

基于决策支持度的不完备信息系统约简算法^{*}

陶 志, 商 维, 李 卫民
(中国民航大学 理学院, 天津 300300)

摘 要: 提出了一种基于决策属性支持度的属性相对约简算法。通过引入决策属性支持度对不完备决策表中属性的重要性进行了定义, 并以此作为启发信息进行属性的选择, 该算法的时间复杂度是多项式的。寻找决策表中最小相对约简问题是典型的 NP-hard 问题, 采用该算法可降低问题复杂度。通过实例说明, 该算法能得到不完备决策表的最小相对约简。
关键词: 粗糙集; 决策属性支持度; 不完备决策表; 属性约简
中图分类号: TP 18 **文献标识码:** A **文章编号:** 0529-6579 (2007) 02-0023-04

粗糙集 (RS) 理论是由波兰学者 Pawlak 于 1982 年提出的^[1], 它作为一种刻划具有不完整性 和不确定性信息的全新数学工具, 已经成为人工智能领域的一个新的学术热点。其主要思想是, 在保持知识库的分类能力不变的前提下, 通过属性 (知识) 约简, 导出问题的决策或分类规则^[2]。目前, 粗糙集理论已被成功的应用于机器学习、决策分析、过程控制、模式识别与数据挖掘等领域^[3-4]。

知识约简是粗糙集理论中的核心内容之一。对于一个信息系统, 人们总是期望找到所有约简或最小约简。然而遗憾的是, 寻找所有约简或最小约简均为 NP 完全问题。近年来有学者提出多种知识约简算法, 绝大多数都是以完备信息系统作为研究对象。然而在现实中, 信息的不完备 (对象的属性值缺损) 现象是广泛存在的, 比如, 在医学诊断智能决策系统中, 由于一些病人不适合做某项仪器检查, 病人 (对象) 的某些症状 (属性值) 就不能够获得, 此时, 如何从不完备信息系统中获取有用的知识就显得更为重要。尽管有 RS 理论研究者曾提出过不完备信息系统的属性约简方法^[5-6], 然而, 这些方法均无法有效解决大规模信息系统的计算复杂性问题^[7]。

本文在相容关系 (自反性、对称性) 的基础上^[8], 首先给出不完备信息系统中条件属性对整体决策支持度的概念, 由此来描述由条件属性所提供的知识对整体决策的完全程度, 并通过决策属性支持度定义了条件属性对决策属性的相对重要性,

以此作为启发式信息, 提出了基于决策属性支持度的启发式知识相对约简算法。通过实例分析, 说明该方法是非常有效的。

1 不完备信息系统与相容关系

定义 1 四元组称为一 $S = (U, A = C \cup D, V, f)$ 个信息系统。其中 U 表示对象的非空有限集合, 称为论域; $A = C \cup D$ 表示属性的非空有限集合, C 为条件属性集合, D 表示决策属性集合, 且 $C \cap D = \phi, \forall a \in A, V_a$ 是属性 a 的值域; 表示 $U \times A \rightarrow V$ 的一个信息函数, 它为每个对象在每个属性上赋予一个信息值, 即 $\forall a \in A, x \in U, f(x, a) \in V_a$ 。若 $D = \phi$, 则称信息系统为数据表, 否则称为决策表。若存在一个 $x \in U, a \in C, f(x, a)$ 未知 (记作 $f(x, a) = *$), 则称信息系统是不完备的; 否则称信息系统是完备的。

对不完备信息系统, Kryszkiewicz 定义了 U 上的一个相容关系 T_P : $\forall \phi \subset P \subseteq A, T_P(x, y) \Leftrightarrow \forall a \in P, f(x, a) = f(y, a) \vee f(x, a) = * \vee f(y, a) = *, x, y \in U$

显然 T_P 满足自反性和对称性。由属性 $a \in A$ 确定的相容关系简记为 T_a 。

令 $T_P(x)$ 表示在 P 下与 x 相容的所有对象的全体, 即 $T_P(x) = \{y \in U \mid T_P(x, y)\}$ 。对于 P 而言, $T_P(x)$ 是与 x 可能不可区分的对象的最大集合。

^{*} 收稿日期: 2006-05-06
基金项目: 国家自然科学基金委员会与中国民用航空总局联合资助项目 (60672178); 中国民航大学博士启动基金资助项目 (05 qd02 9)
作者简介: 陶志 (1963 年生) 男, 博士, E-mail: 86543213@163.com
(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

对 $B \subseteq C$ 用 $U/T_B(x) = \{T_B(x) \mid x \in U\}$ 表示论域 U 的一个分类。 $U/T_B(x)$ 中的任何元素称为相容类。 $U/T_B(x)$ 中的相容类一般不构成 U 的划分, 它们构成 U 的一个覆盖。

对决策表 $S = (U, A = C \cup D, V, f)$ 因为专家在信息不完备的情况下仍能给出决策, 故我们可以假定决策属性值不为空。按照决策属性值可将论域划分为不相交的等价类, 记为: $U/D = \{Y_1, Y_2, \dots, Y_m\}$ 。对 $B \subseteq C$ 记 $D_B(x) = \{f(Y_i, D) \mid T_B(x) \cap Y_i \neq \emptyset, 1 \leq i \leq m\}$, $f(Y_i, D)$ 表示等价类 Y_i 的决策值, 即 $D_B(x)$ 表示 x 的 B 相容类的所有可能的决策值, 称为 x 的 B 分配集。

定义 2 对不完备决策表 $S = (U, C \cup D, V, f)$, $b \in B \subseteq C$ 若 $\forall x \in U$ 有 $D_{B \setminus \{b\}}(x) = D_B(x)$, 则称 b 在 B 是相对于决策属性不必要的 (简称相对不必要的), 否则称 b 在 B 中是相对必要的。若 B 中每个属性都是相对必要的, 则称 B 是相对独立的; 否则称 B 是相对依赖的。

由定义 2 知, 从条件属性集中去掉一个属性不会引起决策新的不一致, 则该属性在属性集中相对于决策是不必要的, 否则是必要的。

定义 3 对不完备决策表 $S = (U, C \cup D, V, f)$, $B \subseteq C$ 中所有相对必要的属性称为 B 的相对核, 记为 $Cor_B(B)$ 。

定义 4 对不完备决策表 $S = (U, C \cup D, V, f)$, 称 $R \subseteq B$ 是 B 的一个相对约简, 如果:

- 1) $D_R(x) = D_B(x)$;
- 2) R 是相对独立的。

可以证明相对核是所有相对约简的交集。即: 性质 1 $Cor_B(B) = \bigcap R \subseteq B$ 。其中 $R \subseteq B$ 表示 B 的相对约简。

2 相关子集与决策属性支持度

首先我们给出相关子集的概念:

定义 5 对不完备决策表 $S = (U, C \cup D, V, f)$, $U = \{x_1, x_2, \dots, x_n\}$, 定义对象 x 关于条件属性子集 $B \subseteq C$ 的相关子集如下:

$S_B(x) = \{y \mid y \in T_B(x) \text{ 且 } f(y, D) \neq f(x, D)\}.$

对于 B 而言, $S_B(x)$ 是那些与 x 可能不可区分的对象中, 其决策属性值与 x 不相同的对象的全体。

定义 6 对不完备决策表 $S = (U, C \cup D, V, f)$, $U = \{x_1, x_2, \dots, x_n\}$, 定义条件属性集 $B \subseteq C$ 对决策属性集 D 的支持度如下:

$K_B(D) = 1 - \frac{1}{|U|^2} \sum_{i=1}^n |S_B(x_i)|$

(其中 $|X|$ 表示集合 X 的基数)。

$\frac{1}{|U|^2} \sum_{i=1}^n |S_B(x_i)|$ 表示在已知 B 的条件下 D

仍具有的不确定性, 因此 $K_B(D)$ 是在 B 已知的条件下 D 所具有的确定性, $K_B(D)$ 越大说明依据 B 可使更多的对象被准确的化入 U/D 的等价类中。

定理 1 对不完备决策表 $S = (U, C \cup D, V, f)$, $U = \{x_1, x_2, \dots, x_n\}$, 若 $B \subseteq E \subseteq C$ 则 $K_B(D) < K_E(D)$ 。

证明: 由于 $B \subseteq E$ 所以有 $T_E(x_i) \subseteq T_B(x_i)$, $1 \leq i \leq n$ 于是, $\forall y \in T_E(x_i), f(y, D) \neq f(x_i, D) \Rightarrow y \in T_B(x_i), f(y, D) \neq f(x_i, D)$,

因此, $S_E(x_i) \subseteq S_B(x_i)$ 即有 $\frac{1}{|U|^2} \sum_{i=1}^n |S_E(x_i)| \leq \frac{1}{|U|^2} \sum_{i=1}^n |S_B(x_i)|$ 。

故 $1 - \frac{1}{|U|^2} \sum_{i=1}^n |S_E(x_i)| \geq 1 - \frac{1}{|U|^2} \sum_{i=1}^n |S_B(x_i)| \Rightarrow K_B(D) < K_E(D)$ 证毕。

3 条件属性对决策重要性的度量

决策属性支持度可以用来描述条件属性相对决策属性的重要程度。

定义 7 属性 $b \in B \subseteq C$ 在 B 中对决策的相对重要度定义为:

$sig_{B \setminus \{b\}}(b) = K_B(D) - K_{B \setminus \{b\}}(D)$

以上定义表明属性 $b \in B$ 在 B 中对决策 D 的重要性可由 B 中去掉 b 后所引起的决策属性支持度的变化大小来度量。由此, 不难得出以下结论。

定理 2 属性 $b \in B$ 在 B 中是相对必要的当且仅当 $sig_{B \setminus \{b\}}(b) > 0$ 。

定理 3 $Cor_B(B) = \{b \in B \mid sig_{B \setminus \{b\}}(b) > 0\}$ 。

定义 8 属性 $b \in C \setminus B$ 对 B 的相对重要度定义为:

$sig_B(b) = K_{B \cup \{b\}}(D) - K_B(D)$ 。

定义 8 表明, 一个属性对一个属性集的相对重要性是由添加它之后引起决策属性支持度的变化大小来度量, 因此可由此作为启发知识通过不断增加相对重要性最大的属性寻找最小约简。

4 基于决策属性支持度的知识约简算法

由定理 3 可以方便地求出不完备决策表的相对

核 $\text{cor}_B(C)$ 。以相对核作为求最小相对约简的起点，依照定义 8 中定义的属性重要性，逐次选择重要的属性添加到相对核中去，直至其决策属性支持度等于整个条件属性集 C 对 D 的支持度时为止。

算法步骤如下：

输入：不完备决策表 $S = (U \cup D, V, f)$ 。

输出： $S = (U \cup D, V, f)$ 的一个最小相对约简。

第一步：计算决策属性支持度 $K_C(D)$ 。

第二步：计算所有属性 $\in C$ 在 C 中的重要性 $\text{sig}_{C \setminus \{g\}}(g)$ 作 $\text{cor}_B(C) = \{c \in C \mid \text{sig}_{C \setminus \{g\}}(g) > 0\}$ 。

令 $B = \text{cor}_B(C)$ ，若 $K_B(D) = K_C(D)$ ，则 $\text{cor}_B(C)$ 为 C 的最小约简，输出 $\text{cor}_B(C)$ ，终止；否则进行第三步。

第三步：对属性集 $C \setminus B$ 重复作：

(1) 对每个 $g \in C \setminus B$ 计算 $\text{sig}_B(g)$ ；

(2) 选择属性 $g \in C \setminus B$ 使 $\text{sig}_B(g) = \max_{g \in C} \text{sig}_B(g)$ 作 $B = B \cup \{g\}$ ；

(3) 判断是否 $K_B(D) = K_C(D)$ ，若成立，则输出 B 否则转 (1)。

需要指出的是，本文给出的算法是尽量找出不完备信息系统的一个最小约简。而一般情况下，约简不是唯一的，最小约简也不是唯一的。

下面分析上述算法的时间复杂度。

第一步：对每个 $x \in U$ 需计算 $T_C(x)$ ，在最坏情况下需比较 $(|U| - 1)$ 个对象在 $|C|$ 个属性上的取值，故时间复杂度是 $O(|C||U|^2)$ ；

第二步：对每个 $g \in C$ 需计算 $K_{C \setminus \{g\}}$ ，故时间复杂度为 $O(|C|^2|U|^2)$ ；

第三步：(1) 的时间复杂度是 $O(|C|^2|U|^2)$ ，在最坏的情况下需循环 $|C|$ 次。故时间复杂度是 $O(|C|^3|U|^2)$ 。

综上可知上述算法时间复杂度是 $O(|C|^3|U|^2)$ 。

4 实例分析

为说明算法的有效性，我们利用文献 [5] 给出的关于小汽车的不完备决策表求最小相对约简 (见表 1)。

表 1 中条件属性集 $C = \{\text{价格, 里程, 空间, 最大速度}\}$ ，决策属性集 $D = \{\text{结论}\}$ 。利用基于决策属性支持度的知识约简算法求相对约简的过程如下：

表 1 不完备决策表
Tab 1 Incomplete decision table

汽车	价格	里程	空间	最大速度	结论
1	高	高	大	低	好
2	低	*	大	低	好
3	*	*	小	高	差
4	高	*	大	高	好
5	*	*	大	高	优
6	低	高	大	*	好

第一步：求决策属性支持度 $K_C(D) = 8/9$ 。

第二步：计算每个属性的相对重要性：

$\text{sig}_{C \setminus \{\text{价格}\}}(\text{价格}) = 0$ $\text{sig}_{C \setminus \{\text{里程}\}}(\text{里程}) = 0$

$\text{sig}_{C \setminus \{\text{空间}\}}(\text{空间}) = 1/12$ $\text{sig}_{C \setminus \{\text{最大速度}\}}(\text{最大速度}) = 1/9$ 于是， $\text{cor}_B(C) = \{\text{空间, 最大速度}\}$ 。令 $B = \{\text{空间, 最大速度}\}$ ，由于 $K_B(D) = K_C(D) = 8/9$ 故原不完备信息系统 (决策表) 的相对约简是 $\{\text{空间, 最大速度}\}$ 。

上述计算结果与本例的正确结果是完全一致的，通过实例分析可以看出，用本文提出的基于决策属性支持度的属性约简算法能找到不完备决策表的最小相对约简。

5 结 语

粗糙集理论通过属性约简，导出问题的分类或决策规则，而属性约简大多是基于完备信息系统的。本文针对基于相容关系的不完备决策表，通过定义相关子集和决策属性支持度引入条件属性对决策重要性的概念，以此作为启发式信息提出不完备决策表的属性约简算法。通过实例说明该算法能找到不完备决策表的最小相对约简。

参考文献：

[1] PAWLAK Z. Rough sets— theoretical aspects of reasoning about data [M]. Dordrecht Kluwer Academic Publishers, 1994: 9—30.

[2] 韩祯祥, 张琦, 文福柱. 粗糙集理论及其应用综述 [J]. 控制理论与应用, 1999, 16(2): 153—157.

[3] PAWLAK Z. Rough set theory and its application to data analysis [J]. Cybernetics and Systems, 1998, 29(9): 661—668.

[4] HU X H. Mining knowledge rules from databases— a rough set approach [A]. Proceedings of IEEE International Conference on Data Engineering [C], Los Alamitos: IEEE Computer Society Press, 1996: 96—105.

[5] LIANG J, XU Z. The algorithm on knowledge reduction in incomplete information systems [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Sys.

tem, 2002 10(1): 95—103

[6] YEE Leung WU Weizhi ZHANG Wenxiu Knowledge acquisition in incomplete information systems: A rough set approach [J]. European Journal of Operational Research, 2006 168 164—183

[7] 魏大宽, 黄兵. 不完备模糊目标信息系统粗集与知识约简[J]. 计算机工程, 2006(8): 48—51.

[8] KRYSZKIEWICZ M. Rough set approach to incomplete information systems [J]. Information Sciences, 1998 112 39—49

Knowledge Reduction Based on Decision Attribute Support Degree under Incomplete Information System

TAO Zhi SHANG Wei LI Weimin
(Civil Aviation University of China Science College, Tianjin 300300, China)

Abstract: A kind of attribute relative reduction for decision attribute support degree was proposed. States the significance of attributes in the incomplete decision table by introducing decision attribute support degree, which is then used as heuristic information for selecting attributes. This algorithm has a polynomial time complexity. Seeking the minimal relative reduction in a decision table is typically a NP-hard problem, and its complexity can be reduced by using the algorithm provided in this paper. An example shows this algorithm can achieve the minimal relative reduction of incomplete decision table.

Key words: rough set; decision attribute support degree; incomplete decision table; attribute reduction

(上接第 22 页)

A Novel Adaptive Group of Pictures Structure

WANG Yuan-ge^{1,2}, LIANG Fan¹, ZHANG Xiu-li¹, XIAO Ming-ming²
(1. Department of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou 510275, China;
2. Department Information Engineering, Zhongkai Agricultural College, Guangzhou 510225, China;
3. Jingshen Technical College, Jingshen 448000, China)

Abstract: Motion compensated temporal filtering (MCTF) decomposed a group of pictures (GOP) consisted of several successive frames to support temporal scalability by using subband/wavelet transform. Compared with fixed-size GOP structure, adaptive GOP structure (AGS) improves the coding performance, but the computational complexity of AGS is also increased significantly. A novel adaptive GOP structure (NAGS) is proposed to estimate the temporal characters of video sequences by inter-prediction. According to the temporal characters, it adopts different size of GOP. NAGS significantly reduces the computational complexity of AGS, and fixes the problem of the jitter of temporal decomposition delay in AGS. The experimental results show that NAGS has almost the same coding performance as AGS.

Key words: SVC; MCTF; lifting; GOP; AGS