文章编号:1673-9469(2008)03-0110-03

基于最大外权重的一种启发式属性约简算法

刘志民
(河北工程大学 理学院,河北 邯郸 056038)

摘要:作为属性对区分元素类别所做贡献大小的度量,定义属性重要性权重,并作为启发性知识构造一种寻求约简的启发式算法。算法复杂度是多项式的,收敛速度快,可得到 Pawlak 约简。

关键词:粗糙集;内权重;外权重;约简
中图分类号:O159 **文献标识码**:A

Heuristic attribute reduction algorithm on biggest outer weight

LIU Zhi-min
(College of Science, Hebei University of Engineering, Handan 056038, China)

Abstract: In this paper, attribute significance weight is defined which measures the contributions of attributes to distinguishing the different element categories. a heuristic algorithm which aims to search reduction is obtained by using the weight as heuristic knowledge. The complexity of the algorithm is polynomial, the convergence is fast and the result is Pawlak reduction.

Key words: rough set; inner weight; outer weight; reduction

粗集[1]是 Z. Pawlak 教授在 1982 年提出的一种不确定性集合,粗集理论是处理不确定、不完整和不一致数据的重要数学方法,是实现数据压缩、知识发现和机器学习的重要工具。属性约简是建立在粗集理论上的约简理论,它可以去掉冗余条件属性而保证分类能力不变。数据分析法是求属性约简的强有力方法[2,3],但数据分析法对属性选择的次数随着条件属性个数的增加,以指数量级剧增,即为 NP 问题。为此,人们不得不从各种角度构造启发式搜索算法,以降低约简的计算复杂度[4,5]。针对决策信息表达系统,文献[6],[7]给出两种不同的属性权重定义。本文依据各条件属性对区分元素类别所做贡献的大小,定义信息表达系统中属性的一种重要性权重,并以此为启发式信息,给出一种求约简的启发式算法。

1 属性的重要性权重

设 $S=(U,A,V,f)$为信息表达系统,其中,U 为论域;A 为属性集;$V=\bigcup_{a\in A}V_a$,V_a 为属性 a 的值域;f 为映射:任意 $x\in U$,$a\in A$,通过 f,x 关于属性 a 的值为 $a(x)\in V_a$。若 $A=C\cup D$,其中 C 为条件属性集,D 为决策属性集,则称 $T=(U,C\cup D,V,f)$为决策信息系统。

对于信息表达系统 $S=(U,A,V,f)$,$P\subset A$,属性 a 的重要性是指:当 $a\in P$ 时,从 P 中删除 a 后,对 U 中个体区分能力减弱了多少;当 $a\notin P$ 且 $a\in A-P$ 时,当把 a 并入 P,对 U 中个体区分能力增强了多少。在此 P 称作参照系,离开参照系孤立地谈 a 的重要性没有实际价值。极端情况:$P=\phi$,空集 ϕ 对论域 U 的划分可以认为是 U 自身,即空集 ϕ 把 U 划分为一个类。将属性 a 并入 ϕ,则 a 对 U 中个体的区分能力就是属性集$\{a\}$对 U 中个体的区分能力,所以,单一属性对 U 中个体的区分能力是针对空集 ϕ 而言的。

属性集 P 对 U 中个体的区分能力可用划分等价类

$$U/P=\{y_2,y_2,\cdots,y_t\} \tag{1}$$

收稿日期:2008-05-23
基金项目:国家自然科学基金资助(60474019),河北省自然科学基金资助(F2005000482) 特约专稿
作者简介:刘志民(1979-),男,河北临漳人,助教,从事粗集理论及其应用研究。

定量描述:

若 $y_j(j=1,2,\cdots,t_1)$中多于一个个体,那么这些 y_j 中的这些个体是知识 P 下无法区分的;令 $|y_j|$表示 y_j 中包含的个体数目,则 y_j 中有 $|y_j|-1$ 个个体无法与 y_j 中的第一个个体区分开。于是,知识 P 区分不开 U 中的个体数目为

$$\sum_{j=1}^{t_1}(|y_j|-1) \tag{2}$$

当 $a\in P$ 时

$$U/P-\{a\}=\{X_1,X_2,\cdots,X_{t_2}\} \tag{3}$$

显然有 $t_2\leq t_1$,令

$$n_P(a)=\sum_{i=1}^{t_2}(|X_i|-1)-\sum_{j=1}^{t_1}(|y_j|-1) \tag{4}$$

则 $n_p(a)$表示从 P 中删除属性 a 后,因区分能力减弱而增加的区分不开的 U 中个体的数目。称

$$\omega_P(a)=n_P(a)/\sum_{b\in P}n_P(b) \tag{5}$$

为属性 a 关于属性集 P 的内重要性权重,简称内权重。

若 P 是A 的一个约简,则可用关于 P 的内重要性权重对 P 中的属性进行重要性大小的排序。

当 $a\notin P$ 时

$$U/P+\{a\}=\{Z_1,Z_2,\cdots,Z_{t_3}\} \tag{6}$$

显然 $t_3\geq t_1$,令

$$n_P(a)=\sum_{j=1}^{t_1}(|y_j|-1)-\sum_{i=1}^{t_3}(|Z_i|-1) \tag{7}$$

则 $n_P(a)$表示 a 并入 P 后,因区分能力增强而减少的区分不开的 U 中个体的数目。称

$$\omega_{P+}(a)=n_P(a)/\sum_{b\in A-P}n_P(b) \tag{8}$$

为属性 a 关于属性集 P 的外权重。称

$$\omega_{P+}(a^*)=\max_{a\in A-P}\{\omega_{P+}(a)\} \tag{9}$$

为关于 P 的最大外权重,称 a^* 为关于 P 的最大外权属性。

从 $A-P$ 中选择一种属性并入P,自然希望并入集对 U 中个体具有最大的区分能力。那么,首选属性就是 a^*,因为 a^* 关于 P 具有最大外权重。通常 P 是A 的核,且允许核 $Core(A)=\phi$。若 P 不是 A 的核,则要求 P 是独立的,并且 P 尚不是A 的一个约简。这样,根据最大外权重概念,可构造基于核的寻求约简的启发式搜索算法。

2 粗集中属性重要性权重的应用

定义最大外权属性,可由此构造寻求约简的一种启发式搜索算法。

2.1 基于最大外权求约简的启发式搜索算法

设 $S=(U,A,V,f)$为信息表,属性集 A 的核为 $Core(A)$,现从核出发求 A 的约简。

设 a_i 是关于 P_{i-1}的最大外权属性,令

$$\begin{cases}P_i=P_{i-1}\cup\{a_i\},(a_i\in A-P_{i-1})\\P_0=Core(A)\end{cases} \tag{10}$$

注意:a_i 的并入不能保证在 P_{i-1}中保留的核外属性不变成冗余属性。

①$P_1=P_0\cup\{a_1\}$。因为 $P_0=Core(A)$不是 A 的约简,所以 P_1 独立。

②$P_2=P_0\cup\{a_1\}\cup\{a_2\}$。因 $P_0+\{a_1\}$较 $P_0+\{a_2\}$对 U 中个体具有更强的区分能力;所以,a_2 的并入不会使保留属性 a_1 变为冗余。

③$P_3=P_0\cup\{a_1\}\cup\{a_2\}\cup\{a_3\}$。易见,$a_3$ 的并入,虽不会使 a_2 变为冗余;但是,却无法保证 a_1 不变成冗余。如 $P=P_0\cup\{a_2,a_3\}$是 A 的一个约简。

所以,当第 i 次$(i\geq 3)$并入属性 a_i 时,在并入集 $P_i=P_{i-1}+\{a_i\}$中,不能保证没有冗余属性,即 P_i 可能不独立。所以,对并入集 $P_i=P_{i-1}+\{a_i\}$ $(i\geq 3)$需作独立性检验。

检验方法:观察 U 中个体 $x_i(i=1,2,\cdots,N)$由核属性值及 $a_1,a_2,\cdots,a_i$ 各值组成的信息表,当去掉 $a_k(1\leq k\leq i-2)$所在列时,看是否出现相同的个体行,若出现,则 a_k 不能约去,说明 P_i 独立;否则 a_k 是冗余属性,即 P_i 不独立。由此得启发式搜索算法。

2.2 算法步骤

步骤 1. 令 $P_0=Core(A)$,$a_1\in A-P_0$ 是关于 P_0 的最大外权属性。

步骤 2. 令 $P_1=P_0\cup\{a_1\}$,若 P_1 是 A 的约简,则停止搜索。否则,P_1 是独立的,设 $a_2\in A-P_1$ 是关于 P_1 的最大外权属性。

步骤 3. 令 $P_2=P_1\cup\{a_2\}$,若 P_2 是 A 的约简则停止搜索;否则,P_2 是独立的。

步骤 4. 设 $a_i\in A-P_{i-1}$是关于 P_{i-1}的最大外权属性,令 $P_i=P_{i-1}+\{a_i\}(i\geq 3)$

①对 P_i 作独立性检验,若 P_i 中 $a_k(1\leq k\leq i-2)$冗余,则将 a_k 删除,同时将 a_k 从核外集中永久删除。验后的 P_i 仍记作 P_i。

②对独立的 P_i 检验其是否是 A 的约简。如

P_i 无非零最大外权属性,则 P_i 是约简,停止搜索并输出结果;否则,令 $i+1=i$ 返回步骤4。

因为 A 中仅含有有限种属性,所以一定存在自然数 K,使得 P_K 是 A 的一个约简。

2.3 算法讨论

因为每次从 $A-P_i$ 中选择关于 P_{i-1} 的最大外权属性 a_i 并入 P_{i-1},$P_i=P_{i-1}+\{a_i\}(i\geq 1)$,不管 a_i 最终是否被保留在约简中,一个不变的事实是:作为选择范围的 $A-P_i$ 每次都减少了一种属性。所以,算法复杂度是多项式的。

上述算法虽然避开了 *NP* 问题,并且能够找到一个 Pawlak 约简,却无法保证得到的是最小约简,只能是一个满意约简。

3 结束语

给定信息表达系统 $S=(U,A,V,f)$ 后,A 的核及 A 的所有约简都已客观存在,问题是采用怎样的算法,才能以尽可能少的计算量求出约简。本文定义一种外权重,并以此为启发性知识构造一种寻求约简的搜索算法。计算复杂度是多项式的,收敛速度快;不足是无法保证求出的是最小约简。

值得注意的是:粗集是一个特殊的系统,粗集中定义的任何一种属性权重都是特定条件下的属性权重。因为粗集与非粗集系统关于分类有完全不同的规则,它与非粗集系统的任何一种属性权重都有不同的内容、不同的算法和不同的应用。或说粗集中的属性权重不能应用于非粗集系统;非粗集系统的任何一种属性权重都不能用粗集算法获得。

参考文献:

[1] PAWLAK Z. Rough sets - theoretical aspects of reasoning about data [M]. Dordrecht: Kluwer Academic Publishers, 1991.

[2] 刘 清. Rough 集及 Rough 推理[M].北京:科学出版社,2001.

[3] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2001.

[4] 苗夺谦,胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究和发展,1999,36(6),681-684.

[5] 星学华,安 静. 基于粗集证据理论的指挥机构指挥能力的评估[J]. 科技信息,2007,(2):13-14.

[6] 曾黄麟. 粗集理论及其应用—关于数据推理的新方法[M]. 重庆:重庆大学出版社,1998.

[7] 王宏开,姚炳学,胡海清. 基于粗集理论的权重确定方法[J]. 计算机工程与应用,2003,39(6):21-22.

(责任编辑 闫纯有)

(上接第109页)

[3] 张 涛,刘土光. 开孔有限板的孔边应力场分析[J].华中科技大学学报,1993,30(1):87-89.

[4] 黄 义,韩建刚.薄板小波有限元理论及其应用[J].计算力学学报,2006,23(1):76-80.

[5] 杜守军,夏亨熹.二次B样条平面单元[J].河北农业大学学报,2002,25(4):143-148.

[6] CHEN W H, WU C W. Adaptable spline element for membrane vibration analysis[J]. Int. J. for Numerical Methods in England,1996,39:2457-2476.

[7] KO J, KURDILA A J. A class of finite element methods based on orthonormal compactly supported wavelets [J]. Computational Mechanics, 1995,(16):235-244.

[8] INGRID DAUBECHIES. 小波十讲[M]. 北京:国防工业出版社,2004.

(责任编辑 刘存英)

基于最大外权重的一种启发式属性约简算法

作者： 刘志民， LIU Zhi-min
作者单位： 河北工程大学,理学院,河北,邯郸,056038
刊名： 河北工程大学学报（自然科学版） ISTIC
英文刊名： JOURNAL OF HEBEI UNIVERSITY OF ENGINEERING(NATURAL SCIENCE EDITION)
年，卷(期)： 2008,25(3)
被引用次数： 2次

参考文献(7条)

1. PAWLAK Z Rough sets-theoretical aspects of reasoning about data 1991

2. 刘清 Rough集及Rough推理 2001

3. 王国胤 Rough集理论与知识获取 2001

4. 苗夺谦;胡桂荣 知识约简的一种启发式算法[期刊论文]-计算机研究与发展 1999(06)

5. 星学华;安静 基于粗集证据理论的指挥机构指挥能力的评估[期刊论文]-科技信息 2007(02)

6. 曾黄麟 粗集理论及其应用-关于数据推理的新方法 1998

7. 王宏开;姚炳学;胡海清 基于粗集理论的权重确定方法[期刊论文]-计算机工程与应用 2003(06)

引证文献(2条)

1. 李志军. 王昊. 马鸣霄. 郭继坤 软硬件协同设计中模拟退火划分算法的改进[期刊论文]-黑龙江科技学院学报 2011(2)

2. 董春游. 黄春楠 改进的差别矩阵属性约简方法[期刊论文]-黑龙江科技学院学报 2010(2)