优势信息系统的属性约简算法

李学文,王小刚
(北方民族大学,宁夏 银川 750021)

摘 要:粗糙集理论是一种新的软计算方法,已成为知识发现和属性约简算法领域的一个研究热点。但经典的粗糙集理论是建立在等价关系基础之上的。从粗糙集理论出发,在优势关系之上给出了优势信息系统中信息量的概念,建立了知识粒度与信息量之间的关系。在此基础上,提出了一种基于信息量的属性约简启发式算法,得到了该算法的时间复杂性为 $O(|A|^3\times|U|^2)$。通过实例分析表明该算法是有效的,为进一步研究约简算法提供了一种可行的方法。

关键词:粗糙集;优势关系;信息量;属性约简
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2009)08-0107-03

Algorithm on Attribute Reduction in Dominance Information System Based on Dominance Relation

LI Xue-wen, WANG Xiao-gang
(North University for Ethics, Yinchuan 750021, China)

Abstract: Rough set theory is a relatively new soft computing technique and has become a topic of general interest in the field of knowledge discovery and algorithm of attribute reduction. The classical rough set theory is based on relation of equivalence. In this paper, the concept of information quantity is given in dominance information system from rough set theory. Relationships between knowledge and information quantity have been built. Accordingly, a new heuristic reduction algorithm based on information quantity, of which time complexity is $O(|A|^3\times|U|^2)$, is proposed. It proves that the algorithm is effective through example analysis, moreover, puts forward new methods for further researching of attribute reduction.

Key words: rough set; dominance relation; information quantity; attribute reduction

0 引 言

粗糙集理论是波兰数学家 Z. Pawlak 于 1982 年提出的一种分析数据的数学理论。该理论是处理不精确和不确定知识的有效工具,近年来已被广泛应用到机器学习、数据挖掘、故障诊断等领域。

属性约简是粗糙集理论的核心问题之一。信息系统中描述知识的属性并不是同等重要的,有些甚至是冗余的,属性约简就是在保持信息系统的分类能力不变的条件下,删除其中的冗余属性。许多学者对属性约简作了深入研究,并取得了丰富的成果[1~6]。这些研究大多是以等价关系为基础的。而现实问题中经常会出现一些有序问题,这就需要研究信息系统上的优势关系。因此, Greco, Matarazzo 和 Slowinski[7]提出了基于优势关系的粗糙集方法(DRSA)。文献[8]从辨识矩阵的角度出发,给出了信息系统的优势关系下的两种属性约简方法。

在优势信息系统中,给出了知识的信息量概念,并建立了知识与信息量之间的关系。在此基础上,提出了一种基于信息量的属性约简启发式算法,该算法的时间复杂性为 $O(|A|^3\times|U|^2)$。通过例子分析表明该算法是有效的。

1 信息系统中的优势关系

定义 1 设 $S=\{U,A,F\}$ 为一个信息系统,其中 U 为对象集,即 $U=\{x_1,x_2,\cdots,x_n\}$,U 中的每个 $x_i(i\leqslant n)$ 称为一个对象。A 为属性集,即 $A=\{a_1,a_2,\cdots,a_n\}$,A 中的每个 $a_i(i\leqslant m)$ 称为一个属性。F 为 U 和 A 的关系集,即 $F=\{f_i,j\leqslant m\}$,其中 $f_j:U\to V_j(j\leqslant m)$,$V_j$ 为属性 a_j 的值域。

$$R_B^{\geqslant}=\{\{x_i,x_j\}:f_1(x_i)\geqslant f_2(x_j),\forall a_i\in B\}$$

称为信息系统上的优势关系。若 $(x,y)\in R_B^{\geqslant}$,则称 x 关于属性集 B 对 y 有优势。

收稿日期:2008-12-05;修回日期:2009-02-04
基金项目:教育部科学技术研究重点项目(206159);宁夏自然科学基金资助项目(NZ0516);北方民族大学校级科研项目(2006Y048)
作者简介:李学文(1980-),女,硕士,研究方向为粗糙集理论。

记$[x_i]_B = \{x_j:(x_i,x_j) \in R_B^{\geq}$。易见优势关系具有以下性质：

(1) $R_B^{\geq}$ 是自反和传递的，不是对称的，所以不再是等价关系；

(2) 当 $B_1 \subseteq B_2 \subseteq A$，有 $R_{B_1}^{\geq} \supseteq R_{B_2}^{\geq} \supseteq R_A^{\geq}$；

(3) 当 $B_1 \subseteq B_2 \subseteq A$ 时，有 $[x_i]_{B_1} \supseteq [x_i]_{B_2} \supseteq [x_i]_A$；

(4) $J_B = \{[x_i]_B : x_i \in U\}$ 是对象集 U 上的覆盖。

设 $B,C \subseteq A$，记

$J_B = J_C$ 表示 $[x_i]_B = [x_i]_C, \forall x_i \in U$。

$J_B \subseteq J_C$ 表示 $[x_i]_B \subseteq [x_i]_C, \forall x_i \in U$。

便于行文方便，以下所指信息系统均指基于优势关系的信息系统。

定义 2 设 $S = \{U,A,F\}$ 为一个信息系统 $\forall B \subseteq A, b \in B$，若 $J_B = J_{B-\{b\}}$，则称 b 在 B 中是可省略的，否则称 b 在 B 中是必要的。若 B 中所有元素均为必要的，则 B 为独立的，否则 B 为依赖的。

定义 3 设 $S = \{U,A,F\}$ 为一个信息系统，$\forall B \subseteq A$，若 B 是独立的，且 $J_B = J_A$，则 B 为信息系统的约简。

2 知识的信息量

在粗糙集理论中，知识被视为对对象进行分类的能力，即关于论域的划分，从而认为知识是有粒度的，即知识粒度越大，知识含量越小。知识和信息量的关系在等价关系下已经被讨论[9]。在基于优势关系的信息系统中，知识被视为关于论域的覆盖，所以很难再用经典的信息量来解释知识。

下面在基于优势关系的信息系统中介绍知识的信息量概念，并建立知识与信息量间的关系。

定义 4 设 $S = \{U,A,F\}$ 为一个信息系统，$\forall B \subseteq A, J_B = \{X_1,X_2,\cdots,X_{|U|}\}$，知识(属性集)$B$ 的信息量定义为：

$$I(B) = |U| - \sum_{i=1}^{|U|} P^2(X_i)$$

其中 $|U|$ 表示集合 X 的基数，$P(X_i) = |X_i| / |U|$ 表示 X_i 在 U 中的概率。

定理 1 设 $S = \{U,A,F\}$ 为一个信息系统，$B,C \subseteq A$，若 $J_B \subseteq J_C$，则 $I_B \geq I_C$。

证明：由 $J_B \subseteq J_C$ 得，$[x_i]_B \subseteq [x_i]_C, \forall x_i \in U$，即 $|[x_i]_B| \leq |[x_i]_C|, \forall x_i \in U$，所以

$$I(B) = |U| - \sum_{i=1}^{|U|} (|[x_i]_B| / |U|)^2 \geq |U| - \sum_{i=1}^{|U|} (|[x_i]_C| / |U|)^2 = I(C)$$

该定理表明，知识的信息量随着知识粒度的变小而单调增加。

定理 2 设 $S = \{U,A,F\}$ 为一个信息系统，对于任意的 $B,C \subseteq A$，若 $J_B \subseteq J_C$ 且 $I(B) = I(C)$，则 $J(B) = J(C)$。

证明：由 $I(B) = I(C)$ 得，

$$|U| - \sum_{i=1}^{|U|} (|[x_i]_B| / |U|)^2 = |U| - \sum_{i=1}^{|U|} (|[x_i]_C| / |U|)^2$$

即 $|[x_i]_B| = |[x_i]_C|$，又由 $J_B \subseteq J_C$，得 $[x_i]_B \subseteq [x_i]_C$，即 $1 \leq |[x_i]_B| \leq |[x_i]_C|$，所以

$$\sum_{i=1}^{|U|} (|[x_i]_B| / |U|)^2 \leqslant \sum_{i=1}^{|U|} (|[x_i]_C| / |U|)^2$$

又由 $[x_i]_B \subseteq [x_i]_C$，可得 $|[x_i]_B| = |[x_i]_C, \forall x_i \in U|$，所以 $J(B) = J(C)$。

该定理表明，如果两个信息系统中存在包含关系，且它们关于知识的信息量相等，则这两个信息系统是等价的。

3 基于信息量的属性约简算法

由知识与信息量之间的关系可以得到下面的定理。

定理 3 设 $S = \{U,A,F\}$ 为一个信息系统，若存在 $B \subseteq A$，满足：

(1) $I(B) = I(A)$

(2) $\forall b \in B, I(B - \{b\}) \neq I(B)$

则 B 为信息系统的一个约简。

定理 3 从信息角度提供了属性约简的依据，这是 IQBARK 算法的基础。该算法以由 A 中去掉 a 后的信息量的大小作为属性 a 对于属性集 A 的参考重要度，$I(A - \{a\})$ 的值越大，属性 a 对于属性集 A 的参考重要度越小，算法的起点是初始属性集 A，采用逐步删除属性来达到约简的目的，它不需要计算属性的核。

属性约简算法(IQBARK)

输入：一个信息系统 $S = \{U,A,F\}$，其中 U 为对象集，A 为属性集。

输出：该信息系统的约简。

步骤 1：计算该信息系统的信息量 $I(A)$；

步骤 2：计算每个 $a_i \in A$ 的知识量 $I(a_i)$，将 a_i 按 $I(a_i)$ 升序排列；

步骤 3：令 $B = A$，按 $I(a_i)$ 递增的顺序对每个 a_i 重复下述操作：

(1) 计算从 A 中去掉 a_i 后的信息量 $I(A - \{a_i\})$；

(2) 如果 $I(A) = I(A - \{a\})$，则 a_i 应约简，$B =$

$B-\{a_i\}$;否则 a_i 不能被约简,B 不变。

最后得到的 B 即为信息系统的约简。

定理 4 IQBARK 算法的时间复杂度是 $O(|A|^3\times|U|^2)$。

证明:

(1) 计算约简需要计算一次 $I(A)$。

(2) 最坏情况下需要计算 $|A|$ 次 $I(a_i)$ 和 $|A|$ 次 $I(A-\{a_i\})$。

(3) 为了计算 $I(A)$ (计算 $I(A)$ 和计算 $I(A-\{a_i\})$ 的时间复杂性相同),需要求下列计算:

① 计算 $|A|$ 个覆盖,所需时间复杂性为 $O(|A|\times|U|^2)$;

② 为了计算 $I(A)$ 和 $I(A-\{a_i\})$,需计算 $|A|$ 和 $|A|-1$ 次交运算,所需时间复杂性为$(|A|+|A|-1)\times O(|U|^2)$。因此,计算一次 $I(A)$ 的时间复杂性为 $O(|A|^2\times|U|^2)$。

所以,整个算法的时间复杂性为:$(|A|+|A|-1)\times O(|A|^2\times|U|^2)=O(|A|^3\times|U|^2)$

例 1　表 1 给出了一个信息系统。

表 1　信息系统

U	a_1	a_2	a_3
x_1	1	2	1
x_2	3	2	2
x_3	1	1	2
x_4	2	1	3
x_5	3	3	2
x_6	3	2	3

解:下面使用所给的 IQBARK 算法求解约简。

步骤 1: 计算 $I(A)$。

由于 $J_A=\{\{x_1\},\{x_1,x_2,x_3\},\{x_3\},\{x_3,x_4\},\{x_1,x_2,x_3,x_5\},\{x_1,x_2,x_3,x_4,x_6\}\}$,所以 $I(A)=40/9$。

步骤 2:计算 $I(a_i)$。

$I(a_1)=91/36, I(a_2)=97/36, I(a_3)=95/36$,所以,$I(a_1)\leqslant I(a_2)\leqslant I(a_3)$,把 a_1 选入;

步骤 3:令 $B=A$,计算 $I(B-\{a_1\})$,由于 $I(B-\{a_1\})=I(\{a_2,a_3\})=I(A)$,所以 a_1 约简,令 $B=B-\{a_1\}=\{a_2,a_3\}$;而对其余的 a_i,$I(B-\{a_1\})\neq I(A)$,所以不能约简。因此,最后约简为$\{a_2,a_3\}$。

4　结束语

从粗糙集理论出发,研究了优势信息系统的属性约简算法,给出了优势关系下信息量的概念,提出了知识粒度与信息量之间的关系,即信息量随着知识粒度的变小而单调增加,并对其进行了量化分析。在此基础上,提出了基于信息量的属性约简算法,为进一步研究约简算法提供了理论依据。

参考文献:

[1] Miao Duoqian, Hu Guirong. A heuristic algorithmfor reduction of knowledge[J]. Journal of Computer Research and Development, 1999,36(6):681-684.

[2] 张文修, 梁　怡, 吴伟志. 信息系统与知识发现[M]. 北京:科学出版社,2003.

[3] 李　鸿. 一种基于粗糙熵的知识约简算法[J]. 计算机工程与应用, 2005,36(14):78-80.

[4] 李红梅,周桂红,王克俭. 基于粗糙集和遗传算法的知识发现方法[J]. 计算机技术与发展,2007,17(8):76-79.

[5] 纪　滨. 粗糙集理论及进展的研究[J]. 计算机技术与发展,2007,17(3):69-72.

[6] Miao Duoqian, Wang Jue. An information-based algorithm for reduction of knowledge[R]. Beijing: Inst. of Autom., Acad. Sinica, 1997:1155-1158.

[7] Matarazzo G S, Slowinski R. Rough approximatiion of a preference relation by dominance relations[J]. European Journal of Operational Research, 1999,117:63-83.

[8] 贾戎莉. 信息系统上的优势关系与保序关系[J]. 山西师范大学学报,2005,19(2):14-16.

[9] Liang Jiye, Li Deyu. Information Measures of Roughness of Knowledge and Significance ofAttibute in Rough Set Theory [J]. Journal of Engineering Mathematics, 2005, 17: 106-108.

(上接第 106 页)

[5] Kennedy J, Eberhart R. Particle swarm optimization[C]// Proc IEEE Int Conf on Neural Networks. Perth: [s. n.], 1995:1942-1948.

[6] 卜艳萍,俞金寿. 离散微粒群优化算法在网格任务中的应用[J]. 计算机仿真,2008,25(4):175-178.

[7] 王雅琳,王　宁,阳春华,等. 求解任务分配问题的一种离散微粒群算法[J]. 中南大学学报,2008,29(3):571-576.

[8] 马　驰,阮秋琦. 基于离散微粒群优化算法的 SVM 参数选择[J]. 计算机技术与发展,2007,17(12):20-23.

[9] 李　钧,王忠群,刘　涛. 基于遗传编程的网格资源调度算法[J]. 计算机技术与发展,2008,18(2):129-132.