文章编号: 1001-0920(2001)03-374-04

基于二元决策系统的粗集知识获取方法研究

王亚英, 邵惠鹤
(上海交通大学 自动化系, 上海 200030)

摘 要: 提出一种新的粗集知识获取方法, 首先将事例集表示成二元决策系统, 然后将其分解成一系列单一二元决策子系统。利用粗集理论对每一子系统进行分析, 推理出最优规则。在对决策系统进行条件属性和规则简化时, 提出了概率最佳简约准则和概率最小规则准则, 按照这两种最优准则可以获得概率意义上数目最小规则集。通过实例分析, 具体说明了该方法的实现步骤, 结果表明该方法具有明显的优越性。

关键词: 粗集; 知识获取; 二元决策系统; 决策规则
中图分类号: TP 18 文献标识码: A

Binary Decision System-based Rough Set Approach for Knowledge Acguisition

WANG Ya-ying, SHAO Hui-he
(Department of Automation, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: A new rough set approach for knowledge acquisition is presented. It firstly transforms the case sets into the binary decision system, then discompose the system into some simple subsystems according to decision attribute. Rough set theory is used to analyze data and induct the optimal rules to every subsystem. During reducing the condition attribute sets and rules, a probability-based optimal reduct criterion and a probability-based minimum rule criterion are presented, which can be used to get the minimum rule sets in a sense of probability. Finally operation procedure is given by analyzing a real example and the result testifies that it is better than other methods.

Key words: rough set; knowledge acquisition; binary decision system; decision rule

1 引 言

知识获取是构造专家系统的“瓶颈”问题, 而专家知识的好坏将直接影响整个系统的性能。归纳学习能从大量分散的事实和蕴含规律的数据中归纳出一般规则, 是解决知识获取问题的有效手段。但是许多专家系统要求计算机人员和领域专家的紧密合作, 并要求人工对专家知识进行编码建立, 不仅效率低下, 而且要在不同的应用领域进行相同的工作。因此, 迫切需要建立一种归纳学习方法, 以便从一个精心选取的专家决策样本集中自动归纳和提炼决策规则, 而与领域无关。

粗集理论是研究不完整数据及不精确知识的表达、学习、归纳的一套方法[1,2], 是以对观察和测量数据进行分类为基础, 通过对数据进行分析、近似分类、推理数据间的关系, 从中发现隐含的知识, 揭示

收稿日期: 1999-12-14; 修回日期: 2000-03-23
作者简介: 王亚英(1973—), 女, 山西广灵人, 博士生, 从事智能控制、粗集理论与应用等研究; 邵惠鹤(1936—), 男, 浙江宁波人, 教授, 博士生导师, 从事过程模型化及优化控制、智能控制等研究。

其潜在的规律。本文首先将事例集表示为二元决策系统, 并将二元决策系统分解为多个单一二元决策子系统; 然后针对每一子系统提取出最能反映其特性的条件属性简约集, 得到简化的子系统; 最后利用核与简化的概念对子系统中的决策规则进行简化, 得到简洁明了的决策规则集。

2 基于二元决策系统的粗集知识获取方法

在决策系统 $S = (U, C \cup D, V, f)$ 中: U 是对象集; $C \cap D = \varnothing$, 其中 C 表示条件属性集, D 表示决策属性集, 如果 $D = \{d\}$, 则称其为单一决策系统; $V = \bigcup_{a \in C \cup D} V_a$ 是属性的值域集, 如果 $V = \{0, 1\}$, 则称其为二元决策系统; f 是信息函数, $f: U \times C \cup D \to V$, 为指定 U 中每个对象的属性值。决策系统也可用数据表形式表示, (二元) 决策系统的数据表形式称为 (二元) 决策表。为了表示简单, 有时用 $(U, C \cup D)$ 表示决策系统。

任何事例集都可表示为二元决策表形式, 如表 1 的二元决策表便是旋转机械故障诊断事例集的决策表表示。表中的数值为 1, 说明该事例含有此症状 (或属于此类); 为 0 则说明不含症状(或不属于此类)。从决策表中提出决策规则主要有以下两步:

表 **1**　二元决策系统 $S = (U, C \cup Q)$

U	1	2	3	4	…	38	39	40	Q_1	Q_2	…	Q_8
1	1	0	0	1	…	0	0	0	1	0	…	0
2	0	0	0	0	…	0	0	0	1	0	…	0
3	0	0	0	1	…	0	0	0	1	0	…	0
4	0	0	0	0	…	0	0	0	1	0	…	0
⋮												⋮
87	0	0	0	0	…	1	0	1	0	0	…	1
88	0	0	0	0	…	1	1	0	0	0	…	1
89	0	0	0	0	…	0	1	1	0	0	…	1
90	0	0	0	0	…	0	0	1	0	0	…	1

1) 决策表中条件属性的简化, 即求取 C 的 D 简约, 得到简化的决策表。将事例集表示为二元决策系统, 然后将决策系统分解为多个单一二元决策子系统, 依据概率最佳简约准则为每个子系统选出最能反映其特性的简约属性集, 为最终得到简洁明了的决策规则提供了可能。

2) 通过消去简化决策表中决策规则的冗余条件属性值, 提取简化的决策规则。在单一二元决策系统 $(U, C \cup \{d\})$ 中, 每个正例代表一条规则, 可表示为 $\{\{C_i \mid C_i(x) = 1\} \to \{d \mid d(x) = 1\} \mid [x]_C \subseteq [x]_d\}$, $s = \mathrm{card}([x]_C)$。由于每条规则具有多种简化形式, 而且多条规则可能拥有相同的规则简化形式, 求取所有简化规则和最小简化规则是相当复杂的, 在属性较多时也是不可能的。为此提出一种概率最小规则准则, 在计算出每条决策规则的核属性的基础上, 依照它选择相应规则简化形式, 能够得出概率意义上的最小规则集。

下面给出概率最佳简约准则和概率最小规则准则。在此基础上, 利用一般的简约求取算法和规则粗集获取算法[3, 4], 即可求得概率最佳简约集和概率最小规则集。

概率最佳简约准则　在单一二元决策系统 $S = (U, C \cup \{d\})$ 中, 设正例数为 $\mathrm{Pos}N$, 其反例数为 $\mathrm{Neg}N = |U| - \mathrm{Pos}N$, 如果分别统计出各个条件属性 C_i 在该决策类的正例和反例中出现的频度 PO_i, $\mathrm{NO}_i (i = 1, 2, \cdots, n)$, 则其属性出现的相对频度 ($\mathrm{PO}_i/\mathrm{Pos}N$ 或 $\mathrm{NO}_i/\mathrm{Neg}N$) 大小, 在某种意义上便反映了判别该类决策的能力大小。将它作为属性的简约重要度, 即 $\mathrm{Con\ Sig}(C_i) = \max(\mathrm{PO}_i/\mathrm{Pos}N, \mathrm{NO}_i/\mathrm{Neg}N)$。在求取简约时, 以核为起点, 依次添加重要度较大的属性, 便可求得概率意义上的最佳简约集。

概率最小规则准则　在单一二元决策系统中, 决策表中的每个正例都代表一条规则, 如果将各属性在所有正例中属于核属性的频度 $P_i (i = 1, 2, \cdots, n)$ 作为该属性的规则重要度, 以任一条规则的核属性为起点, 依次添加出现规则重要度较大的其它属性, 便可能得到覆盖更多正例的规则简化形式, 从而使最终得到的规则数目较小。这种最小数目的决策规则是概率意义上得到的, 称为概率最小规则准则。

3 应用实例

现以文献[5 ～ 7] 的旋转机械故障诊断知识获取为例阐述本文方法的可行性。旋转机械常见故障有 8 类, 每类故障已获取了一批典型事例及其对应症状, 如表 2 所示(其各症状的具体意义略)。

1) 将故障事例集表示为二元决策系统 $S = (U, C \cup Q)$, 如表 1 所示。$U = \{1, 2, \cdots, 90\}$ 为对象集, 按照表 1 顺序排列; 条件属性集 $C = \{1, 2, \cdots, 40\}$ 为症状集; 决策属性集 $Q = \{Q_1, Q_2, \cdots, Q_8\}$ 表示故障类别。此二元决策系统可分解成 8 个单一子决

策系统 $S_i = (U, C \cup \{Q_i\}), i = 1, 2, \cdots, 8$。

2) 对于每个二元子决策系统, 根据概率最佳简约准则求其最佳简约集, 可得简化的子决策系统。以子决策系统 $S_1 = (U, C \cup \{Q_1\})$ 为例, 求得其最佳简约集为{5, 17}。简化子决策系统如表 3 所示, 其中 s 表示简化形式所覆盖的原决策系统的对象个数, 亦即支持度。

表 2 故障及其事例

代号	故障类别	故 障 事 例
Q_1	油膜振荡	(1, 4, 5) (15, 17) (4, 5) (17, 28, 29) (4, 5) (17, 28) (5, 17, 24) (4, 5, 15) (4, 17, 24) (5, 17, 24) (4, 5, 15) (4, 17, 24) (5, 17, 29) (4, 5, 17)
Q_2	初始质量不平衡	(2, 6, 11, 18) (6, 11, 18) (11, 14, 18, 25) (6, 11, 18) (2, 11, 14, 18) (6, 11, 18) (6, 11, 14, 18) (11, 14, 18, 25) (3, 11, 18) (2, 11, 18, 25) (2, 11, 18) (11, 18, 25) (6, 11, 14, 18) (6, 11, 18)
Q_3	叶片结垢	(6, 11, 34, 35) (6, 15, 19, 34, 37) (6, 11, 34, 36, 37) (6, 11, 35) (6, 11, 15, 36) (15, 19, 35, 36) (6, 11, 34, 37) (6, 11, 15, 35, 36)
Q_4	转子碰摩	(3, 10, 16, 31, 33) (10, 16, 20, 22) (10, 16, 20) (3, 10, 31, 33) (10, 20, 22) (10, 16, 20) (10, 16, 20, 22) (10, 15, 20) (10, 20, 22)
Q_5	转子弯曲	(12, 15, 23, 32) (12, 19, 23, 32) (6, 12, 23, 32) (3, 15, 19, 25) (6, 15, 23) (12, 19, 23) (19, 25, 32) (3, 12, 23) (3, 6, 15, 19) (6, 12, 19, 25) (12, 18, 23) (15, 19, 32) (3, 15, 19, 25) (13, 23, 32)
Q_6	转轴裂纹	(8, 13, 23, 27) (3, 13, 23, 26) (13, 26, 27) (3, 8, 13, 27) (8, 15, 23) (8, 13, 26, 27) (13, 23, 26) (15, 23, 27) (8, 13, 27) (15, 26, 27)
Q_7	转轴不对中	(13, 23, 25) (7, 13, 23) (7, 13, 23, 25) (7, 13, 23) (3, 7, 15, 23) (8, 23, 25) (3, 13, 23, 25) (3, 7, 13, 23) (7, 13, 23) (15, 23, 25) (7, 13, 23) (13, 23, 25)
Q_8	轴承座松动	(16, 20, 38, 39) (10, 20, 38, 40) (10, 16, 20) (10, 16, 30, 39) (10, 16, 20, 39) (10, 16, 38, 40) (16, 38, 39) (20, 39, 40) (10, 16, 20, 40)

表 3 简化的子决策系统 S_1

U	5	17	Q_1	s
1	1	0	1	5
2	0	1	1	5
3	1	1	1	4
4	0	0	0	76

3) 根据简化子决策系统计算出规则的核, 并依据概率最小规则准则求得规则最佳简化形式。对于表 3 的简化子决策系统 S_1, 求得其核如表 4 所示, 规则最佳简化形式为: $5 \rightarrow Q_1$, $s = 9$; $17 \rightarrow Q_1$, $s = 9$。

表 4 S_1 的规则核

U	5	17	Q_1	s
1	1	–	1	5
2	–	1	1	5
3	1	1	1	4
4	0	0	0	76

类推其它子系统的规则简化形式, 可得到表 5 所示的概率最小规则集。

文献[5, 6] 分别采用基于模拟退火算法和退火演化算法的知识获取方法, 尽管所得规则一致, 但与文献[7] 的基于遗传算法的知识获取方法相比, 在判别故障类别 Q_4 和 Q_8 时, 具有明显的不一致性, 这是由于在诊断事例集中存在不一致事例的缘故。例如事例(10, 16, 20) 既属于故障 Q_4, 又属于故障 Q_8, 使得这 3 种算法产生的规则不仅不一致, 而且也影响了对其它事例的判别性能。例如给定事例(20, 39, 40), 依照文献[5, 6] 和文献[7] 并不能判断出它属于 Q_4 还是属于 Q_8, 但根据本文方法可得出它属于 Q_8, 这与给定事例集的诊断结果是一致的。

表 5 概率最小规则集

规则类别	规 则
1	$5 \rightarrow Q_1, s = 9$; $17 \rightarrow Q_1, s = 9$
2	$(11, 18) \rightarrow Q_2, s = 14$
3	$34 \rightarrow Q_3, s = 4$; $35 \rightarrow Q_3, s = 4$ $(11, 15) \rightarrow Q_3, s = 2$
4	$22 \rightarrow Q_4, s = 4$; $(3, 10) \rightarrow Q_4, s = 2$ $(10, 15) \rightarrow Q_4, s = 1$
5	$12 \rightarrow Q_5, s = 7$; $(3, 19) \rightarrow Q_5, s = 3$; $(6, 23) \rightarrow Q_5, s = 2$; $32 \rightarrow Q_5, s = 6$
6	$26 \rightarrow Q_6, s = 5$; $27 \rightarrow Q_6, s = 7$; $(8, 15) \rightarrow Q_6, s = 1$
7	$7 \rightarrow Q_7, s = 7$; $(23, 25) \rightarrow Q_7, s = 6$
8	$39 \rightarrow Q_8, s = 5$; $40 \rightarrow Q_8, s = 4$

粗集方法具有一定的容错性, 能够更好地解决不确定不一致数据, 具有其它方法难以比拟的优越性。但是如果将该事例集表示为普通的决策系统, 利用粗集方法处理所得的规则不仅多而且复杂, 推理效果也远远不及本文方法。这是由于尽管故障事例集的症状较多, 但每个事例具有的症状却不多, 利用普通决策系统求取简约并不能选出对每种故障类型特别有利的属性集, 而将决策系统分解为多个单一二元子决策系统, 却可以克服这一缺点。

4 结　　语

基于二元决策系统, 本文提出一种新的粗集知识获取技术。它通过将事例集表示为多个单一二元决策系统, 充分利用粗集理论的优点, 依照概率最佳简约准则, 提取出能反映每类决策特点的最佳属性子集, 然后按照概率最小规则准则获得概率意义上数目最小规则集, 所得规则简洁明了。大量仿真事例表明, 本文方法不仅行之有效, 而且具有比其它方法明显的优越性, 可以实现知识库的自动生成, 为解决专家系统中知识获取的 "瓶颈"问题提供了一条有效途径。

参考文献:

[1] Pawlak Z. Rough sets- Theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic Publishers, 1991. 68-162.

[2] 曾黄麟. 粗集理论及其应用[M]. 重庆: 重庆大学出版社, 1998.

[3] Guan J W, Bell D A. Rough computational methods for information systems[J]. Artificial Intelligence, 1998, 105(1,2): 77-103.

[4] 吴福保, 李奇, 宋文忠. 基于粗集理论知识表达系统的一种归纳学习方法[J]. 控制与决策, 1999, 14(3): 206-211.

[5] 张雪江, 朱向阳, 钟秉林, 等. 基于模拟退火算法的知识获取方法的研究[J]. 控制与决策, 1997, 12(4): 327-331.

[6] 张雪江, 朱向阳, 钟秉林, 等. 基于退火演化算法的知识获取机制的研究[J]. 控制理论与应用, 1998, 15(1): 93-99.

[7] 彭志刚, 张纪会, 徐心和. 基于遗传算法的知识获取及其在故障诊断中的应用研究[J]. 信息与控制, 1999, 28(5): 391-395.