文章编号: 1001-0920(2001)06-0886-04

启发式知识约简算法的研究与应用

王亚英, 张春慨, 邵惠鹤
(上海交通大学 自动化系, 上海 200030)

摘 要: 从信息角度对决策系统中的属性重要度进行度量, 在此基础上, 提出一种知识约简的启发式算法, 它以信道容量为启发式信息, 减小了知识约简过程中的搜索空间。实例分析表明, 本算法能够获得决策系统的一种良好的相对约简。

关键词: 粗集; 信道容量; 知识约简; 启发式算法
中图分类号: TP 18 文献标识码: A

Research and Application of Heuristic Algorithm for Reduction of Knowledge

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

Abstract: The significance of attributes in decision systems is defined from the viewpoint of information. Then a heuristic algorithm for reduction of knowledge is presented, which uses the capacity as heuristic information to reduce the searching space during the reduction of knowledge. The experimental results show that this algorithm can find the better relative reduction of the decision system.

Key words: rough set; capacity; reduction of knowledge; heuristic algorithm

1 引 言

粗集理论是 Pawlak Z 等学者提出的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], 无需任何先验信息, 就能以对观察和测量所得数据进行分类的能力为基础, 通过对数据进行分析、近似分类以及推理数据间的关系, 从中发现隐含的知识, 揭示潜在的规律。知识约简是粗集理论的核心内容之一。知识库中的知识(属性)并不是同等重要, 而且某些知识是冗余的。冗余知识的存在, 不仅会造成资源的浪费(需要存储空间), 而且干扰人们作出正确而简洁的决策。所谓知识约简, 就是在保持知识库的分类或决策能力不变的条件下, 删除其中不相关或不重要的知识。

在粗集理论中, 决策系统是知识库的一种表示形式。一般来说, 决策系统的知识约简不是唯一的, 即对同一个决策系统可能存在多个相对约简。知识约简的目的是导出关于决策系统的决策规则, 因而相对约简中的属性性质与多少将直接影响决策规则的效果与繁简。然而找到最小约简却是 NP-hard 问题[3], 解决这类问题的一般方法是采用启发式搜索。本文从信息角度对决策系统中的属性重要度进行度

收稿日期: 2000-06-26; 修回日期: 2000-10-13
作者简介: 王亚英(1973—), 女, 山西大同人, 博士生, 主要从事人工智能与智能控制、粗集理论及应用等研究; 邵惠鹤(1936—), 男, 浙江宁波人, 教授, 博士生导师, 主要从事工业过程模型化与优化控制等研究。

量, 提出一种基于信息容量的知识相对约简算法, 以信道容量作为启发式信息, 来减小知识约简过程中的搜索空间。实验表明, 本算法能够获得决策系统的一种良好的相对约简。

2　基于信息容量的知识相对约简算法

2.1　决策系统与决策表

基于粗集理论的观点, 一个决策系统可表示为 $S = (U, C \cup D, V, f)$, 其中, U 是论域, 为非空有限集合; $C \cap D = \varnothing$, C 表示条件属性集, D 表示决策属性集, 均为非空有限集合; $V = \bigcup_{a \in C \cup D} V_a$ 是属性的值域集, 而 V_a 是属性 $a \in C \cup D$ 的值域; $f: U \times C \cup D \rightarrow V$ 是信息函数, 指定 U 中的每一个对象的属性值。

决策系统也可用数据表表示, 称之为决策表, 其中行代表对象, 列代表属性, 每行表示对象的一条信息, 对象 x 与属性 b 的交会点就是对象 x 在属性 b 下的值 $b(x)$, 如实例中表 1 所示。为了便于表达, 决策系统可以表示为 $(U, C \cup D)$。

2.2　互信息与信道容量

信息论是 C E Shannon 为解决信息传递(通信)过程问题而建立的一系列理论[4,5]。一个传递信息的系统是由发送端(信源)和接受端(信宿)以及连接两者的通道(信道)组成。信息是用来消除(随机)不确定性的东西, 信息量的大小由所消除的不确定性的大小来计量。在决策表中, 人们关心的是哪些条件属性对于决策更重要, 而条件属性与决策属性之间的互信息和信道容量便反映了信息量的大小。

定义 1　在决策表 $S = (U, C \cup \{d\})$ 中, 设可以按照决策属性 d, 将论域 U 分为 n 类, 即 $\{X_1, X_2, \cdots, X_n\}$, 作为传递信息系统的输入端 X; 而依照条件属性 $R \subseteq C$, 将论域 U 分为 m 类, 即 $\{Y_1, Y_2, \cdots, Y_m\}$, 作为系统的输出端 Y, 则其互信息的计算公式为

$$I(X, Y) = H(X) - H(X \mid Y)$$

其中, $H(X) = -\sum_{i=1}^{n} P(X_i) \log P(X_i)$ 为信息熵,

$$H(X \mid Y) = -\sum_{j=1}^{m} \sum_{i=1}^{n} P(Y_j) P(X_i \mid Y_j) \log P(X_i \mid Y_j)$$

为条件熵, 而 $P(X_i) = |X_i| / |U|$, $P(Y_i) = |Y_i| / |U|$ 为分布概率, $P(X_i \mid Y_j) = |X_i \cap Y_j| / P(Y_j)$ 为条件分布概率。

由信息论可知, 给定信道的互信息 $I(X, Y)$ 是给定输入的概率分布 $P(X)$ 的 $\cap$ 型函数, 而由 $\cap$ 型函数的性质可知, 一定存在一概率分布 $P(X)$, 使得互信息达到最大, 这个最大的互信息就称为信道容量(Capacity)。

定义 2　决策表中条件属性 R 与决策属性 D 之间的信道容量定义为

$$\text{Capacity}(R, D) = \max_{P(X)} \{I(X, Y)\}$$

无论 $P(X)$ 如何变化, 总不会大于 Capacity, 因此 Capacity 对于给定信道是个常数, 只与信道的统计特性有关, 是完全描述信道特性的参量, 也是信道能够传输的最大信息量, 比互信息更能反映客观情况。Capacity 的计算比较复杂, 一般采用迭代法计算[5]。

文献[6] 采用基于互信息的知识相对约简算法, 取得了很好的效果。但是互信息作为属性的重要度度量, 存在一个假设, 即训练例子集中的决策类别比例应与实际问题领域的类别比例相同[5]。而在一般情况下却不能保证相同, 这样计算训练集的互信息就有偏差。而信道容量, 不仅是不依赖于决策类别比例的属性度量值, 而且又是完全描述信道特性的参量, 是信道能够传输的最大信息量, 所以 Capacity 大的属性, 必是信息量大的属性, 对决策会更有利。因此, 我们选用信道容量作为属性的重要度度量, 可以克服互信息依赖于决策类别比例的缺点, 进行条件属性集的约简, 从而得到更好的更为客观的知识约简。

定义 3　设决策表 $S = (U, C \cup D)$, 其中, U 是论域, C 和 D 分别为条件和决策属性集, 且 $R \subset C$, 则对于任意属性 $p \in C - R$ 的重要度 $\text{SGF}(p, R, D)$ 定义为

$$\text{SGF}(p, R, D) = \text{Capacity}(R \cup \{p\}, D) - \text{Capacity}(R, D)$$

2.3　基于信道容量的知识相对约简算法

算法: CBARK(Capacity-based algorithm for reduction of knowledge)。

输入: 一个决策表 $S = (U, C \cup D)$, 其中, U 是论域, C 和 D 分别为条件和决策属性集。

输出: 该决策表的一个相对约简。

具体方法如下:

步 1: 计算 C 相对于 D 的核 $C_0 = \text{CORE}_D(C)$;

步 2: 令 $C_{\text{res}} = C - C_0$, 计算决策表 S 中的条件属性集 C_{res} 中的各属性的重要度, 并按其重要度的大小对 C_{res} 降序排序;

步 3: 令 $B = C_0$, 分别计算 D 相对于 C 和 B 的正域 $\text{POS}_C(D)$ 和 $\text{POS}_B(D)$, 如果 $\text{POS}_C(D) \neq$

$POS_B(D)$, 对条件属性集 C_{res} 进行如下重复:

1) 选择具有属性重要度的属性 $p \in C_{res}$, $C_{res} = C_{res} - p$, 计算 $POS_{B\cup\{p\}}(D)$;

2) 如果 $POS_{B\cup\{p\}}(D) = POS_C(D)$, $B = B \cup \{p\}$, 结束; 否则转 3);

3) 如果 $POS_{B\cup\{p\}}(D) = POS_B(D)$, 转 1); 否则 $B = B \cup \{p\}$, 转 1);

步 4: 最后得到的 B 就是 C 相对于 D 的一个约简。

利用信道容量作为属性的重要度, 与互信息作为属性的重要度一样, 会出现倾向于值域中含有较多值的属性, 这种倾向并不都合理, 为此, 对于属性值域含有值数目不等的决策表, 我们可以将其属性统统化为二值属性, 如属性天气的值域为{晴, 多云, 雨}, 可以分解成 3 个分属性: 天气 — 晴、天气 — 多云、天气 — 雨, 其值域都为{是, 否}, 这样就不存在偏向问题了; 然后分别求其分属性的属性重要度, 最后选择所包含的分属性的最大属性重要度作为该属性的重要度。具体方法为:

步 1: 计算 C 相对于 D 的核 $C_0 = CORE_D(C)$;

步 2: 令 $C_{res} = C - C_0$, 将 C_{res} 中的各个属性化为二元属性, 得到新的属性集 C';

步 3: 计算决策表 S 中的条件属性 $C \in C'$ 的重要度, 并依照所包含的分属性的最大属性重要度对 C 降序排序;

步 4: 令 $B = C_0$, 分别计算 C 和 B 关于 D 的正域 $POS_C(D)$ 和 $POS_B(D)$, 如果 $POS_C(D) \neq POS_B(D)$, 对条件属性集 C 进行如下重复:

1) 选择具有最大属性重要度的属性 $p \in C$, $C = C - p$, 计算 $POS_{B\cup\{p\}}(D)$;

2) 如果 $POS_{B\cup\{p\}}(D) = POS_C(D)$, $B = B \cup \{p\}$, 结束; 否则转 3);

3) 如果 $POS_{B\cup\{p\}}(D) = POS_B(D)$, 转 1); 否则 $B = B \cup \{p\}$, 转 1);

步 5: $B = B \cup C_0$, 最后得到的 B 就是 C 关于 D 的一个相对约简。

2.4 算法的复杂性

寻找知识相对约简的复杂性主要由决策表中的属性组合所决定。对于此算法, 在最坏情况下, 每次所考虑的属性数依次为 $M, M-1, \cdots, 1$(M 为决策表中的条件属性数), 故总次数为

$$M + (M-1) + \cdots + 1 = M(M+1)/2$$

如果忽略对象数对计算约简的影响, 那么, 在最坏情况下, 该算法能够在 $O(M^2)$ 时间复杂性内找到满意的约简。

3 应用实例

现以一气象状况实例系统作为决策系统, 如表 1 所示, 论域 $U = \{1, 2, \cdots, 14\}$, 条件属性集为 $C = \{a_1, a_2, a_3, a_4\}$, 决策属性为 d。

表 1 气象决策表

U	天气(a_1)	温度(a_2)	湿度(a_3)	有风(a_4)	类别(d)
1	晴	热	高	假	N
2	晴	热	高	真	N
3	多云	热	高	假	P
4	雨	温暖	高	假	P
5	雨	凉快	正常	假	P
6	雨	凉快	正常	真	N
7	多云	凉快	正常	真	P
8	晴	温暖	高	假	N
9	晴	凉快	正常	假	P
10	雨	温暖	正常	假	P
11	晴	温暖	正常	真	P
12	多云	温暖	高	真	P
13	多云	热	正常	假	P
14	雨	温暖	高	真	N

首先求得其核属性为 $C_0 = \{a_1, a_4\}$; 而后计算出属性集 $C_{res} = C - C_0 = \{a_2, a_3\}$ 中各属性的重要度分别为 0.021 885, 0.115 829, C_{res} 按属性重要度降序排序为$\{a_3, a_2\}$; 令 $B = C_0$, 因为 $POS_B(\{d\}) \neq POS_C(\{d\})$, 故从 C_{res} 中选出属性重要度较大的属性 a_3, 计算 $POS_{B\cup\{a_3\}}(\{d\})$, 因为 $POS_{B\cup\{a_3\}}(\{d\}) = POS_C(\{d\})$, 则 $B = B \cup \{a_3\} = \{a_1, a_3, a_4\}$ 即为所求的相对约简, 从而得到如表 2 所示的简化决策表。

表 2 简化决策表

U	天气(a_1)	湿度(a_3)	有风(a_4)	类别(d)
1, 8	晴	高	假	N
2	晴	高	真	N
3	多云	高	假	P
4	雨	高	假	P
5, 10	雨	正常	假	P
6	雨	正常	真	N
7	多云	正常	真	P
9	晴	正常	假	P
11	晴	正常	真	P
12	多云	高	真	P
13	多云	正常	假	P
14	雨	高	真	N

在此简化决策表的基础上, 利用规则提取方法[7] 可从中得出决策规则有 5 条:

1) (a_1, 晴) 且(a_3, 高), 则类别 = N;

2) (a_1, 雨) 且(a_4, 真), 则类别 = N;

3) (a_1, 多云), 则类别 = P;

4) (a_1, 雨) 且(a_4, 假), 则类别 = P;

5) (a_1, 晴) 且(a_3, 正常), 则类别 = P。

该决策表还有另一属性约简{a_1, a_2, a_4}, 以此约简为条件属性所得的决策规则有 7 条, 即:

1) (a_1, 晴) 且(a_2, 热), 则类别 = N;

2) (a_1, 雨) 且(a_4, 真), 则类别 = N;

3) (a_1, 多云), 则类别 = P;

4) (a_1, 雨) 且(a_4, 假), 则类别 = P;

5) (a_2, 温暖) 且(a_4, 真), 则类别 = P;

6) (a_2, 凉快) 且(a_4, 假), 则类别 = P;

7) (a_1, 晴) 且(a_2, 温暖) 且(a_4, 假), 则类别 = N。

因此, 原决策表的最佳属性约简为{a_1, a_3, a_4}, 从而说明本算法能够得到良好的属性约简。

4 结　　语

本文从信息角度对决策系统中的属性重要度进行度量, 提出一种新的知识约简的启发式算法。在此算法中, 以信道容量作为启发式信息, 来减小决策系统中知识约简过程中的搜索空间, 克服了互信息依赖于决策类别比例的缺点, 最终能够获得良好的相对约简。实验表明, 本算法不失为一种良好的知识约简方法。

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, 1996.

[3] Wong S K M, Ziarko W. On optimal decision rules in decision tables[M]. Poland: Bulletin of Polish Academy of Science, 1985. 693-696.

[4] 王育民, 梁传甲. 信息与编码理论[M]. 西安: 西北电讯工程学院出版社, 1986. 80-111.

[5] 陈文伟. 智能决策技术[M]. 北京: 电子工业出版社, 1998. 156-171.

[6] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 26(6): 681-684.

[7] 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.

(上接第 885 页)

表 3 给出了改进的 Hamming 网络对 5 个检验样本的识别结果, 其正确识别率为 100%。另外, 经过多次随机选出 26 个样本, 以各类项目的聚类中心形成该类的模式项目向量, 用其余 5 个样本进行验证, 正确率都为 100%。操作过程表明, 该网络计算较为简单, 利用计算机进行计算时所需存储单元较少, 且收敛速度较快。

表 3　5 个检验样本的实际类别及其经过 Hamming 神经网络迭代后输出的类别

检验样本	1	2	3	4	5
实际类别	A^1	A^1	A^2	A^3	A^4
网络识别结果	A^1	A^1	A^2	A^3	A^4

5 结　　语

上述案例分析表明, 改进的 Hamming 神经网络方法在 R&D 项目的中止决策时能够得到相当精确的结果, 且在应用上有较大的灵活性, 即样本向量的各元素的取值状态不必局限于双极性值。因此各指标值较容易取得; 从计算方法上看迭代输出一定能得到一个收敛的结果, 而且迭代收敛所需次数较少(一般仅需几步即可), 计算量也较少, 简单易行。

参考文献:

[1] 屈交胜, 官建成. R&D 项目中止决策的 Fuzzy 动态综合评判[J]. 科研管理, 1996, 17(9): 36-42.

[2] K Brockhoff. R&D project termination decisions by discriminant analysis —— An international comparison[J]. IEEE Trans on EM, 1994, 41(3): 245-254.

[3] 钟义信, 潘新安, 杨义先. 智能理论与技术 —— 人工智能与神经网络[M]. 北京: 人民邮电出版社, 1992.

[4] R Balachandra. Early warning signals for R&D projects [M]. MA: Lexington Books, 1989.

[5] V Kumar, A Persaud, U Kumar. To terminate or not ongoing R&D project: A managerial dilemma[J]. IEEE Trans on EM, 1996, 43(3): 273-284.