文章编号: 1001-0920(2005)08-0878-05

一种基于新的条件信息熵的高效知识约简算法

刘启和, 李 凡, 闵 帆, 叶 茂, 杨国纬
(电子科技大学 计算机科学与工程学院, 成都 610054)

摘 要: 分析了在知识约简过程中现有条件信息熵的不足, 给出一种新的条件信息熵, 由此定义新的属性重要性, 将其与基于正区域和基于现有条件信息熵的属性重要性进行比较, 结果表明新的属性重要性是一种更准确、更全面的启发信息. 以新的属性重要性为启发信息设计约简算法, 并给出计算新的条件信息熵的高效算法. 理论分析和实验结果表明, 与基于现有条件信息熵的约简算法相比, 该约简算法时间复杂度较低, 且在搜索最小或次优约简方面更优.

关键词: Rough 集理论; 知识约简; 条件信息熵
中图分类号: TP301 **文献标识码**: A

An Efficient Knowledge Reduction Algorithm Based on New Conditional Information Entropy

LIU Qi-he, LI Fan, MIN Fan, YE Mao, YANG Guo-wei
(College of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China. Correspondent: LIU Qi-he, E-mail: qiheliu@uestc.edu.cn)

Abstract: The disadvantages of the current conditional information entropy are analyzed. A new conditional information entropy is proposed. Based on this entropy the new significance of an attribute is defined and compared with two significances of this attribute based on the positive region and the current conditional information entropy respectively. The result shows that when used as heuristic information, the proposed significance of the attribute is better than the other two. Finally, a heuristic algorithm for knowledge reduction is designed and an efficient algorithm for computing conditional information entropy is proposed. Theoretical analysis and experimental results show that time complexity of this reduction algorithm is less than that of the algorithm based on the current conditional information entropy. Also, this reduction algorithm is more capable of finding the minimal or optimal reducts.

Key words: Rough sets theory; Knowledge reduction; Conditional information entropy

1 引 言

知识约简是Rough 集理论中的核心问题之一. 搜索所有约简或最小约简被证明是一个NP 完全问题[1], 因此一般采用启发式算法搜索最优或次优约简. 人们已提出了基于正区域[2]、基于区分矩阵属性频率[3]和基于条件信息熵[4,5]的约简算法. 然而, 使用现有条件信息熵设计约简算法, 存在一定的不足: 其一是现有的条件信息熵无法等价表示知识约简; 其二是约简算法的时间复杂度较高[4~6].

当前存在两种搜索约简的策略[5]: 1) 以所有条件属性集为起点, 自顶向下逐步去掉不必要的属性; 2) 以核或相对核为起点, 自底向上逐步增加属性. 本文采用第 2 种方法, 分析在核的基础上自底向上逐步增加属性时等价类如何被逐步细化和分离, 并进一步分析了现有条件信息熵的不足. 在此基础上给出一个新的条件信息熵, 由它表示的知识约简与代数表示的知识约简等价, 从而解决了文献[5, 6]提出的信息论表示与代数表示下约简结果不一致的问

收稿日期: 2004-09-20; **修回日期**: 2005-02-23.
基金项目: 国家自然科学基金(天元)项目(A 0324638).
作者简介: 刘启和(1973—), 男, 重庆人, 博士生, 从事Rough 集理论等研究; 杨国纬(1939—), 男, 重庆人, 教授, 博士生导师, 从事人工智能、计算机网络等研究.

题

本文根据新的条件信息熵定义新的属性重要性, 并以它为启发信息设计约简算法 为克服基于现有条件信息熵的约简算法[2, 5]时间复杂度高的缺点, 文中给出计算条件信息熵的高效算法, 降低了约简算法的时间复杂度 理论分析和实验结果表明, 该约简算法时间复杂度低于基于现有条件信息熵的约简算法[5], 即该算法是高效的, 并且该约简算法比现有方法更容易搜索到最小或次优约简

2　Rough 集的基本概念

下面简要介绍 Rough 集理论的相关概念, 其详细定义可参阅文献[1, 5, 7].

在决策表 $S = \langle U, C, D \rangle$ 中, 对于任意 $B \subseteq C \cup D$, 由 B 确定的不可区分关系为 $\mathrm{IND}(B)$, 对象 x 在 $\mathrm{IND}(B)$ 中的等价类为 $[x]_B$. $\mathrm{IND}(B)$ 在 U 上导出的划分记为 $U/\mathrm{IND}(B)$, 简记为 U/B.

定义 1[7]　在决策表 $S = \langle U, C, D \rangle$ 中, $A \subseteq C$, $X \subseteq U$, 用 $\underline{A}(X)$ 或 $\underline{A}X$ 表示 X 的 A - 下近似集, 决策属性 D 的 A - 正区域 $\mathrm{POS}_A(D)$ 定义为

$$\mathrm{POS}_A(D) = \bigcup_{X \in U/D} \underline{A}X, \tag{1}$$

并记

$$r_A(D) = |\mathrm{POS}_A(D)| / |U| \tag{2}$$

定义 2[5] (属性重要性的代数观点定义)　在决策表 $S = \langle U, C, D \rangle$ 中, $A \subseteq C$, 则任意 $a \in C - A$ 的属性重要性为

$$\mathrm{SGF}_1(a, A, D) = r_{A \cup \{a\}}(D) - r_A(D). \tag{3}$$

定义 3[5] (属性重要性的信息论观点定义)　在决策表 $S = \langle U, C, D \rangle$ 中, $A \subseteq C$, 则任意 $a \in C - A$ 的属性重要性为

$$\mathrm{SGF}_2(a, A, D) = H(D|A) - H(D|A \cup \{a\}), \tag{4}$$

其中 $H(D|A)$ 是条件信息熵

3　条件信息熵

本节分析现有条件信息熵的局限性, 给出新的条件信息熵和新的属性重要性, 并将其与基于正区域和基于现有条件信息熵的属性重要性进行比较

在决策表 $S = \langle U, C, D \rangle$ 中, 决策属性集 D 导出的 U 上划分记为 $U/D = \{Y_1, Y_2, \cdots, Y_m\}$.

定义 4　在决策表 S 中, 称 $\mathrm{POS}_C(D)$ 中的元素为相容对象, $U - \mathrm{POS}_C(D)$ 中的元素为矛盾对象

3.1　现有条件信息熵的局限性

为说明现有条件信息熵的局限性, 首先给出如下例子:

例 1　在表 1 的决策表中, $S = \langle U, C, D \rangle$, 其中 $U = \{1, 2, \cdots, 10\}$, $C = \{a, b, c, e, f\}$, $D = \{d\}$.

表 1　一个决策表系统 S

U	a	b	c	e	f	d
1	0	0	0	0	1	0
2	0	1	1	1	0	1
3	1	1	0	1	1	1
4	0	1	1	1	0	0
5	0	0	1	0	1	0
6	1	1	0	1	0	1
7	0	1	1	1	1	1
8	1	1	1	0	1	1
9	1	1	0	1	1	0
10	0	1	1	1	1	0

在决策表 S 中, $U/D = \{\{1, 4, 5, 9, 10\}, \{2, 3, 6, 7, 8\}\}$, $\mathrm{POS}_C(D) = \{1, 5, 6, 8\}$, $U - \mathrm{POS}_C(D) = \{2, 3, 4, 7, 9, 10\}$, $\mathrm{CORE}_D(C) = \{f\}$, 最小约简 (即属性个数最少) 为 $\{a, e, f\}$, $\mathrm{POS}_{\{f\}}(D) = \varnothing$. 则

$U/\{f\} = \{\{1, 3, 5, 7, 8, 9, 10\}, \{2, 4, 6\}\}$,

$U/\{e, f\} = \{\{3, 7, 9, 10\}, \{1, 5, 8\}, \{2, 4, 6\}\}$,

$\mathrm{POS}_{\{e,f\}}(D) = \varnothing$.

一方面, 在核 $\{f\}$ 的基础上增加属性 e 后, $\mathrm{POS}_{\{e,f\}}(D)$ 为空集 根据定义 2 和定义 3, 相对于属性 a, b, c 而言, 属性 e 的重要性最小, 分别为

$\mathrm{SGF}_1(e, \{f\}, D) = 0$, $\mathrm{SGF}_2(e, \{f\}, D) = 0.01$.

进而以 SGF_1 和 SGF_2 为启发信息的约简算法[2, 5], 搜索结果不是最小约简 $\{a, e, f\}$, 而是 $\{a, b, c, f\}$.

另一方面, 在核属性集 $\{f\}$ 上增加属性 e 后, 可使等价类 $\{1, 3, 5, 7, 8, 9, 10\}$ 中所有矛盾对象 $\{3, 7, 9, 10\}$ 与相容对象 $\{1, 5, 8\}$ 分离 将所有矛盾对象从相容对象中分离开来, 有助于搜索最小或次优约简, 因此属性 e 非常重要 这说明 SGF_1 和 SGF_2 没有准确地描述属性 e 的重要性, 其原因在于 SGF_1 和 SGF_2 的定义中没有区分矛盾对象与相容对象这两个重要概念

3.2　新的条件信息熵

本节提出新的条件信息熵, 以克服现有信息熵的局限性

定义 5　在决策表 S 中, 对于任意 $A \subseteq C$, 显然有 $\underline{A}(U - \mathrm{POS}_A(D)) = U - \mathrm{POS}_A(D)$, 因此记 $\underline{A}Y_0 = U - \mathrm{POS}_A(D)$. 特别地, 当 $A = C$ 时, 有 $\underline{C}Y_0 = U - \mathrm{POS}_C(D)$.

定理 1　在决策表 S 中, $A, B \subseteq C$, 则 $\mathrm{POS}_A(D) = \mathrm{POS}_B(D)$ 的充分必要条件是

$$\underline{A}Y_i = \underline{B}Y_i, \quad \forall i \in \{1, 2, \cdots, m\}. \tag{5}$$

如果集簇 $\{\underline{A}Y_0, \underline{A}Y_1, \cdots, \underline{A}Y_m\}$ 中没有空集, 则该集簇是 U 上的一个划分; 如果该集簇中有空集, 则去掉空集后仍是 U 上的一个划分. 若没有特别说明, 不妨假设该集簇中没有空集

由定理 1 可直接得到如下推论:

推论 1 在决策表 S 中,$A \subseteq C$, 则 $POS_A(D) = POS_C(D)$ 的充分必要条件是

$$\underline{A}Y_i = \underline{B}Y_i, \quad \forall i \in \{0, 1, \cdots, m\}. \tag{6}$$

由推论 1 知, 如果属性集 A 是约简的, 则由 A 导出的划分不仅把属于不同决策类的相容对象分离成不同的划分块, 而且把所有矛盾对象从相容对象中分离出来, 作为一个单独的划分块. 然而, 在以核为起点, 自底向上逐步增加属性的约简过程中, 如果属性集 A 不是约简的, 那么应如何对 A 逐步增加属性使其变为约简的?由于 A 不是约简的, 说明由 A 决定的基本知识的粒度较粗, 以致于相容对象与矛盾对象在 A 的同一等价类中, 或属于不同决策类的相容对象在 A 的同一等价类中.

在例 1 中令 $A = \{f\}$, 则相容对象 6 与矛盾对象 2 和 4 在 A 的等价类 $\{2, 4, 6\}$ 中; 属于不同决策类的相容对象 5 和 8 在 A 的等价类 $\{1, 3, 5, 7, 8, 9, 10\}$ 中. 因此, 增加属性的目的有两个: 一是使 A 的等价类中的相容对象与矛盾对象分离; 二是使 A 的等价类中属于不同决策类的相容对象分离.

$SGF_1(a, A, D)$ (定义 2) 只对正区域基数进行定量描述; $SGF_2(a, A, D)$ (定义 3) 只描述了 A 的等价类中属于不同决策类的对象分离情况, 因为它没有考虑相容对象和矛盾对象, 所以无法描述 A 的等价类中决策属性值相同的相容对象与矛盾对象的分离. 正因为如此, 在不一致决策表中, 由于矛盾对象的存在, 使用条件熵 $H(D|A)$ 无法等价表示知识约简. 划分 $\{\underline{C}Y_0, \underline{C}Y_1, \cdots, \underline{C}Y_m\}$, 既将属于不同决策类的相容对象分开, 又将相容对象与矛盾对象分开, 因此在其上定义条件信息熵, 可以克服现有条件信息熵的上述不足.

定义 6 在决策表 S 中,$A \subseteq C$, 则集簇 $\{\underline{A}Y_0, \underline{A}Y_1, \cdots, \underline{A}Y_m\}$ 是 U 上的一个划分, 由此划分得到的等价关系记为 R_A. 特别地, R_C 表示由划分 $\{\underline{C}Y_0, \underline{C}Y_1, \cdots, \underline{C}Y_m\}$ 得到的等价关系.

R_C 是根据 S 中的所有数据计算出来的等价关系, 但 Rough 集理论一般根据属性集来获得对应的等价关系. 为方便起见, R_C 可以看作一个属性, 由这个属性导出的等价关系便是 R_C.

定义 7 (新的条件信息熵) R_A 在 U 上的子集组成的 σ 代数上的概率分布定义为

$$[R_A: p] = \begin{bmatrix} \underline{A}Y_0 & \underline{A}Y_1 & \cdots & \underline{A}Y_m \\ p(\underline{A}Y_0) & p(\underline{A}Y_1) & \cdots & p(\underline{A}Y_m) \end{bmatrix}. \tag{7}$$

其中

$$p(\underline{A}Y_i) = |\underline{A}Y_i| / |U|, \quad i = 0, 1, \cdots, m.$$

特别地, 如果 $A = C$, 则对任意的 $P \subseteq C$, 设 $U/P = \{X_1, X_2, \cdots, X_n\}$, 可得新的条件信息熵

$$H(R_C|P) = -\sum_{i=1}^{n} p(X_i) \sum_{j=0}^{m} p(\underline{C}Y_j|X_i) \log(p(\underline{C}Y_j|X_i)). \tag{8}$$

其中

$$p(\underline{C}Y_j|X_i) = |\underline{C}Y_j \cap X_i| / |X_i|,$$
$$i = 1, 2, \cdots, n, \quad j = 0, 1, \cdots, m.$$

因为属性集 A 的确定划分比 R_A 的确定划分更精细, 所以有 $H(R_A|A) = 0$. 特别地, 有 $H(R_C|C) = 0$.

3.3 属性重要性比较

定义 8 (新的属性重要性定义) 设在决策表 $S = \langle U, C, D \rangle$ 中,$A \subseteq C$, 则任意 $a \in C - A$ 的属性重要性为

$$SGF_3(a, A, D) = H(R_C|A) - H(R_C|A \cup \{a\}). \tag{9}$$

在一致决策表 $S = \langle U, C, D \rangle$ 中, $CY_0 = \varnothing$, $R_C = U/D$, 所以易证如下定理:

定理 2 在一致决策表 $S = \langle U, C, D \rangle$ 中,$A \subseteq C$, 任意 $a \in C - A$, 则有

$$SGF_3(a, A, D) = SGF_2(a, A, D). \tag{10}$$

根据定义 3 和定义 8, $SGF_2(a, A, D)$ 用条件熵 $H(D|A)$ 来定义, 而 $SGF_3(a, A, D)$ 用 $H(R_C|A)$ 来定义. 二者的差异主要体现在决策划分 U/D 与 R_C 上. 在一致决策表中, $R_C = U/D$, 所以二者相等; 在不一致决策表中, R_C 将 U/D 中所有矛盾对象作为一个独立的块 $\underline{C}Y_0$, 这是 R_C 与 U/D 的不同点. 因为如此, $SGF_3(a, A, D)$ 能够描述约简过程中矛盾对象与相容对象的分离, 这是 $SGF_3(a, A, D)$ 不同于 $SGF_2(a, A, D)$ 之处.

根据文献[5]中引理 1, 可得如下定理:

定理 3 设在决策表 $S = \langle U, C, D \rangle$ 中,$A \subseteq C$, $a \in C - A$, 如果 $SGF_3(a, A, D) = 0$, 则 $SGF_1(a, A, D) = 0$.

由定义 2, $SGF_1(a, A, D)$ 只对正区域基数进行定量描述. 定理 3 说明, 如果增加属性后正区域基数变大, 即 $SGF_1(a, A, D) \neq 0$, 则这种信息必然会在 $SGF_3(a, A, D)$ 上有所反映. 此时 $SGF_3(a, A, D) \neq 0$, 说明 $SGF_3(a, A, D)$ 包含了比 $SGF_1(a, A, D)$ 更多的信息.

由以上讨论可以看出 $SGF_3(a, A, D)$ 有以下优点:

1) $SGF_3(a, A, D)$ 定义在 R_C 上, R_C 已区分开矛盾对象与相容对象. 因此, 如果增加属性 a 后能够分离矛盾对象与相容对象, $SGF_3(a, A, D)$ 就能准确地

描述出来. 如例 1, 在核$\{f\}$的基础上增加属性e后, 尽管正区域没有增加, 但矛盾对象 3, 7, 9, 10 被分离出来, 因此 $SGF_3(e, \{f\}, D) = 0.69$, $SGF_1(e, \{f\}, D) = 0$. 同理, 由定义 3, $SGF_2(e, \{f\}, D)$ 只描述了各决策类的分离情况, 而无法描述矛盾对象与相容对象的分离, 所以 $SGF_2(e, \{f\}, D) = 0.01$.

2) 由于 $SGF_3(a, A, D)$ 能更准确全面地描述信息, 以它为启发信息的约简算法更有可能获得最小或次优约简. 如例 1, 以 $SGF_3(a, A, D)$ 为启发信息的约简算法, 会获得例 1 中决策表S的最小约简$\{a, e, f\}$, 而以 $SGF_1(a, A, D)$ 和 $SGF_2(a, A, D)$ 为启发信息的约简算法, 均获得约简$\{a, b, c, f\}$(参见后面表 3).

3) 使用 $SGF_3(a, A, D)$ 和定义 7 可以等价表示知识约简的代数定义, 这保证了以 $SGF_3(a, A, D)$ 为启发信息的约简算法, 所获得的约简一定是代数表示下的约简.

表 2 给出了例 1 中条件属性相对于核$\{f\}$的属性重要性.

表 2　例 1 中条件属性相对于核$\{f\}$的属性重要性

属性	SGF_1	SGF_2	SGF_3
a	0.1	0.17	0.57
b	0.2	0.20	0.60
c	0.1	0.09	0.37
e	0.0	0.01	0.69

由此可见, $SGF_3(e, \{f\}, D)$ 更能准确全面地描述属性e的重要性, 说明新的条件信息熵克服了现有条件信息熵的不足.

4　知识约简算法

以 $SGF_3(a, A, D)$ 为启发信息的约简算法, 必须计算条件熵 $H(R_C|A \cup \{a\})$. 为降低算法的时间复杂度, 应研究计算 $H(R_C|A)$ 的高效算法. R_C 是决策表$S = \langle U, C, D\rangle$中论域$U$上的等价关系, 由文献[5]可得到计算 $H(R_C|A \cup \{a\})$ 的高效算法如下:

算法 1

输入: 决策表$S = \langle U, C, D\rangle$, R_C 和划分U/A;

输出: $H(R_C|A \cup \{a\})$ 和划分$U/A \cup \{a\}$.

Step 1: 计算划分$U/A \cup \{a\}$;

Step 2: 计算划分$U/\{R_C\} \cup A \cup \{a\}$;

Step 3: 计算$H(\{R_C\} \cup A \cup \{a\})$;

Step 4: 计算$H(A \cup \{a\})$, 获得

$$H(R_C|A \cup \{a\}) = H(\{R_C\} \cup A \cup \{a\}) - H(A \cup \{a\}). \quad (11)$$

利用文献[2]中的方法计算划分, 在已知划分U/A的情况下, Step 1 和 Step 2 的时间复杂度均为 $O(|U|\log|U|)$, Step 3 和 Step 4 的时间复杂度均为 $O(|U|)$, 则该算法的时间复杂度为 $O(|U|\log|U|)$, 与文献[2]中正区域渐增式计算的时间复杂度相同.

在以 $SGF_3(a, A, D)$ 为启发信息的约简算法中, 每次循环时条件属性子集A均不变, 这使得 $SGF_3(a, A, D)$ 最大的a就是$H(R_C|A \cup \{a\})$最小的a. 因此只需计算$H(R_C|A \cup \{a\})$, 这样可避免计算 $H(R_C|A)$, 减少了计算量. 由此在算法 1 的基础上, 设计了知识约简算法如下:

算法 2

输入: 决策表$S = \langle U, C, D\rangle$, 其中$U$为论域, C和D分别为条件属性集和决策属性集;

输出: 决策表S的一个相对约简A.

Step 1: 计算决策表S中条件属性集C相对于决策集D的正区域 $POS_C(D)$, 并由此得到R_C.

Step 2: 计算条件属性集C相对于决策属性集D的核属性集$CORE_D(C)$, 并令$B = C - CORE_D(C)$.

Step 3: 令$A = CORE_D(C)$:

1) 如果$|A| != 0$, 则计算条件熵$H(R_C|A)$, 转 4);

2) 对每个属性$a \in B$, 计算R_C相对条件属性集$A \cup \{a\}$的条件熵$H(R_C|A \cup \{a\})$;

3) 选择使$H(R_C|A \cup \{a\})$最小的属性a, 把a从B中删除, 并把a增加到A的尾部;

4) 如果$H(R_C|A) = 0$, 则转 Step 4, 否则转 2).

Step 4: 从A的尾部开始, 从后向前判别每个属性a是否可约: 如果$a \in CORE_D(C)$, 则从a开始向前的属性都是核中属性, 不可约, 算法终止; 否则, 如果$H(R_C|A - \{a\}) = 0$, 则a是可约简的, 把a从A中删除.

算法 2 以 $SGF_3(a, A, D)$ 为启发信息, Step 1 只需计算正区域 $POS_C(D)$ 一次, 并获得 R_C; 在 Step 3 和 Step 4 中, 只计算条件熵 $H(R_C|A \cup \{a\})$, Step 4 保证算法 2 是完备的, 即得到的结果不能再约简. 文献[5]的约简算法以 $SGF_2(a, A, D)$ 为启发信息, 该约简算法是不完备的[2], 这是因为现有条件熵无法等价表示约简, 所以不能从信息熵的角度给出一个完备的约简算法; 文献[2]提出一种基于正区域的完备约简算法, 它以 $SGF_1(a, A, D)$ 为启发信息.

使用文献[2]的计算正区域和核方法, Step 1 的时间复杂度为 $O((|C|+1)|U|\log|U|)$, Step 2 的时间复杂度为 $O((|C|+1)^2|U|\log|U|)$; 由算法 1, Step 3 为 $O((|C|)^2|U|\log|U|)$, Step 4 为 $O((|C|)^2|U|\log|U|)$. 所以算法 2 的时间复杂度为 $O((|C|)^2|U|\log|U|)$, 低于文献[5]中基于现有条

表 3 约简结果与执行时间比较

决策表	是否为一致决策表	实例数	约简前条件属性数	最小约简属性数	算法A		算法B		算法 2	
					约简后条件属性数	执行时间/s	约简后条件属性数	执行时间/s	约简后条件属性数	执行时间/s
例 1	否	10	5	3	4	0.06	4	0.05	3	0.04
voting-records	是	435	16	9	9	0.54	9	0.26	9	0.29
tic-tac-toe	是	958	9	8	8	1.42	8	0.48	8	0.66
zoo	否	101	17	10	11	0.33	10	0.14	10	0.15
mushroom	是	8 124	22	4	4	17.67	5	6.42	4	6.67
chess end-game	是	3 196	36	29	29	24.75	29	4.67	29	5.60

件信息熵的约简算法的时间复杂度, 与文献[2]中基于正区域的约简算法时间复杂度相等.

5 实验结果与分析

本文选用例 1 和UCI机器学习数据库[8]中的部分决策表, 在 PC 机(Celeron 1.7 G, 256 M RAM, WINXP)上进行实验, 分别采用文献[5]的CEBRKCC 算法(简称算法A)、文献[2]的约简算法(简称算法B)和本文算法(算法 2)进行知识约简. 3 种算法都使用文献[2]的方法计算划分和核, 并使用 Java 语言实现. 其运行结果如表 3 所示.

从表 3 可以看出, 算法 2 的执行时间低于算法 A 的执行时间, 与算法B 的执行时间大致相同. 算法 2 约简后, 条件属性数小于或等于算法A 和算法 B 约简后的条件属性数, 并能搜索到最小约简. 实验结果与前面的分析相符合, 验证了算法 2 的高效性, 表明在搜索最小或次优约简方面, 算法 2 比算法A 和算法B 更优. 算法 2 之所以更容易找到最小约简, 是因为在不一致决策表中, $SGF_3(a,A,D)$ 描述了在约简过程中自底向上逐步增加属性时, 等价类中矛盾对象与相容对象的分离.

6 结 论

本文首先分析了现有条件信息熵的不足, 在此基础上给出一种新的条件信息熵, 它可等价地表示知识约简; 然后定义新的属性重要性, 并分析了新的属性重要性与其他两种属性重要性(定义 2 和定义 3)的联系和差异, 理论分析和实例显示, 新的属性重要性能较好地描述当逐步增加属性时, 等价类中矛盾对象与相容对象的分离, 是知识约简算法中理想的启发信息; 最后以新的属性重要性为启发信息设计约简算法, 并给出计算条件熵的高效算法, 该约简算法的时间复杂度低于基于现有条件信息熵的约简算法, 与基于正区域的高效约简算法的时间复杂度相同. 理论分析和实验结果均说明该约简算法不仅是高效的, 而且在搜索最小或次优约简方面, 比基于现有条件信息熵和正区域的约简算法更优.

参考文献(References)

[1] Pawlak Z, Grzymala B J, Slowinski R, et al. Rough Sets[J]. *Communication of the ACM*, 1995, 38(11): 89-95.

[2] 刘少辉, 盛秋戬, 吴斌, 等. Rough 集高效算法研究[J]. *计算机学报*, 2003, 26(5): 524-529.
(Liu S H, Sheng Q J, Wu B, et al. Research on Efficient Algorithms for Rough Set Methods[J]. *Chinese J of Computer*, 2003, 26(5): 524-529.)

[3] Wang J, Wang J. Reduction Algorithms Based on Discernibility Matrix: The Ordered Attributes Method[J]. *J of Computer Science and Technology*, 2001, 16(6): 489-504.

[4] 苗夺谦, 王珏. 粗糙集理论中概念与运算的信息表示[J]. *软件学报*, 1999, 10(2): 113-116.
(Miao D X, Wang J. An Information Representation of the Concept and Operations in Rough Set Theory[J]. *J of Software*, 1999, 10(2): 113-116.)

[5] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. *计算机学报*, 2002, 25(7): 759-766.
(Wang G Y, Yu H, Yang D C. Decision Table Reduction Based on Conditional Information Entropy[J]. *Chinese J of Computer*, 2002, 25(7): 759-766.)

[6] Wang G Y. Algebra View and Information View of Rough Sets Theory[A]. *Proc of SPIE*[C]. Orlando, 2001: 200-207.

[7] 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.
(Zhang W X, Wu W Z, Liang J Y, et al. *Rough Set Theory and Methods*[M]. Beijing: Science Press, 2001.)

[8] Blake C L, Merz C J. *UCI Repository of Machine Learning Databases*[R]. Irvine: University of California at Irvine, 1998.