文章编号:1000-6788(2007)01-0131-06

一种新的信息系统属性约简算法

高学东,丁　军
(北京科技大学 管理学院,北京 100083)

摘要:　在分析目前已有基于 Rough Set 的属性约简算法后,给出了一个新的度量属性重要性的计算公式,分析了该计算公式的性质,然后给出了一个时间复杂度为 $\max\{O(|A||U|\log|U|),O(|A|^2|U|)\}$ 的快速属性约简算法,最后用一个实例说明了算法的有效性.

关键词:　Rough Set;信息系统;属性重要度;属性约简;算法复杂度
中图分类号:　TP18　　**文献标志码**:　A

A New Attribute Algorithm for Reduction of Information System

GAO Xue-dong, DING Jun
(School of Management, Science and Technology University of Beijing, Beijng 100083, China)

Abstract:　After analyzing the attribute reduction algorithm based on Rough Set that has arisen at present, a new formula for measuring the importance of attribution was given, and the property of this formula was analyzed. Then a new algorithm for attribution reduction was provided. The time complexity of the provided algorithm is $\max\{O(|A||U|\log|U|),O(|A|^2|U|)\}$. At last, the efficiency of the new algorithm was illustrated with an example.

Key words:　Rough Set; Information system; attribute importance; attribute reduction; algorithm complexity

1　引言

粗糙集理论[1,2]是一种新的处理模糊和不确定知识的数学工具,其主要思想是在保持信息系统的分类能力不变的前提下,通过属性约简,导出问题的决策或分类规则,广泛地应用于机器学习、决策分析、过程控制、模式识别和数据挖掘等领域[1,2].

信息系统的最佳属性约简是 NP-hard 问题,解决这一问题通常采用启发式搜索方法[3]. 文献[4,5]给出了较好的启发函数,使得属性约简算法的最坏时间复杂度为 $O(|A|^3|U|^2)$ (其中 $|A|$ 表示属性的个数,$|U|$ 表示信息系统对象的个数),文献[6]给出了一个最坏时间复杂度为 $O(|A|^2|U|^2)$ 的属性约简算法,使得算法的复杂度降低一个数量级;文献[7]通过快速排序方法又将属性约简算法的最坏时间复杂度降为 $O(|A|^2|U|\log|U|)$;文献[8]在研究属性重要性的计算时,给出了新的策略,使得算法的效率提高,但最坏时间复杂度仍为 $O(|A|^2|U|\log|U|)$. 本文对文献[5]的启发函数和属性的重要性计算进行充分分析后,给出了一个新的、较好的、度量属性重要性的计算公式,并分析了其性质,然后给出了一个最坏时间复杂度为 $\max\{O(|A||U|\log|U|),O(|A|^2|U|)\}$ 的属性约简算法,最后用一实例说明了新算法的有效性.

2　信息系统

定义 2.1　四元组 $S=(U,A,V,f)$ 是一个信息系统,其中 U 表示对象的非空有限集合,称为论域;A 表示属性的非空有限集;$V=\bigcup_{a\in A}V_a$,V_a 是属性 a 的值域;f 表示 $U\times A\to V$ 是一个信息函数,它对一个对象的每一个属性赋予一个信息值,即 $\forall a\in A$,$x\in U$,有 $f(x,a)\in V_a$,每一个属性子集 $P\subseteq A$ 决定了一个二元不可区分关系 $IND(P)$:

收稿日期:2005-01-26

作者简介:高学东(1963-),男,教授,博士生导师,研究方向:管理过程优化,数据挖掘;丁军(1978-),男,博士研究生,研究方向:数据挖掘,粗糙集理论及应用.

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

关系 $IND(P)$, $P \subseteq A$, 构成了 U 的一个划分, 用 $U/IND(P)$ 表示, 简记为 U/P, U/P 中的任何元素 $[x]_P = \{y \in U \mid \forall a \in P, f(x,a) = f(y,a)\}$ 称为等价类.

令 $P, Q \subseteq A$, 记 $U/P = \{P_1, P_2, \cdots, P_t\}$, $U/Q = \{Q_1, Q_2, \cdots, Q_s\}$, 若 $\forall P_i \in U/P \Rightarrow \exists Q_j \in U/Q$ 使 $P_i \subseteq Q_j$, 并且 $\forall Q_j \in U/Q \Rightarrow \exists P_i \in U/P$, 使 $Q_j \subseteq P_i$, 则称 $U/P = U/Q$; 若 $\forall P_i \in U/P \Rightarrow \exists Q_j \in U/Q$ 使 $P_i \subseteq Q_j$, 则称 U/P 为 U/Q 的加细, 常记为 $U/P \subseteq U/Q$.

引理 2.1 在信息系统 $S = (U, A, V, f)$, $\forall Q \subseteq P \subseteq A$, 则有 $U/P \subseteq U/Q$.

证明 记 $U/P = \{P_1, P_2, \cdots, P_t\}$, $U/Q = \{Q_1, Q_2, \cdots, Q_s\}$, 任取 $P_i = [x]_P \in U/P$, 则 $P_i = [x]_P = \{y \mid \forall a \in P, f(y,a) = f(x,a)\}$, 由于 $Q \subseteq P$, 故有 $Q_j = [x]_Q = \{y \mid \forall a \in Q, f(y,a) = f(x,a)\}$ 使得 $P_i \subseteq Q_j$. 由 P_i 的任意性知: $U/P \subseteq U/Q$.

定义 2.2 设 $S = (U, A, V, f)$ 为一个信息系统, $\forall P \subseteq A$, 若 $a \in P$ 使得 $IND(P - \{a\}) = IND(P)$, 则称属性 a 在 P 中是不必要的; 否则称属性 a 在 P 中是必要的.

定义 2.3 设 $S = (U, A, V, f)$ 为一个信息系统, $\forall P \subseteq A$, 若 $\forall a \in P$ 在 P 中都是必要的, 则称 P 为独立的, 否则, 称 P 为相依的.

定义 2.4 设 $S = (U, A, V, f)$ 为一个信息系统, $\forall P \subseteq A$, 如果满足 $IND(P) = IND(A)$, 且 P 为独立的, 则称 P 为 A 的一个属性约简.

3 属性的重要性分析及其计算

定义 3.1[5] 设 $S = (U, A, V, f)$ 为信息系统, $\forall P \subseteq A$, $U/P = \{P_1, P_2, \cdots, P_t\}$, 知识 P 的信息量定义为:

$$I(P) = 1 - \frac{1}{|U|^2} \sum_{i=1}^{t} |P_i|^2.$$

定义 3.2[5] 设 $S = (U, A, V, f)$ 为信息系统, 设 $P \subseteq A$, $\forall a \in A - P$, 属性 a 的重要性定义为:

$$sig_P(a) = I(P \cup \{a\}) - I(P). \tag{1}$$

下面以表 1 所示信息系统说明上述度量属性重要性的公式不能较好地反映信息系统中属性的重要性.

属性 a 将信息系统分为: $\{x_1, x_3, x_6\}$, $\{x_2, x_4, x_7\}$, $\{x_5, x_8, x_9, x_{10}\}$; 属性 b 将信息系统分为: $\{x_1, x_3, x_6\}$, $\{x_2, x_4, x_5, x_7, x_9\}$, $\{x_8, x_{10}\}$; 属性 c 将信息系统分为: $\{x_1, x_2, x_3, x_4\}$, $\{x_6, x_7\}$, $\{x_5, x_8, x_9, x_{10}\}$; 属性 d 将信息系统分为: $\{x_1, x_2, x_3, x_4, x_5, x_7, x_9\}$, $\{x_6\}$, $\{x_8, x_{10}\}$; 属性 e 将信息系统分为: $\{x_1, x_2, x_3, x_4, x_6\}$, $\{x_5, x_7, x_8, x_9, x_{10}\}$. 用公式(1)计算它们的重要性见表 2.

由表 2 可知属性 d 的重要性小于 e 的重要性, 从直觉上这是不合理的, 因为 d 把信息系统分为三块, 而 e 将信息系统分为二块. 从直觉上讲, 当然希望属性能将信息系统划分的块越多越好, 因为这样, 属性约简的速度就会更快, 也较大可能性得到最佳约简[7].

另一方面, 记 $U/A = \{A_1, A_2, \cdots, A_k\}$ 中的块 A_i ($i = 1, 2, \cdots, k$) 为基本块, 由引理 1 可知, 任何属性不可能再将 A_i 分细(即将 A_i 的元素加以区分), 即任何属性不可区分 A_i 中的任何两个对象. 因此, 若有属性在信息系统上的划分中有基本块, 则基本块在下一次划分之前可以去掉(因为不可能再将其细分), 这样将 U 缩小.

表 1 信息系统例

U	a	b	c	d	e
x_1	1	2	1	3	1
x_2	2	3	1	3	1
x_3	1	2	1	3	1
x_4	2	3	1	3	1
x_5	3	3	3	3	2
x_6	1	2	2	1	1
x_7	2	3	2	3	2
x_8	3	1	3	2	2
x_9	3	3	3	3	2
x_{10}	3	1	3	2	2

表 2　表 1 中信息系统属性的重要性

	a	b	c	d	e
重要性	66/100	62/100	64/100	46/100	50/100

例如在表 1 中 A 将信息系统划分为:$\{x_1, x_3\}, \{x_2, x_4\}, \{x_5, x_9\}, \{x_6\}, \{x_7\}, \{x_8, x_{10}\}$,有 6 个基本模块. 在 d 的划分中有 2 个基本模块$\{x_6\}, \{x_8, x_{10}\}$,如果选中 d 时,在 d 之后加入任何属性与 d 一起构成的划分中$\{x_6\}, \{x_8, x_{10}\}$不会再被划细,这样,$\{x_6\}, \{x_8, x_{10}\}$在下一次划分时,就不必要再被划分,因而,可以直接从 U 中去掉,需要进一步划分的是$\{x_1, x_2, x_3, x_4, x_5, x_7, x_9\}$. 这样论域由原来的 10 个对象变为 7 个对象了,在 7 个对象上的划分自然比 10 个对象上的划分要简单且容易,公式 1 是无法能体现这一性质的,因而,我们基于如下观点提出一个新的度量属性重要性的计算公式:要求保证分类的个数尽量多,且尽可能去掉更多的对象.

定义 3.3　(P 的信息量) 设 $S = (U, A, V, f)$ 为信息系统, $P \subseteq A$,则 P 的信息量定义为:

$$\gamma(P) = \frac{|U/P|}{|U/A|} + \omega\frac{|U_P|}{|U|}, \text{ 其中 } 0 < \omega < \frac{1}{|U/A|}, U_P = \bigcup_{P_i \in U/P \text{且} |P_i/A| = 1} P_i.$$

定理 3.1　设 $S = (U, A, V, f)$ 为信息系统, $P \subseteq A$, $U/P = U/A \Leftrightarrow \gamma(P) \geq 1$.

证明　若 $U/P = U/A$,显然有 $\gamma(P) \geq 1$;当 $\gamma(P) \geq 1$ 时,假设 $U/P \neq U/A$,由于 $P \subseteq A$,由引理 2.1 知 $U/A \subseteq U/P$,即至少存在 $P_i \in U/P$ 使得 $\forall A_j \in U/A$ 有 $P_i \not\subseteq A_j$ (否则有 $U/A \supseteq U/P$,从而有 $U/A = U/P$,这与假设矛盾). 故 $|U/P| \leq |U/A| - 1$,从而有

$$\frac{|U/P|}{|U/A|} \leq \frac{|U/A| - 1}{|U/A|} = 1 - \frac{1}{|U/A|},$$

而 $\gamma(P) = \frac{|U/P|}{|U/A|} + \omega\frac{|U_P|}{|U|} \leq \frac{|U/P|}{|U/A|} + \omega < \frac{|U/P|}{|U/A|} + \frac{1}{|U/A|} \leq 1 - \frac{1}{|U/A|} + \frac{1}{|U/A|} = 1$,与已知条件矛盾,故命题成立.

从 P 的信息量定义公式可知:该公式是保证分类的个数尽可能多,同时尽可能去掉更多的对象.

定理 3.2　设 $S = (U, A, V, f)$ 为信息系统,若 $Q \subseteq P \subseteq A$,则 $\gamma(P) \geq \gamma(Q)$.

证明　由引理 2.1 可知:$U/P \subseteq U/Q$,即 $|U/P| \geq |U/Q|$,另一方面显然有 $|U_P| \geq |U_Q|$,故有 $\gamma(P) \geq \gamma(Q)$.

定义 3.4　设 $S = (U, A, V, f)$ 为信息系统, $P \subseteq A$, $\forall a \in (A - P)$ 的重要性定义为:$\mathrm{sig}_P(a) = \gamma(P \cup \{a\}) - \gamma(P)$.

定理 3.3　设 $S = (U, A, V, f)$ 为信息系统, $P \subseteq A$,若 $\gamma(P) = \gamma(A)$,且 $\forall a \in P$ 有 $\mathrm{sig}_{P-\{a\}}(a) > 0$,则 P 为 A 的一个属性约简.

证明　由定理 3.1 和 3.2 即得.

4　属性重要性的计算方法

在文献[5]中给出的属性重要性的计算复杂度为 $O(|U|^2)$,文献[7]中给出的属性重要性的计算复杂度为 $O(|U| \log |U|)$,文献[8]中的属性重要性的计算复杂度为 $O(n_1 \log n_1 + n_2 \log n_2 + \cdots + n_k \log n_k)$,其中 $n_1 + n_2 + \cdots + n_k \leq |U|$,本文中给出一个属性重要性的计算复杂度为 $O(n_1' + n_2' + \cdots + n_k')$ 的计算方法. 为说明该计算方法,先引入如下的定理.

定理 4.1[7]　设 $S = (U, A, V, f)$ 为信息系统, $P \subseteq A$, $\forall a \in (A - P)$,则有:$U/(P \cup \{a\}) = \bigcup_{X \in U/P} (X/\{a\})$.

这个定理说明 $P \cup \{a\}$ 在 U 上的划分是用 a 对划分 U/P 中的每一块上进一步划分得到,这样使得划分可以用增量式方法进行.

定理 4.2　设 $S = (U, A, V, f)$ 为信息系统, $P \subseteq A$, $\forall a \in (A - P)$,则有:

$$U/(P \cup \{a\}) = \{X \mid X \in U/P \text{且} |X/A| = 1\} \cup \{\bigcup_{X \in U/P \text{且} |X/A| \neq 1} (X/\{a\})\}.$$

证明　由定理 4.1 知:

$$U/(P\cup\{a\}) = \bigcup_{X\in U/P}(X/\{a\})$$
$$=\Big\{\bigcup_{X\in U/P\text{且}|X/A|=1}(X/\{a\})\Big\}\cup\Big\{\bigcup_{X\in U/P\text{且}|X/A|\neq 1}(X/\{a\})\Big\}$$
$$=\{X\mid X\in U/P\text{且}\mid X/A\mid=1\}\cup\Big\{\bigcup_{X\in U/P\text{且}|X/A|\neq 1}(X/\{a\})\Big\}.$$

该定理说明:在 U/P 上的块 X 若满足 $|X/A|=1$,即 $X\in U/A$ 为基本块(U/A 中的块为基本块),基本块在任意属性 a 上不可区分,因而 $X/\{a\}=X$,这时,X 在 $U/P\cup\{a\}$ 上不可改变,从而不可能导致出现新的等价类,故可从 U 中去掉,以减少搜索空间,同时又不会影响 $\gamma(P\cup\{a\})$ 的值.

定理 4.3 设 $S=(U,A,V,f)$ 为信息系统,$P\subseteq A$,$\forall a\in(A-P)$,则有:

$$\gamma(P\cup\{a\})=\gamma(P)+\frac{\Big|\bigcup_{X\in U/P\text{且}|X/A|\neq 1}(X/\{a\})\Big|-|\{X\mid X\in U/P\text{且}\mid X/A\mid\neq 1\}|}{|U/A|}+$$
$$\omega\frac{\Big|\Big\{\bigcup_{X\in U/P\text{且}|X/A|\neq 1}\{\cup y\mid y\in X/\{a\}\text{且}\mid y/A\mid=1\}\Big\}\Big|}{|U|}.$$

证明

$$\mathrm{sig}_P(a)=\gamma(P\cup\{a\})-\gamma(P)=\frac{|U/(P\cup\{a\})|}{|U/A|}+\omega\frac{|U_{P\cup\{a\}}|}{|U|}-\frac{|U/P|}{|U/A|}-\omega\frac{|U_P|}{|U|}$$
$$=\frac{\Big|\bigcup_{X\in U/P\text{且}|X/A|=1}X/\{a\}\Big|+\Big|\bigcup_{X\in U/P\text{且}|X/A|\neq 1}(X/\{a\})\Big|}{|U/A|}-$$
$$\frac{|\{X\mid X\in U/P\text{且}\mid X/A\mid=1\}|+|\{X\mid X\in U/P\text{且}\mid X/A\mid\neq 1\}|}{|U/A|}+$$
$$\omega\frac{\Big|\bigcup_{X\in U/(P\cup\{a\})\text{且}|X/A|=1}X\Big|}{|U|}-\omega\frac{\Big|\bigcup_{X\in U/P\text{且}|X/A|=1}X\Big|}{|U|}$$
$$=\frac{\Big|\bigcup_{X\in U/P\text{且}|X/A|\neq 1}(X/\{a\})\Big|-|\{X\mid X\in U/P\text{且}\mid X/A\mid\neq 1\}|}{|U/A|}+$$
$$\omega\frac{\Big|\Big\{\bigcup_{X\in U/P\text{且}|X/A|=1}\{\cup y\mid y\in X/\{a\}\text{且}\mid y/A\mid=1\}\Big\}\cup\Big\{\bigcup_{X\in U/P\text{且}|X/A|\neq 1}\{\cup y\mid y\in X/\{a\}\text{且}\mid y/A\mid=1\}\Big\}\Big|}{|U|}-$$
$$\omega\frac{\Big|\bigcup_{X\in U/P\text{且}|X/A|=1}X\Big|}{|U|}$$
$$=\frac{\Big|\bigcup_{X\in U/P\text{且}|X/A|\neq 1}(X/\{a\})\Big|-|\{X\mid X\in U/P\text{且}\mid X/A\mid\neq 1\}|}{|U/A|}+$$
$$\omega\frac{\Big|\Big\{\bigcup_{X\in U/P\text{且}|X/A|\neq 1}\{\cup y\mid y\in X/\{a\}\text{且}\mid y/A\mid=1\}\Big\}\Big|}{|U|}.$$

定理 4.3 说明:任意 $a\in(A-P)$ 的重要性可以在 $\gamma(P)$ 的基础计算,即在 U/P 上计算. 它是 U/P 上增加的等价类个数和在从 U/P 中非基本块分离出来的所有基本块的元素的总个数来决定其属性重要性的. 由定理 4.3 可知,计算 $\mathrm{sig}_P(a)$ 只需计算 $X/\{a\}$($X\in U/P$ 且 $|X/A|\neq 1$),即 a 在 U/P 中的非基本块 X 上的划分. 由此得到的等价类增加的个数和分离出来的基本块的所有元素的总个数.

4.1 下面给出信息系统 S=(U,A,V,f) 中计算 X/{a} 的算法

输入:$X(X\in U/P$ 且 $|X/A|\neq 1)$ 和 $a\in A-P$.

输出:由 $X/\{a\}$ 导致增加的等价类个数 S-1 和所有 $X/\{a\}$ 中的基本块的元素 B.

取桶的个数为 $t=|X|$

Calculate(X, a)

{for ($i=1$; $i<|X|+1$; $i++$)

将 $x_i\in X$ **放入编号为** $f(X_i,a)$ **的桶里**;

统计非空桶的个数,记为 S,**则增加的等价类的个数为** S-1;

令 $B = \varnothing$;//B 存放基本块的元素.

for ($i = 1; i < S + 1; i + +$)

判断第 i 个非空桶内的元素是否构成一个基本块, 若是则将其所有元素放入 B;

}

易知,算法 Calculate(X, a)的的复杂度是 $O(|X|)$.

4.2　下面给出信息系统 S = (U, A, V, f) 的属性约简算法

输入:信息系统 $S = (U, A, V, f)$

输出:属性约简

算法 RedBasedSig()

1) 用快速排序方法计算出 U/A.

2) $R = \varnothing$;//R 存放信息系统的属性约简 $\gamma(R) = 0$;

3) 若 $\gamma(R) \geq 1$,则输出 R,否则

4) 对任意 $a \in A - R$ 做如下处理:

a. 令 $h = 0$;//h 存放由 U/R 到 $U/(R \cup \{a\})$ 增加等价类个数.

$B_{\{a\}} = \varnothing$;//$B_{\{a\}}$ 存放由 U/R 到 $U/(R \cup \{a\})$ 增加的所有基本块的元素

b. 对任意 $y \in \{X \mid X \in U/R$ 且 $|X/A| \neq 1\}$ 做如下处理.

调用 Calculate(y, a);

$h = h +$ 由 Calculate(y, a) 计算出的增加的等价类个数.

$B_{\{a\}} = B_{\{a\}} \cup$ 由 Calculate(y, a) 计算出的所有基本块的元素组成的集合.

c. 计算 $\mathrm{sig}_R(a) = \dfrac{h}{|U/A|} + \omega\dfrac{|B_{\{a\}}|}{|U|}$.

5) 记 $\mathrm{sig}_R(a') = \max\limits_{a \in (A - R)} \mathrm{sig}_R(a)$,若这样的属性不只一个时,则任取其一;

$\gamma(R) = \gamma(R) + \mathrm{sig}_R(a')$; $R = R \cup \{a'\}$; $U = U - B'_{\{a\}}$;转步骤 3.

4.3　算法复杂度分析

算法 RedBasedSig() 的第一步时间复杂度为 $O(|A||U|\log|U|)$;算法 RedBasedSig() 的第四步中计算 $\mathrm{sig}_R(a)$ 的时间复杂度,是由 4.2 中 Calculate(y, a) 引起. 由 Calculate(y, a) 时间复杂度分析知,计算 $\mathrm{sig}_R(a)$ 的时间复杂度为: $\sum\limits_{X \in U/R \text{且} |X/A| \neq 1} O(|X|) = O(|\text{砚}_R|)$,其中 $\text{砚}_R = \bigcup\limits_{X \in U/R \text{且} |X/A| \neq 1} X \subseteq U$. 因而,算法 RedBasedSig() 的第四步的时间复杂度为 $O(|A - R||\text{砚}_R|)$. 算法从第 3 步到第 5 步总的时间复杂度为 $O(|A||U|) + O(|A - 1||\text{砚}_{R_1}|) + \cdots + O(|A - R_K||\text{砚}_{R_K}|)$ (R_K 为属性约简),因而算法最坏的时间复杂度为 $O(|A||U|) + O(|A - 1||U|) + \cdots + O(|U|) = O(|A|^2|U|)$,故算法 RedBasedSig() 最坏的时间算法复杂度为 $\max\{O(|A||U|\log|U|), O(|A|^2|U|)\}$.

5　实例分析

以表 1 所示信息系统为例说明新算法的计算过程.

由表 1 可知:

$$U/\{a, b, c, d, e\} = \{\{x_1, x_3\}, \{x_2, x_4\}, \{x_5, x_9\}, \{x_6\}, \{x_7\}, \{x_8, x_{10}\}\};$$

$$U/\{a\} = \{\{x_1, x_3, x_6\}, \{x_2, x_4, x_7\}, \{x_5, x_8, x_9, x_{10}\}\}.$$

故 $\dfrac{|U/\{a\}|}{|U/\{a, b, c, d, e\}|} = \dfrac{3}{6}$, $\dfrac{1}{|U/\{a, b, c, d, e\}|} = \dfrac{1}{6}$,取 $\omega = \dfrac{1}{10} = 0.1 < \dfrac{1}{6}$, $B_{\{a\}} = \varnothing$;

故 $\mathrm{sig}_\varnothing\{a\} = \dfrac{|U/\{a\}|}{|U/\{a, b, c, d, e\}|} + 0.1\dfrac{|B_{\{a\}}|}{|U|} = \dfrac{3}{6} + 0.1\dfrac{0}{10} = 0.5$.

同理可求 $\mathrm{sig}_\varnothing(b)$, $\mathrm{sig}_\varnothing(c)$, $\mathrm{sig}_\varnothing(d)$, $\mathrm{sig}_\varnothing(e)$,如表 3 所示.

表 3 各属性的重要性

	a	b	c	d	e
$\mathrm{sig}_{\varnothing}(\cdot)$	0.5	0.52	0.5	0.53	0.33
$B_{\{\cdot\}}$	Ø	$\{x_8, x_{10}\}$	Ø	$\{x_6, x_8, x_{10}\}$	Ø

由表 3 可得: $\mathrm{sig}_{\varnothing}\{d\} = \max_{y\in\{a,b,c,d,e\}} \mathrm{sig}_{\varnothing}(y)$ **,故有**: $R = \{d\}$; $\gamma(\{d\}) + 0.53$; $U = U - B_{\{d\}} = \{x_1, x_2, x_3, x_4, x_5, x_7, x_9\}$; $\bigcup_{X\in U/\{d\}\text{且}|X/\{a,b,c,d,e\}|\neq 1} X = \{x_1, x_2, x_3, x_4, x_5, x_7, x_9\}$. **由于** $\gamma(\{d\}) = 0.53 < 1$ **,故算法转入下一轮计算. 取** $a \in A - \{d\}$ **,则** $\{x_1, x_2, x_3, x_4, x_5, x_7, x_9\}/\{a\} = \{\{x_1, x_3\}, \{x_2, x_4, x_7\}, \{x_5, x_9\}\}$ **,由此分离出的基本块为**: $\{x_1, x_3\}$, $\{x_5, x_9\}$. **故** $B_{\{a\}} = \{x_1, x_3, x_5, x_9\}$. **由此导致增加的等价类个数为** 2. **故** $\mathrm{sig}_{\{d\}}\{a\} = \frac{2}{6} + 0.1\frac{|B_{\{a\}}|}{10} = \frac{1}{3} + 0.2\frac{4}{10} = 0.33 + 0.04 = 0.37$ **同理可求** $\mathrm{sig}_{\{d\}}(b)$, $\mathrm{sig}_{\{d\}}(c)$, $\mathrm{sig}_{\{d\}}(e)$ **,如表 4 所示.**

表 4 各属性的重要性

	a	b	c	e
$\mathrm{sig}_{\{d\}}(\cdot)$	0.37	0.19	0.36	0.17
$B_{\{\cdot\}}$	$\{x_1, x_3, x_5, x_9\}$	$\{x_1, x_3\}$	$\{x_5, x_7, x_9\}$	Ø

由表 4 可得: $\mathrm{sig}_{\{d\}}\{a\} = \max_{y\in\{a,b,c,e\}} \mathrm{sig}_{\{d\}}(y)$ **,故**: $R = \{d, a\}$, $\gamma(\{d, a\}) = 0.90$, $U = U - B_{\{a\}} = \{x_2, x_4, x_7\}$, $\bigcup_{X\in U/\{a,d\}\text{且}|X/\{a,b,c,d,e\}|\neq 1} X = \{x_2, x_4, x_7\}$. **由于** $\gamma(\{d, a\}) = 0.90 < 1$ **,故算法转入下一轮计算. 取** $b \in A - \{a, d\}$ **,则** $\{x_2, x_4, x_7\}/\{b\} = \{x_2, x_4, x_7\}$ **,由此分离出的基本块为** 0. **由此导致增加的等价类个数为** 0. **故** $\mathrm{sig}_{\{a,d\}}\{b\} = 0$. **同理可求**: $\mathrm{sig}_{\{a,d\}}\{c\} = 0.20$, $\mathrm{sig}_{\{a,d\}}\{e\} = 0.20$. **由于** $\gamma(\{a, d, c\}) = \gamma(\{a, d, e\}) = 0.90 + 0.20 = 1.10 \geq 1$. **故** $\{a, d, c\}$, $\{a, d, e\}$ **为信息系统的约简**.

6 结论

本文提出的属性约简算法 RedBasedSig() **与其它算法相比较有以下优点**: 1) **本算法是目前属性约简算法中时间复杂度最好的算法**; 2) **在本算法中,没有用到求核的算法,实际上求一个信息系统的核是一个很费时的算法**[4]. 3) **在本算法中,随着算法的运行,被搜索空间** U **(对象集) 是以最快速度减少,从而使算法的效率得到提高**. 4) **因为在大多数的属性约简算法中,都要求出** U/A **,如果不记这一步的时间复杂性,与其它一些相关的属性约简算法相比,我们的算法在求属性约简过程中的时间复杂度是最好的**.

参考文献:

[1] Pawlak Z, et al. Rough set [J]. Communication of the ACM, 1995, 38(11): 89 - 95.

[2] Pawlak Z, et al. Rough set theory and its application to data analysis [J]. Cybernetics and Systems, 1998, 9: 661 - 668.

[3] Miao Duoqian, Wang Jue. An information-based algorithm for reduction of knowledge[C]//IEEE ICIPS'97, 1997, 1155 - 1158.

[4] Guan J, Bell D. Rough computational methods for information systems [J]. Artificial Intelligence, 1998, 105: 77 - 103.

[5] **梁吉业, 曲开社, 徐宗本. 信息系统的属性约简**[J]. **系统工程理论与实践**, 2001, 21(12): 76 - 80.
Liang Jiye, Qu Kaishe, Xu Zongben. Reduction of attribute in information systems[J]. Systems Engineering - Theory & Practice, 2001, 21(12): 76 - 80.

[6] **叶东毅. Jelonek 属性约简算法的一个改进**[J]. **电子学报**, 2000, 28(12): 81 - 82.
Ye Dongyi. An improvement to Jelonek's attribution reduction algorithm[J]. Acta Electronica Sinica, 2000, 28(12): 81 - 82.

[7] **刘少辉, 盛秋戬, 等. Rough 集高效算法的研究**[J]. **计算机学报**, 2003, 26(5): 524 - 529.
Liu Shaohui, Sheng Qiujian, et al. Research on efficient algorithm for rough set method[J]. Chinese Journal of Computer, 2003, 26(5): 524 - 529.

[8] **杜金莲, 迟忠先, 翟巍. 基于属性重要性的逐步约简算法**[J]. **小型微型计算机系统**, 2003, 24(6): 976 - 978.
Du Jinglian, Chi Zhongxian, Zhai Wei. An improved algorithm for reduction of knowledge based on significance of attribution[J]. Mini-micro System, 2003, 24(6): 976 - 978.