

## 以优势关系为基础的粗糙集在地震数据挖掘中的应用

朱冰冰<sup>1</sup>, 吴绍春<sup>1</sup>, 王 炜<sup>2</sup>

(1. 上海大学 计算机工程与科学学院, 上海 200072; 2. 上海市地震局, 上海 200062)

**摘 要:** 在地震数据挖掘应用中, 可用粗糙集方法进行对震例数据的属性约减。但是, 经典的粗糙集理论建立在由等价关系对对象集划分的基础上, 而震例数据是有序的而不是分类的对象。现对经典粗糙集理论进行扩展, 提出一种用优势关系代替等价关系的粗糙集方法, 并在此基础上提出一种基于差别矩阵的属性约简算法。实验结果表明, 用这种方法能得出一些采用传统粗糙集理论所无法得到的结果。

**关键词:** 粗糙集理论; 优势关系; 地震学

**中图分类号:** TP311.13 **文献标识码:** A

## Application of dominance rough set in seismology

ZHU Bing-bing<sup>1</sup>, WU Shao-chun<sup>1</sup>, WANG Wei<sup>2</sup>

(1. School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China;

2. Earthquake Administration of Shanghai Municipality, Shanghai 200062, China)

**Abstract:** In seismology, rough set can be used in the reduction of attributes. But classical rough set are based on indiscernibility relation or similarity relation. The problems in Seismology deal with ordering objects instead of classifying objects. Now we presented an extension of the classical rough set called dominance-based rough set, and an attribute-reduction algorithm based on dominance matrix. The experiment indicates that the results will be more meaningful than those induced by the classical rough set theory.

**Key words:** rough set; dominance relations; seismology

### 0 引言

在现实世界中, 很多属性或类别间是具有优势关系的。以地震中的震例数据为例, 它的属性包括地震活动性指标、地震强度因子  $M_f$  值、 $b$  值、 $h$  值、 $D$  值、震情指数、波速、断层总面积、震级等。把前面若干项当作条件属性, 而把震级当作决策属性。条件属性不但是以一种有序状态存在的, 而且存在着一种优势关系。例如, 地震强度因子  $M_f$  值越大, 它的震级一般来说也越大, 如果忽略这一点可能就会得出错误的结论。假设现在有两条数据  $A$  和  $B$ , 如果  $A$  的  $M_f$  值较  $B$  大, 一般来说它的震级也应该较  $B$  大。如果此时  $A$  的震级却小于  $B$ , 这显然是不太正常的。但是如果用传统的粗糙集方法就无法发现这一点, 甚至可能得出错误的结论。为此, 在地震数据挖掘中引入基于优势关系的粗糙集理论, 从而充分利用地震数据本身的特点, 有效去除噪音, 从大量属性中找出最有决定意义的属性。

### 1 传统的粗糙集基本概念

粗糙集把客观世界或对象世界抽象成一个信息系统, 也称属性一值系统。

**定义 1** 信息系统。一个信息系统  $I$  是一个四元组:  $I = (U, A, V, f)$ 。

其中,  $U$  就是论域, 设  $U$  有  $n$  个对象, 则  $U$  可表示为  $U =$

$\{x_1, x_2, \dots, x_n\} \subseteq A$  是有限个属性的集合, 设有  $m$  个属性, 则其可表示成  $A = \{a_1, a_2, \dots, a_m\}$ 。 $A$  又可进一步划分成两个不相交的集合: 条件属性集  $AT$  和决策属性集  $D$ ,  $AT$  和  $D$  满足  $AT \cap D = \emptyset$ ,  $D$  一般只有一个属性, 即使有多个属性, 一般也可用等价类的方法转化成一个。 $V$  是属性的值域集,  $V = \{V_1, V_2, \dots, V_m\}$ , 其中  $V_i$  是属性  $a_i$  的值域。 $f$  是信息函数:  $U \times A \rightarrow V$ ,  $f(x_i, a_j) \in V_j$ 。信息系统通常以决策表的形式给出, 如表 1 所示, 该表是一个关于苹果颜色、大小与成熟度关系的例子。

表 1 决策表示例

$U$	颜色(C)	大小(S)	熟否(M)
$O_1$	红	大	是
$O_2$	红	中	是
$O_3$	红	小	否
$O_4$	青	大	是
$O_5$	青	中	否
$O_6$	青	小	否

粗糙集理论就是将用于分类的知识嵌入到集合内, 作为集合的一个组成部分。根据现有的知识判断, 一个对象  $a$  是否属于集合  $X$  有三种情况: (1) 对象  $a$  肯定属于集合  $X$ ; (2) 对象  $a$  肯定不属于集合  $X$ ; (3) 对象  $a$  可能属于也可能不属于集合  $X$ 。由此出现了粗糙集中的两个重要概念: 下近似和上近似。设  $X \subseteq U$  是一组对象, 对于一个等价关系  $R$ , 即  $R \subseteq AT$  是一组

收稿日期: 2006-06-28; 修订日期: 2006-08-28

基金项目: 国家地震科学联合基金资助项目(104090); 上海市自然科学基金资助项目(7A05468)

作者简介: 朱冰冰(1977-), 女, 山东兖州人, 助理馆员, 硕士, 主要研究方向: 数据挖掘; 吴绍春(1965-), 女, 江西宜春人, 副教授, 博士, 主要研究方向: 数据挖掘; 王炜(1947-), 男, 江苏南京人, 研究员, 主要研究方向: 地震预报、数据挖掘。

条件属性,则上下近似定义如下:

**定义 2** 下近似。 $X$  相对于  $R$  的下近似是:

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

**定义 3** 上近似。 $X$  相对于  $R$  的上近似是:

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

$\underline{R}(X)$  是一定能归入  $X$  的记录集合,即所有包含于  $X$  的  $[x]_R$  的并集。 $\overline{R}(X)$  是  $U$  中一定或可能归入  $X$  的对象的集合,即所有与  $X$  的交集不为零的  $[x]_R$  的并集。

由上近似和下近似,可以得到正域、负域和边界域的概念。

**定义 4** 正域。集合  $X$  相对于  $R$  的正域就是  $X$  的下近似,即:

$$POS_R(D) = \underline{R}(X)$$

**定义 5** 负域。集合  $X$  相对于  $R$  的负域是:

$$NEG_R(X) = U - \overline{R}(X)$$

**定义 6** 边界域。集合  $X$  相对于  $R$  的边界域是:

$$BND_R(X) = \overline{R}(X) - \underline{R}(X)$$

边界域中的元素是可能属于也可能不属于  $X$  的对象组成的集合,如果边界域为空,则称集合  $X$  是关于  $R$  的精确集,反之,称  $X$  是关于  $R$  的粗糙集。

**定义 7** 省略和不可省略。设一等价关系为属性集合  $R \subseteq C$ , 属性  $r \in R$ , 当等价类  $Ind(R) = Ind(R - \{r\})$ , 称  $r$  为  $R$  中可省略的, 否则  $r$  为  $R$  中不可省略的。

**定义 8** 约简。属性子集  $R \subseteq C$ , 如果  $POS_R(D) = POS_C(D)$  对于  $R$  的任意子集有 (1) 不真, 则称  $R$  为  $C$  的一个约简, 记为  $Red(D)$ , 简记为  $Red$ 。

**定义 9** 核。 $C$  中所有不可省略的属性的集合称为  $C$  的核, 即:

$$Core(D) = \cap Red(D)$$

以上面的表 1 为例,  $R = S \in AT$ , 有:  $U/S = \{S_1, S_2, S_3\}$ , 其中  $S_1 = \{O_1, O_4\}$ ,  $S_2 = \{O_2, O_5\}$ ,  $S_3 = \{O_3, O_6\}$ 。  $D = M$ ,  $U/M = \{M_1, M_2\}$ ,  $M_1 = \{O_1, O_2, O_4\}$ ,  $M_2 = \{O_3, O_5, O_6\}$ 。

根据上面的定义:

$$\underline{R}(M_1) = S_1 = \{O_1, O_4\}$$

$$\overline{R}(M_1) = S_1 \cup S_2 = \{O_1, O_2, O_4, O_5\}$$

$$\underline{R}(M_2) = S_3 = \{O_3, O_6\}$$

由此可以获得这样的结论, 如果苹果是大的, 一定是熟的, 如果苹果是小的, 一定是不熟的, 中等大小的苹果可能是熟的, 也可能是不熟的。

$$\overline{R}(M_2) = S_2 \cup S_3 = \{O_2, O_3, O_5, O_6\}$$

$$POS_R(M) = \underline{R}(M_1) \cup \underline{R}(M_2) = \{O_1, O_3, O_4, O_6\}$$

$$NEG_R(M_1) = \{O_3, O_6\}$$

$$NEG_R(M_2) = \{O_1, O_4\}$$

$$BND_R(M_1) = \overline{R}(M_1) - \underline{R}(M_1) = \{O_2, O_5\}$$

$$BND_R(M_2) = \overline{R}(M_2) - \underline{R}(M_2) = \{O_2, O_5\}$$

由上可知  $Core(M) = \{C, S\}$ 。

## 2 扩展的粗糙集基本概念

**定义 10** 有序决策表。有序决策表是一个四元集合,  $I = (U, AT \cup \{d\}, V, f)$  ( $d \notin AT, * \notin V_d$ )。

$AT$  是有限非空条件属性集,  $d$  是决策属性,  $V$  是属性的值域,  $f$  是映射 ( $U \times AT \cup \{d\} \rightarrow V$ ),  $f(x, a) \in V_a$  ( $\forall a \in AT \cup \{d\}, \forall x \in U$ )。在这里它同一般决策表的不同在于, 属性的值域是降序或升序有序的。 $*$  表示某条记录在对应属性上的值不可知。如果  $\exists * \in V_a, a \in AT$ , 称此决策表为非完备有序决策表, 否则就称为完备的有序决策表。

**定义 11** 优势关系。称  $\geq_a$  为相对于  $a$  具有优势,  $x \geq_a y$  表示  $x$  在  $a$  上的值大于或等于  $y$  ( $a \in A, x, y \in U$ )。  $x \geq_a y \Leftrightarrow \forall a \in A, x \geq_a y$ 。在这里, 属性的值域可以是连续的, 也可以是离散的, 但是它们之间必须是有序的。对一个给定的有序决策表, 如果  $A \subseteq AT, x \geq_A y$ , 那么  $x$  在  $A$  上对于  $y$  具有优势, 写成  $xR_A^{\geq} y$ 。

$$R_A^{\geq} = \{(x, y) \in U \times U \mid x \geq_A y\}$$

由此可以定义  $x$  在属性集  $A$  上的优势集合:

$$[x]_A^{\geq} = \{y \in U \mid f(y, a_1) \geq f(x, a_1) (\forall a_1 \in A_1) \text{ and } f(y, a_2) \leq f(x, a_2) (\forall a_2 \in A_2)\}$$

$$[x]_A^{\leq} = \{y \in U \mid f(y, a_1) \geq f(x, a_1) (\forall a_1 \in A_1) \text{ and } f(y, a_2) \leq f(x, a_2) (\forall a_2 \in A_2)\}$$

$A = A_1 \cup A_2$ , 其中  $A_1$  是上向的属性集,  $A_2$  是下向的属性集。

在非完备有序决策表中:

$$R_A^{*\geq} = \{(y, x) \in U \times U \mid (\forall a_1 \in A_1, f(y, a_1) \geq f(x, a_1) \text{ or } f(x, a_1) = * \text{ or } f(y, a_1) = *) \text{ and } (\forall a_2 \in A_2, f(y, a_2) \leq f(x, a_2) \text{ or } f(x, a_2) = * \text{ or } f(y, a_2) = *)\}$$

**定理 1**

1) 优势关系是自反、传递和非对称的, 所以不是等价关系。

2) 如果  $B \subseteq A \subseteq AT$ , 那么  $R_B^{\geq} \supseteq R_A^{\geq} \supseteq R_{AT}^{\geq}$ ;

3) 如果  $B \subseteq A \subseteq AT$ , 那么  $[x]_B^{\geq} \supseteq [x]_A^{\geq} \supseteq [x]_{AT}^{\geq}$ ;

4) 如果  $x_j \in [x_i]_A^{\geq}$ , 那么  $[x_j]_A^{\geq} \subseteq [x_i]_A^{\geq}$ , 并且  $[x_i]_A^{\geq} = \cup \{[x_j]_A^{\leq} : x_j \in [x_i]_A^{\geq}\}$ ;

5)  $[x_i]_A^{\geq} = [x_j]_A^{\geq}$ , 当且仅当  $f(x_i, a) = f(x_j, a) (\forall a \in A)$ ;

6)  $\{[x]_A^{\geq} \mid x \in U\}$  形成  $U$  上的一个覆盖。

以上定理也适用于非完备有序决策表。

**证明:**

1) 由优势关系的定义:  $x \geq_a y$  表示  $x$  在  $a$  上的值大于或等于  $y$ , 易知它是自反、传递和非对称的, 所以不是等价关系。

2) 因为  $B \subseteq A$ , 所以  $\forall x, y \in U$ , 如果  $x \geq_A y$ , 则  $x \geq_B y$ , 由此可得  $R_B^{\geq} \supseteq R_A^{\geq}$ ; 同理可得  $R_A^{\geq} \supseteq R_{AT}^{\geq}$ , 所以  $R_B^{\geq} \supseteq R_A^{\geq} \supseteq R_{AT}^{\geq}$ 。

3) 由上文  $[x]_A^{\geq}$  的定义可以得出  $[x]_A^{\geq} = \{y \in U \mid y \geq_A x\}$ 。因为  $B \subseteq A$ , 则  $\forall y \in [x]_A^{\geq}, y \in [x]_B^{\geq}$ , 所以  $[x]_B^{\geq} \supseteq [x]_A^{\geq}$  同理可证, 如果  $A \subseteq AT, [x]_A^{\geq} \supseteq [x]_{AT}^{\geq}$ 。由此结论得证。

4) 如果  $x_j \in [x_i]_A^{\geq}$ , 那么  $x_j \geq_A x_i$ ; 如果  $y \in [x_j]_A^{\geq}$ , 那么  $y \geq_A x_j \geq_A x_i$ 。所以  $[x_j]_A^{\geq} \subseteq [x_i]_A^{\geq}$ ,  $[x_i]_A^{\geq} = \cup \{[x_j]_A^{\leq} : x_j \in [x_i]_A^{\geq}\}$ 。

5)  $\forall y \in [x_i]_A^{\geq}, y \in [x_j]_A^{\geq}$ , 则  $y \geq_A x_i, y \geq_A x_j$ 。所以  $x_i =_A x_j$ 。

6) 由  $x \in [x]_A^{\geq}$ , 可知  $U \subseteq \{[x]_A^{\geq} \mid x \in U\}$ , 又由  $[x]_A^{\geq}$  的定义可知  $[x]_A^{\geq} \subseteq U$ 。所以参照覆盖的定义: 设  $U$  为论域,  $C$  非空且  $\cup C = U$ , 则称  $C$  是  $U$  的一个覆盖, 可知结论  $\{[x]_A^{\geq} \mid x \in U\}$  形成  $U$  上的一个覆盖成立。证毕。

$\forall X \subset U, \forall A \subset AT, X$  对于优势关系  $R_A^{\geq}$  的上下近似集定义如下:

$$R_A^{\geq}(X) = \{x \in X \mid [x]_A^{\geq} \subseteq X\}$$

$$\overline{R_A^{\geq}}(X) = \{x \in X \mid [x]_A^{\geq} \cap X \neq \emptyset\}$$

由此可以看出下近似集是一定属于  $X$  的, 而上近似集包括可能属于  $X$  和一定属于  $X$  的元素。

**定理 2** 假设四元集  $I = (U, AT, V, f)$  是非完备集,  $X \subseteq U, A \subseteq B \subseteq AT$ , 那么:

$$(1) \underline{R_A^{\geq}}(X) \subseteq X \subseteq \overline{R_A^{\geq}}(X)$$

$$(2) \underline{R_A^{\geq}}(X) \supseteq \underline{R_B^{\geq}}(X)$$

$$\overline{R_A^{\geq}}(X) \supseteq \overline{R_B^{\geq}}(X)$$

证明:

1) 由  $X$  对于优势关系  $R_A^{\geq}$  的上下近似集, 定义可以轻易的得出此结论。

2) 因为  $A \subseteq B$ , 所以  $[x]_B^{\geq} \subseteq [x]_A^{\geq}$ , 再由  $X$  对于优势关系  $R_A^{\geq}$  的上下近似集的定义可以得出  $\underline{R_A^{\geq}}(X) \supseteq \underline{R_B^{\geq}}(X)$ ,  $\overline{R_A^{\geq}}(X) \supseteq \overline{R_B^{\geq}}(X)$ 。证毕。

另外, 根据决策属性  $d$  将  $U$  划分为有限个类。假设  $CL = \{CL_t, t \in T\}, T = \{1, 2, \dots, n\}$ , 表示这些有序类, 也就是说, 如果  $r > s, (r, s \in T)$   $CL_r$  优先于  $CL_s$ 。决策类的  $t$ -上并集和  $t$ -下并集分别定义为:

$$CL_t^{\geq} = \cup_{s \geq t} CL_s, \text{ 声明 } x \in CL_t^{\geq} \text{ 表明 } x \text{ 至少属于类 } CL_t;$$

$$CL_t^{\leq} = \cup_{s \leq t} CL_s, \text{ 声明 } x \in CL_t^{\leq} \text{ 表明 } x \text{ 至多属于类 } CL_t$$

$CL_t^{\geq}$  对优势关系  $R_A^{\geq}$  的上下近似集被定义为:

$$\underline{R_A^{\geq}}(CL_t^{\geq}) = \{x \in U \mid [x]_A^{\geq} \subseteq CL_t^{\geq}, \overline{R_A^{\geq}}(CL_t^{\geq}) =$$

$$\bigcup_{x \in CL_t^{\geq}} [x]_A^{\geq}, t = 1, \dots, n$$

$CL_t^{\geq}$  的临界域被定义为:

$$Bn_A(CL_t^{\geq}) = \overline{R_A^{\geq}}(CL_t^{\geq}) - \underline{R_A^{\geq}}(CL_t^{\geq})$$

### 3 有序决策表的约简

下面从震例数据中随机地抽出一些属性和数据, 说明如何构造差别矩阵。

表 2 有序决策表示例

样本	属性				
	$Mf$	$DM$	震情指数	波速	震级
X1	1.0	0.96	0.12	*	5.3
X2	1.7	0.91	0.15	3.9	5.1
X3	2.7	0.89	0.13	2.8	4.9
X4	1.4	*	0.07	3.1	4.8
X5	3.1	0.85	0.16	5.2	6.0
X6	2.9	0.74	0.16	4.1	6.6
X7	2.1	0.79	0.17	4.4	5.7

表 2 中数据的意义如下: 一般来讲, 根据先验经验, 对决策属性震级来讲,  $Mf$ 、震情指数和波速都是上向属性, 即  $Mf$ 、

震情指数、波速越大, 震级越大; 而  $DM$  是下向属性, 即  $DM$  越小, 震级越大(其中的 \* 表示数据缺失)。

首先定义决策属性上的优势关系:

$$R_d^{\geq} = \{(x, y) : f(d, x) \geq f(d, y)\}$$

由此定义不完备差别矩阵如下:

$$D_p^*(x, y) = \begin{cases} \{a \in AT : (x, y) \notin R_a^{\geq}\} & (x, y) \in R_d^{\geq} \\ \emptyset & (x, y) \notin R_d^{\geq} \end{cases}$$

根据上面对差别矩阵的定义, 可以得到上述决策表的差别矩阵如表 3(为了表示方便, 设  $b$  表示  $Mf$ ,  $c$  表示  $DM$ , 震情指数为  $d$ , 波速为  $e$ , 震级为  $f$ )。

表 3 差别矩阵

$x$	$y$						
	X1	X2	X3	X4	X5	X6	X7
X1		$bcd$	$bcd$	$bce$			
X2			$bc$	$c$			
X3				$ce$			
X4							
X5	$e$			$c$			$cd$
X6	$e$			$c$	$bde$		$de$
X7	$e$		$b$	$c$			

现定义差别矩阵的约简函数:

$$\Delta^* = \bigwedge_{(x, y) \in U \times U} \bigvee D_p^*(x, y)$$

所以对于上述矩阵可以得到其约简函数为:

$$b \wedge c \wedge e \wedge (b \vee c) \wedge (c \wedge d) \wedge (c \vee e) \wedge (d \vee e) \wedge (b \vee c \vee e) \wedge (b \vee d \vee e) \wedge (b \vee c \vee d \vee e) = (b \wedge c \wedge e) \vee (b \wedge c \wedge d \wedge e)$$

从上面的结果可以看出,  $bce$  和  $bcd$  是两个约简, 而  $bce$  是核。

这里还要给出一个定理, 由此找出不协调的数据记录。

**定理 3** 如果差别矩阵的某一项包括全部条件属性, 那么就意味着这两条记录不协调, 其中有一条可能是噪音。

证明 如果他们不相等, 由差别矩阵的定义可知,  $\forall a \in AT, f(a, x) < f(a, y), f(d, x) \geq f(d, y)$ , 所以这两项不协调。

在上面的例子中, 我们认为记录 1 和 2, 3 是不协调的。由于记录 1 与两条记录相悖, 所以认为记录 1 有问题, 将其删掉。重构差别矩阵如表 4。

表 4 改进后的差别矩阵

$x$	$y$						
	X2	X3	X4	X5	X6	X7	
X2		$bc$	$c$				
X3			$ce$				
X4							
X5			$c$				$cd$
X6			$c$	$bde$			$de$
X7		$b$	$c$				

这时得到其约简函数:

$$b \wedge c \wedge (b \vee c) \wedge (c \vee d) \wedge (c \vee e) \wedge (d \vee e) \wedge (b \vee d \vee e) = (b \wedge c \wedge d) \vee (b \wedge c \wedge e) \vee (b \wedge c \wedge d \wedge e)$$

从上面的结果可以得出三个约简  $bce$ ,  $bcd$  和  $bcd$ ,  $bc$  是核。即利用  $Mf$ 、 $DM$  和震情指数就可推断出震级的大概范围。其中  $Mf$ 、 $DM$  对决定震级大小有至关重要的作用。

决策规则总结如下:

- 4.8)  $X2 \sim X7$  支持  
 $(b, \geq, 1.4) \wedge (c, \leq, 0.96) \wedge (d, \geq, 0.07) \rightarrow (f, \geq,$   
 4.8)  $X2 \sim X7$  支持  
 $(b, \geq, 1.4) \wedge (c, \leq, 0.96) \wedge (e, \geq, 2.8) \rightarrow (f, \geq,$   
 4.8)  $X2 \sim X7$  支持  
 $(b, \geq, 2.9) \wedge (d, \geq, 0.16) \wedge (e, \geq, 4.1) \rightarrow (f, \geq,$   
 6.0)  $X5, X6$  支持  
 $(c, \leq, 0.85) \wedge (d, \geq, 0.16) \rightarrow (f, \geq, 5.7) \quad X5, X6,$   
 $X7$  支持

仔细观察上面的数据可以发现,由于非完备决策表中优势关系的定义:

$$R_A^{\geq} = \{(y, x) \in U \times U \mid (\forall a_1 \in A_1, f(y, a_1) \geq f(x, a_1) \text{ or } f(x, a_1) = * \text{ or } f(y, a_1) = *) \text{ and } (\forall a_2 \in A_2, f(y, a_2) \leq f(x, a_2) \text{ or } f(x, a_2) = * \text{ or } f(y, a_2) = *)\}$$

我们认为,如果  $f(x, a) = *$ , 那么  $f(x, a) = f(y, a) (\forall y \in U)$ 。这样会造成噪音数据增加,对不协调的限定过于严格,相应得到的决策规则也会变少。而在地震数据中可能存在一些不为人知的因素,使得在数据集中存在一些轻微的不协调,所以在地震数据挖掘中把条件放宽,将决策属性上的优势关系定义为:  $R_d^{\geq} = \{(x, y) : f(d, x) > f(d, y)\}$ 。

同时将差别矩阵中的项定义为:

$$\{a \in AT : (x, y) \notin R_{|a|}^{\geq}\}$$

这样,只有当记录  $A$  的决策属性  $>$  记录  $B$  的决策属性,但是  $A$  的对应条件属性都小于  $B$  时,我们才认为它们不协调。

## 4 实验

上面给出一个简单的实例来说明差别矩阵的生成和属性的约简,对于震例数据这种数据集大、属性多的情况,提出如下改进措施:

依据决策属性,首先对所有的数据记录按升序进行排序。由于计算差别矩阵某一项的条件是  $(x, y) \in R_d^{\geq}$ , 否则这一项就为空。所以可以根据已排好序的队列依次计算差别矩阵的每一项。首先读入第一条和第二条数据,如果  $(x, y) \in R_d^{\geq}$ , 则依次比较它们对应条件属性的大小,如果第一条数据的属性小于第二条,就把它存入差别矩阵的相应项中。然后再读入第三条,同第一条作比较。如此逐个进行,直到全部比较完毕。如此就可以得到差别矩阵。

### 4.1 算法步骤

#### 1) 生成区别矩阵

如果某一项属性个数  $|c| = |AT|$ , 那么  $g(x) = g(x) + 1$ ,  $x$  是两条记录中决策属性较小的那一条,即列项。 $g(x)$  表示与记录  $x$  相悖的记录数。

如果某一项  $c$  的属性个数为 1, 就把该属性直接列入最终的属性约简集合  $Red$  中;

如果区别矩阵中包含核属性,那么就删除此项。否则更新属性频度函数  $f(a) = f(a) + |AT| / |c|, \forall a \in c$ 。

#### 2) 合并相同项

如果某一项的  $g(c)$  大于某一给定值  $A$ , 就将其删除,否则将其合并,按每一项的长度和频度进行排序(即首先按长度排序,长度相同的情况下按频度排序),生成  $M$ 。

### 3) 求解约简

```
For  $M$  中的每一项  $m$  do
  If ( $m \cap Red = \emptyset$ )
     $Red = Red \cup \{a\}$  // 选择  $m$  中  $f(a)$  较大的  $a$  加入  $Red$ 
  Endif
Endfor
Return  $Red$ 
```

### 4.2 算法的复杂度分析

步骤 1 中生成差别矩阵的代价是  $O(|AT| |U|^2)$ , 因此步骤 2 中合并相同项并排序的时间复杂度是  $O(|U|^2 \log |U|^2) = O(2 |U|^2 \log |U|)$ 。但这只是最坏情况,实际上差别矩阵的项要远小于最坏情况。在步骤 3 遍历并生成约简中,由于差别矩阵中最多有  $|U|(|U| - 1)/2$  项,每一项最多有  $|AT|$  个属性,因此最坏时间复杂度为  $O(|AT| |U|^2)$ 。而且经过步骤 1(去除包含核属性的项)和步骤 2(去噪及合并)后,只剩下得多的项。

### 4.3 实验结果

在实验中采用了 Java 中的聚集框架及其中的算法,从而加强了程序的规范性、代码的复用性及可读性。通过实验,我们从现有的五百多条震例数据中去除了大约 1/5, 并从 27 个条件属性中得到其核属性集为  $\{EmaxEall, DM, Mmax, A\}$ 。从而说明这四个属性对震级大小具有至关重要的作用。利用算法步骤中的第三步对 UCI 中的标准数据集进行属性约简,并与 ROSETTA 得出的结果进行比较发现,利用这种算法进行属性约简无法保证最终的属性约简结果是最佳的,但是在大多数情况下,它是可以得到最小约简的。即使得到的结果不是最佳的,在绝大部分情况下也是次佳的。

### 参考文献:

- [1] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11(5): 341-356.
- [2] GRECO S, MATARAZZO B, SLOWINSKI R. Rough approximation of a preference relation by dominance relations[R]. Warsaw: Warsaw University of Technology, 1996.
- [3] GRECO S, MATARAZZO B, SLOWINSKI R. Rough sets methodology for sorting problems in presence of multiple attributes and criteria[J]. European Journal of Operational Research, 2002, 138(2): 247-259.
- [4] GRECO S, MATARAZZO B, SLOWINSKI R. Rough sets theory for multicriteria decision analysis[J]. European Journal of Operational Research 2001, 129(1): 11-47.
- [5] GRECO S, MATARAZZO B, SLOWINSKI R. A new rough set approach to multicriteria and multiattribute classification[A]. Rough sets and Current Trends in Computing (RSTCTC'98), Lecture Notes in Artificial Intelligence[C]. Berlin: Springer-Verlag, 1998. 60-67.
- [6] SLOWINSKI R, GRECO S, MATARAZZO B. Rough set analysis of preference-ordered data[A]. RSTCTC 2002, LNAI 2475[C]. Berlin Heidelberg: Springer-Verlag, 2002. 44-59.
- [7] DEMBICZYNSKI K, PINDUR R, SUSMAGA R. Dominance-based rough set classifier without induction of decision rules[J]. Electronic Notes in Theoretical Computer Science, 2003, 82(4): 147-161.
- [8] 安利平, 陈增强, 袁著祉. 基于粗集理论的多属性决策分析[J]. 系统工程学报, 2004, 19(6): 559.
- [9] 贾戎莉. 信息系统上的优势关系与保序关系. 山西师范大学学报, 2005, 19(2): 14.