相同。

i)若相同, $POS_{(\bigcup_{i=1}^k \{r_i\}-\{r_i\})}(D)=POS_{(\bigcup_{i=1}^k \{r_i\})}(D)$ ,则继续执行步骤7。

ii)若不同, $POS_{(\bigcup_{i=1}^k \{r_i\}-\{r_i\})}(D) \neq POS_{(\bigcup_{i=1}^k \{r_i\})}(D)$ ,则转到步骤8。

步骤7  $r_i$ 是不必要的特征,在 $\Gamma$ 中删除该中间特征; $Red=Red-R_i$ ;合并相同的对象,转到步骤6。

步骤8 将一个小规模信息表 $T^{R_i}$ 与主表 $\Gamma$ 合并成中间表 $T^{M_i}$ 。

步骤9 对过渡表 $T^{M_i}=\langle U^{M_i}, A^{M_i}, V^{M_i}, f^{M_i} \rangle (i=1,2,\dots,k)$ 中的每个特征 $a$ 进行判断,计算去掉该特征 $a$ 的 $\tau$ 的正区域和决策特征集 $D$ 相对条件特征集 $C$ 的 $\tau$ 的正区域是否相同。

i)若相同, $PSO_{(A^{M_i}-D-\{a\})}(D)=PSO_{(A^{M_i}-D)}(D)$ ,则继续执行步骤10。

ii)若不同, $PSO_{(A^{M_i}-D-\{a\})}(D) \neq PSO_{(A^{M_i}-D)}(D)$ ,则转到步骤9。

步骤10  $a$ 是不必要的特征,在 $T^{M_i}$ 中删除该特征; $Red=Red-\{a\}$ ;  $Red=Red-\{a\}$ ;转到步骤9。

步骤11 循环结束,输出约简结果 $Red$ 。

## 5 实例分析

为说明提出的相容关系的分割知识约简算法,本文对表1所列的信息表进行约简。表1包含4个条件特征、1个决策特征,表中的特征值都是连续的。设 $\tau=0.2, A=C \cup D, C=\{H, P, Q, S\}, D=\{W\}$ 。

表1 数据集

| 对象    | H   | P   | Q  | S  | W  |
|-------|-----|-----|----|----|----|
| $o_1$ | 0.8 | 133 | 4  | 54 | 5  |
| $o_2$ | 0.9 | 38  | 18 | 12 | 34 |
| $o_3$ | 0.8 | 40  | 19 | 13 | 6  |
| $o_4$ | 5   | 134 | 3  | 55 | 36 |

设小规模信息表个数为 $k=2$ 。将信息表分割成一个缩减表和2个小规模信息表。缩减表中包含的特征集有两个中间特征和一个决策特征 $\{r_1, r_2, W\}$ ;第一个小规模信息表 $T^{R_1}$ 包含的特征集是信息表1的前两个特征和一个中间特征 $\{H, P, r_1\}$ ;第二个小规模信息表 $T^{R_2}$ 包含的特征集是信息表1的后两个特征和一个中间特征 $\{Q, S, r_2\}$ 。

两个小规模信息表的相容类分别是:

$$U/B_{D,\tau}=\{\{o_1, o_3\}, \{o_2, o_4\}\}$$

$$U^{R_1}/B_{H,\tau}=\{\{o_1, o_2, o_3\}, \{o_4\}\}$$

$$U^{R_1}/B_{P,\tau}=\{\{o_1, o_4\}, \{o_2, o_3\}\}$$

$$U^{R_1}/B_{(H,P),\tau}=\{\{o_1\}, \{o_2, o_3\}, \{o_4\}\}$$

$$U^{R_2}/B_{Q,\tau}=\{\{o_1, o_4\}, \{o_2, o_3\}\}$$

$$U^{R_2}/B_{S,\tau}=\{\{o_1, o_4\}, \{o_2, o_3\}\}$$

$$U^{R_2}/B_{(Q,S),\tau}=\{\{o_1, o_4\}, \{o_2, o_3\}\}$$

缩减表的相容类是:

$$U^\Gamma/B_{W,\tau}=\{\{o_1\}, \{o_2, o_3\}, \{o_4\}\}$$

根据上面相容类的结果确定的缩减表如表2所列,2个小规模信息表如表3、表4所列。

表2 缩减表

| 对象    | $r_1$   | $r_2$   | W     |
|-------|---------|---------|-------|
| $o_1$ | $r_1^1$ | $r_2^1$ | $d^1$ |
| $o_2$ | $r_1^2$ | $r_2^2$ | $d^2$ |
| $o_3$ | $r_1^1$ | $r_2^2$ | $d^1$ |
| $o_4$ | $r_1^3$ | $r_2^1$ | $d^2$ |

表3 第一个小规模信息表

| 对象    | H   | P   | $r_1$   |
|-------|-----|-----|---------|
| $o_1$ | 0.8 | 133 | $r_1^1$ |
| $o_2$ | 0.9 | 38  | $r_1^2$ |
| $o_3$ | 0.8 | 40  | $r_1^2$ |
| $o_4$ | 5   | 134 | $r_1^3$ |

表4 第二个小规模信息表

| 对象    | Q  | S  | $r_2$   |
|-------|----|----|---------|
| $o_1$ | 4  | 54 | $r_2^1$ |
| $o_2$ | 18 | 12 | $r_2^2$ |
| $o_3$ | 19 | 13 | $r_2^2$ |
| $o_4$ | 3  | 55 | $r_2^1$ |

令 $Red=C$ 。

对缩减表 $\Gamma=\langle U^\Gamma, A^\Gamma, V^\Gamma, f^\Gamma \rangle$ 中的每个中间特征 $r_i (i=1,2,\dots,k)$ 进行判断,计算去掉该中间特征 $r_i$ 的 $\tau$ 的正区域和决策特征集 $D$ 相对条件特征集 $\bigcup_{i=1}^k \{r_i\}$ 的 $\tau$ 的正区域是否相同。

相对于条件特征集 $\bigcup_{i=1}^k \{r_i\}$ 的 $\tau$ 的下近似为:

$$\underline{\bigcup_{i=1}^k \{r_i\}}_\tau \{o_1, o_3\} = \{r_1, r_2\}_\tau \{o_1, o_3\} = \{o_1\}$$

$$\underline{\bigcup_{i=1}^k \{r_i\}}_\tau \{o_2, o_4\} = \{r_1, r_2\}_\tau \{o_2, o_4\} = \{o_4\}$$

因此,决策特征集 $D$ 相对于的条件特征集 $\bigcup_{i=1}^k \{r_i\}$ 的 $\tau$ 的 $\beta$ -正区域为:

$$POS_{\bigcup_{i=1}^k \{r_i\}-\{r_1\},\tau}(D) = \bigcup_{X \in U^\Gamma/B_{D,\tau}} \underline{R_\tau X} = \underline{\bigcup_{i=1}^k \{r_i\}}_\tau \{o_1, o_3\} \cup$$

$$\underline{\bigcup_{i=1}^k \{r_i\}}_\tau \{o_2, o_4\} = \{o_2, o_4\} = POS_{C,\tau}(D)$$

对于条件特征子集 $\bigcup_{i=1}^k \{r_i\} - \{r_1\}$ ,决策特征集 $D$ 相对条件特征子集 $\bigcup_{i=1}^k \{r_i\} - \{r_1\}$ 的 $\tau$ 的 $\beta$ -正区域为:

$$POS_{\bigcup_{i=1}^k \{r_i\}-\{r_1\},\tau}(D) = \bigcup_{X \in U^\Gamma/B_{D,\tau}} \underline{R_\tau X} = \{r_2\}_\tau \{o_1, o_3\} \cup$$

$$\{r_2\}_\tau \{o_2, o_4\} = \emptyset$$

$$POS_{\bigcup_{i=1}^k \{r_i\}-\{r_1\},\tau}(D) \neq POS_{\bigcup_{i=1}^k \{r_i\},\tau}(D)$$

其中, $r_1$ 是必要的中间特征,不可被删除。

对于条件特征子集 $\bigcup_{i=1}^k \{r_i\} - \{r_2\}$ ,决策特征集 $D$ 相对于

条件特征子集  $\bigcup_{i=1}^k \{r_i\} - \{r_2\}$  的  $\tau$  的  $\beta$  正区域为:

$$POS_{\bigcup_{i=1}^k \{r_i\} - \{r_2\}, \tau}(D) = \bigcup_{x \in U^T / B_{D, \tau}} R_x X = \{r_1\}_\tau \{o_1, o_3\} \cup \{r_1\}_\tau \{o_2, o_4\} = \{o_2, o_4\}$$

$$POS_{\bigcup_{i=1}^k \{r_i\} - \{r_2\}, \tau}(D) = POS_{\bigcup_{i=1}^k \{r_i\}, \tau}(D)$$

其中,  $r_2$  是不必要的中间特征, 可被删除。

根据性质 1 和性质 3,  $Red = Red - \{r_2\} = Red - R_2 = Red - \{Q, S\} = \{H, P\}$ , 合并相同的对象。

将一个小规模信息表  $T^{R_1}$  与主表  $\Gamma$  合并成中间表  $T^{M_1}$  (见表 5), 并去掉不必要的特征 (见表 6)。

表 5 去掉  $r_2$  之前的过渡表  $T^{M_1}$

| 对象    | H   | P   | $r_2$   | W  |
|-------|-----|-----|---------|----|
| $o_1$ | 0.8 | 133 | $r_2^1$ | 5  |
| $o_2$ | 0.9 | 38  | $r_2^2$ | 34 |
| $o_3$ | 0.8 | 40  | $r_2^3$ | 6  |
| $o_4$ | 5   | 134 | $r_2^1$ | 36 |

表 6 去掉  $r_2$  之后的过渡表  $T^{M_1}$

| 对象    | H   | P   | W  |
|-------|-----|-----|----|
| $o_1$ | 0.8 | 133 | 5  |
| $o_2$ | 0.9 | 38  | 34 |
| $o_3$ | 0.8 | 40  | 6  |
| $o_4$ | 5   | 134 | 36 |

对过渡表  $T^{M_1}$  中的每个特征进行判断, 计算去掉该特征后的  $\tau$  的正区域和决策特征集  $D$  相对于条件特征集  $C$  的  $\tau$  的正区域是否相同。

$$POS_{\{H, P\} - \{H\}, \tau}(D) = \{o_1\}$$

$$POS_{\{H, P\} - \{P\}, \tau}(D) = \emptyset$$

所以,  $POS_{\{H, P\} - \{H\}, \tau}(D) \neq POS_{C, \tau}(D)$ ,  $POS_{\{H, P\} - \{P\}, \tau}(D) \neq POS_{C, \tau}(D)$ ,  $H$  和  $P$  都是必要特征, 不可被删除, 因此  $Red = \{H, P\}$ 。

**结束语** 伴随着新兴技术的兴起, 大数据在我们的生活中无处不在, 而且数据结构复杂。针对大数据环境下出现的复杂信息表, 本文提出了相容关系下的分割知识约简算法, 该方法不仅能够处理连续值的信息, 而且可以把大规模的、复杂的、不好处理的信息表分割成小规模、简单的、更容易处理的信息表, 能够有效提高粗糙集理论对复杂数据信息表的计算效率。

## 参考文献

- [1] Pawlak Z. Rough sets[J]. International Journal of Information Computer Science, 1982, 11(5): 341-356
- [2] Wu W Z, Zhang W X, Xu Z B. Characterizing Rough Fuzzy Sets in Constructive and Axiomatic Approaches[J]. Chinese Journal of Computers, 2004, 27(2): 197-203(in Chinese)  
吴伟志, 张文修, 徐宗本. 粗糙模糊集的构造与公理化方法[J]. 计算机学报, 2004, 27(2): 197-203
- [3] Zhang Xiu-wen, Wu Wei-zhi, Liang Ji-ye, et al. Rough set theory and method[M]. Beijing: Science Press, 2001(in Chinese)  
张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001

- [4] Wang G Y. Rough Set Theory and Knowledge Acquisition[M]. Xi'an: Xi'an Jiaotong University Press, 2001(in Chinese)  
王国胤. 粗糙集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001
- [5] Miao D Q, Hu G R. A Heuristic Algorithm for Reduction of Knowledge[J]. Journal of Computer Research & Development, 1999, 36(6): 681-684(in Chinese)  
苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(6): 681-684
- [6] Li J Z, Liu X M. An Important Aspect of Big Data: Data Usability[J]. Journal of Computer Research & Development, 2013, 50(6): 1147-1162(in Chinese)  
李建中, 刘显敏. 大数据的一个重要方面: 数据可用性[J]. 计算机研究与发展, 2013, 50(6): 1147-1162
- [7] Meng X F, Ci X. Big Data Management: Concepts, Techniques and Challenges[J]. Journal of Computer Research & Development, 2013, 50(1): 146-169(in Chinese)  
孟小峰, 慈祥. 大数据管理: 概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1): 146-169
- [8] Miao D Q. A New Method of Discretization of Continuous Attributes in Rough Sets[J]. ACTA Automatica SINICA, 2001, 27(3): 296-302(in Chinese)  
苗夺谦. Rough set 理论中连续属性的离散化方法[J]. 自动化学报, 2001, 27(3): 296-302
- [9] Grzymala-Busse J W. Discretization of numerical attributes [M] // Klösgen W, Zytkow J, eds. Handbook of Data Mining and Knowledge Discovery. Oxford University Press, 2002: 218-225
- [10] Parthalaín N M, Shen Q. Exploring the boundary region of tolerance rough sets for feature selection[J]. Pattern Recognition, 2009, 42: 655-667
- [11] Bazan J G, Latkowski R, Szczuka M. Missing template decomposition method and its implementation in rough set exploration system[C] // Proceedings of the Fifth International Conference on Rough Sets and Current Trends in Computing. Kobe, LNAI, 2006: 254-263
- [12] Zhang Q Z. An Approach to Rough Set Decomposition of Incomplete Information Systems[C] // IEEE Conference on Industrial Electronics and Applications. Harbin, IEEE, 2007: 2455-2460
- [13] Lin G P, Liang J Y, Qian Y H. An information fusion approach by combining multigranulation rough sets and evidence theory [J]. Information Sciences, 2015, 314(9): 184-199
- [14] Li H, Li L, Zhang J X. Multi-focus image fusion based on sparse feature matrix decomposition and morphological filtering[J]. Optics Communications, 2015, 342(5): 1-11
- [15] Fischer B. Decomposition of time series—comparing different methods in theory and practice[R]. Eurostat Working Paper, 1995
- [16] Wang M, Liu B, Tang J H, et al. Metric learning with feature decomposition for image categorization [J]. Neurocomputing, 2010, 73(10-12): 1562-1569