知识约简的一种启发式算法

苗夺谦[①②] 胡桂荣[①]
① (山西大学数学系 太原 030006)
② (中国科学院自动化研究所模式识别国家重点实验室 北京 100080)

摘 要 知识约简是 Rough Set理论研究中的核心内容之一,现已证明寻找决策表的最小约简是 NP-hard问题.文中首先从信息的角度,对决策表中属性的重要性给出度量;在此基础上,提出了一种基于互信息的知识相对约简的启发式算法,并指出该算法的复杂性是多项式的;最后,通过实例分析表明,在多数情况下该算法能够得到决策表的最小约简.

关键词 Rough Set理论,知识约简,启发式算法,算法复杂性

中图法分类号 TP18

A HEURISTIC ALGORITHM FOR REDUCTION OF KNOWLEDGE

MIAO Duo-Qian[①②] and HU Gui-Rong[①]
① (*Department of Mathematics, Shanxi University, Taiyuan* 030006)
② (*National Laboratory of Pattern Recognition, Insitute of Automation, Chinese Academy of Sciences, Beijing* 100080)

Abstract Reduction of knowledge is one of the important topics in the research on rough set theory. It has been proven that computing the optimal (minimal) reduction of decision table is a NP-hard problem. In the paper here, first, the significance of attributes in decision table is defined from the viewpoint of information; then, a heuristic algorithm based on mutual information for reduction of knowledge is proposed, and the complexity of this algorithm is analyzed; Finally, the experimental results show that this algorithm can find the minimal reduction for most decision tables.

Key words Rough set theory, reduction of knowledge, heuristic algorithm, complexity of algorithm

1 引 言

Rough Set理论是一种处理不精确、不相容和不完全数据的新的数学工具[1].目前,它正在被广泛应用于人工智能、模式识别与智能信息处理等领域,并取得了令人可喜的成果[2,3].

知识约简是 Rough Set理论的核心内容之一[1,4,5].众所周知,知识库中的知识(属性)并不是同等重要的,甚至其中某些知识是冗余的.特别,当知识库数据是随机采集的时,其冗余性更为普遍.冗余知识的存在,一方面是对资源的浪费(需要存储空间);另一方面,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 Rough Set理论中,依据是否考虑决策属性而将知识库分别表示为信息系统或决策表的形式.对信息

原稿收到日期: 1998-08-10;修改稿收到日期: 1998-11-26.本课题得到国家自然科学基金(项目编号 69805004)和山西省青年基金(项目编号 981017)资助.苗夺谦,男,1964年 4月生.博士,副教授,主要研究方向为 Rough Set理论与应用、KDD Data Mining、人工智能和模式识别.

系统的知识约简,文献 [6]已作了研究.本文主要讨论对决策表的知识约简,又称为知识的相对约简.

一般来讲,一个决策表的知识相对约简不是唯一的,即对同一个决策表可能存在多个相对约简.因为知识约简的目的是导出关于决策表的决策规则,约简中属性的多少直接影响着决策规则的繁简.因此,人们期望找到具有最少属性的约简,即最小约简.然而,遗憾的是 Wong. S. K. M和 Ziarko. W已经证明找出一个决策表的最小约简是 NP-hard问题[7].

导致 NP-hard的主要原因是属性的组合爆炸问题.在人工智能中,解决这类问题的一般方法是采用启发式搜索[8].本文首先从信息角度对决策表中的属性的重要性进行了定义,以此作为启发式信息,减小知识约简过程中的搜索空间.在此基础上,提出了基于互信息的知识相对约简算法.通过实验说明,在多数情况下本算法能够获得决策表的最小相对约简.

2 基于互信息的知识相对约简算法

一个决策表 T可以定义为四元组 $T=\langle U,C\cup D,V,f\rangle$,其中 U为论域;C,D分别为关于 U的条件和决策属性集;$V=\bigcup_{a\in C\cup D}V_a$,$V_a$ 表示属性 a的值域;$f: U\times(C\cup D)\to V$是一个信息函数,即对 $\forall x\in U,a\in C\cup D$,有 $f(x,a)\in V_a$.

在文献 [6]中,我们将 Rough Set理论中的知识看作是定义在论域 U的子集组成的 σ-代数上的随机变量.从而,引入了知识熵与互信息的概念.所有这些定义将是本文展开讨论的基础.

2.1 决策表中属性重要性的定义

在决策表中,人们关心的是哪些条件属性对于决策更重要.这就启示我们考虑条件属性与决策属性之间的互信息.我们认为,在决策表中添加某个属性所引起的互信息的变化大小可以作为该属性重要性的度量.

设 $T=\langle U,C\cup D,V,f\rangle$为一个决策表,且 $R\subset C$.那么,在 R中添加一个属性 $a\in C$之后互信息的增量为:

$$I(R\cup\{a\};D)-I(R;D)=H(D|R)-H(D|R\cup\{a\})$$

这里,$I(x;y)$表示 x与 y的互信息;$H(y|x)$表示已知 x时,y的条件熵.该增量越大,说明在已知属性 R的条件下,属性 a对决策 D就越重要.

定义 1. 设 $T=\langle U,C\cup D,V,f\rangle$是一个决策表,且 $R\subset C$.则对于任意属性 $a\in C-R$的重要性 $SGF(a,R,D)$定义为:

$$SGF(a,R,D)=H(D|R)-H(D|R\cup\{a\}).$$

若 $R=\Phi$,则 $SGF(a,R,D)$变为:

$$SGF(a,D)=H(D)-H(D|a)=I(a;D)$$

即为属性 a与决策 D的互信息.

$SGF(a,R,D)$的值越大,说明在已知 R的条件下,属性 a对于决策 D就越重要.本文将把 $SGF(a,R,D)$作为寻找最小知识约简时的启发式信息,减少搜索空间.

2.2 基于互信息的知识相对约简算法——MIBARK算法

由 Rough Set理论知道,任何决策表的相对核是唯一的,而且它包含在所有的相对约简之中,所以,相对核可以作为求最小知识约简的起点.在文献 [9]中,我们给出了知识相对约简的信息表示,并且证明了它与 Pawlak. Z的代数表示的等价性.由这些结论可知,互信息相等可以作为寻找知识相对约简的终止条件.

下面给出基于互信息的知识相对约简(MIBARK)算法.本算法是以 bottom-up的方式求相对约简的.它以决策表的相对核为起点,依据第 2.1节定义的属性重要性,逐次选择最重要的属性添加到相对核中,直到终止条件满足.

算法 1. MIBARK(mutual information-based algorithm for reduction of knowledge):

输入: 一个决策表 $T=\langle U,C\cup D,V,f\rangle$,其中,$U$为论域,$C,D$分别为条件和决策属性集.

输出: 该决策表的一个相对约简.

步骤 1. 计算决策表 T中条件属性 C与决策属性 D的互信息 $I(C;D)$;

步骤 2. 计算 C相对于 D的核 $C_0= CORE_D(C)$;

一般来说, $I(C_0; D) < I(C; D)$;有时,相对核 $C_0= \emptyset$,此时 $I(C_0; D)= 0$.

步骤 3. 令 $B= C_0$,对条件属性集 $C- B$重复:

① 对每个属性 $p\in C- B$,计算条件互信息 $I(p; D| B)$;

② 选择使条件互信息 $I(p; D| B)$最大的属性,记作 p(若同时有多个属性达到最大值,则从中选取一个与 B的属性值组合数最少的属性作为 p);并且 $B\leftarrow B\cup \{p\}$;

③ 若 $I(B; D)= I(C; D)$,则终止;否则,转①;

步骤 4. 最后得到的 B就是 C相对于 D的一个相对约简.

2.3 算法复杂性

寻找最小知识相对约简是 NP-hard问题,其复杂性主要是由决策表中的属性组合引起的.对于 MIBARK算法而言,在最坏情况下,每次所考虑的属性数依次为 $M, M- 1, \cdots, 1$(M为决策表的条件属性数).故总次数为

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

因此,如果忽略对象数对计算时间的影响,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 实例分析

为了考察 MIBARK算法的有效性,本节选择了一个已知其最小相对约简的决策表进行对比分析.该决策表见文献 [7].其中,论域 $U= \{1, 2, \cdots, 21\}$,条件属性集 $C= \{size, cyl, turbo, fuelsys, displace, comp, power, trans, weight\}$,决策属性 $D= \{mileage\}$.

Ziarko. W 在文献 [7]中讨论了该决策表的相对约简问题.他计算出了该决策表的所有约简 (共 7个),这些约简所含属性的个数分别为 4, 5, 6和 7,所以该决策表的最小相对约简为 $R= \{size, fuelsys, displace, weight\}$.这就意味着,如果以该决策表提供的信息为基础,预测汽车的 *mileage*,不必考虑该决策表中的 11个因素;只需考虑汽车的 *size*, *fuel* 的类型,引擎的 *displacement*和汽车的 *weight* 四个因素就足够了.这在实际应用中是极有价值的.

下面利用本文的 MIBARK算法对该决策表进行约简.

因为本算法以决策表的相对核为起点,所以,首先求出该决策表的相对核 $C_0= CORE_D(C)= \{fuelsys, weight\}$.

① 对该决策表,计算得 $I(C; D)= 0.5143$.

② 对该决策表的相对核 C_0,计算得 $I(C_0; D)= 0.3952$.

③ 令 $B= C_0$,对 $\forall p\in C- B$,计算条件互信息 $I(p; D| B)$:

表 1

属性 p	*size*	*comp*	*power*	*cyl*	*displace*	*turbo*	*trans*
$I(p; D| B)$	0.0874	0.0821	0.0725	0.0533	0.0533	0.0296	0.0210

可以看出,使条件互信息最大的属性为“*size*”,所以,更新后的 $B= \{fuelsys, weight, size\}$,并且,新的

$$\begin{aligned} I(B; D) &= I(Co; D)+ I(p; D| B) \\ &= 0.3952+ 0.0874 \\ &= 0.4826 \end{aligned}$$

下一步,用同样的方法可求得,使 $I(p; D| B)$最大的属性为“*displace*”,所以,更新后的 $B= \{fuelsys, weight, size, displace\}$;并且,新的 $I(B; D)= 0.5143$.

此时, $I(B; D)= I(C; D)$,故程序终止.因此,属性集 $\{fuelsys, weight, size, displace\}$就是该决策表的一个相对约简,约简后的决策表为表 2.

表 2 约简后的关于 car的决策表

条件属性				决策属性
size	*fuel*	*displace*	*weight*	*mileage*
compact	EFI	medium	medium	medium
compact	EFI	medium	light	high
compact	2BBL	medium	heavy	low
compact	EFI	medium	heavy	low
subcompact	2BBL	small	light	high
compact	2BBL	small	medium	medium
subcompact	EFI	small	light	high
subcompact	EFI	medium	medium	high
compact	2BBL	medium	medium	medium
subcompact	EFI	small	medium	high
subcompact	2BBL	small	medium	high
compact	EFI	small	medium	high

通过上面的分析可以看出,本文所给 MIBARK算法对该决策表能够找出最小相对约简.但该算法对最小约简的完备性还需进一步探讨.

4 结束语

Rough Set理论为开发自动规则生成系统提供了一种工具.它通过对决策表进行知识约简,从而导出其决策规则.由于知识约简的不唯一性,使得知识约简的优劣直接影响着决策规则的繁简.人们期望得到关于决策表的最简洁的规则,这就需要计算决策表的最小约简.然而,遗憾的是已经证明找出一个决策表的最小约简是 NP-hard问题.

本文首先从信息的角度,对决策表中属性的重要性给出度量;在此基础上,提出了一种基于互信息的知识相对约简的启发式算法,并指出该算法的复杂性是多项式的;最后,通过实例分析表明,在多数情况下该算法能够得到决策表的最小约简.但是,关于该算法对最小约简的完备性问题还需从理论上作进一步的探讨.

参 考 文 献

1 Pawlak Z. Rough Sets—— Theoretical Aspects of Reasoning About Data. Kluwer Academic Pub, 1991

2 王珏,苗夺谦,周育健. 关于 Rough Set理论与应用的综述. 模式识别与人工智能, 1996, 9(4): 337~ 344
(Wang Jue, Miao Duoqian, Zhou Yujian. Rough set theory and its application: A survey. Pattern Recognition and Artificial Intelligence (in Chinese), 1996, 9(4): 337~ 344)

3 王珏,王任,苗夺谦等. 基于 Rough Set理论的"数据浓缩". 计算机学报, 1998, 21(5): 393~ 400
(Wang Jue, Wang Ren, Miao Duoqian *et al*. Data enriching based on Rough set theory. Chinese Journal of Computers(in Chinese). 1998, 21(5): 393~ 400)

4 苗夺谦. Rough Set理论及其在机器学习中的应用研究 [博士学位论文], 中国科学院自动化研究所,北京, 1997
(Miao Duoqian. Rough set theory and its application in machine learning [Ph D dissertation](in Chinese). Institute of Automation, Chinese Academy of Sciences, Beijing, 1997)

5 Wang Jue, Miao Duoqian. Analysis on attribute reduction strategies of Rough set. Journal of Computer Science and Technology, 1998, 13(2): 189~ 192

6 Miao Duoqian, Wang Jue. An information-based algorithm for reduction of knowledge. IEEE ICIPS 97. 1997. 1155~ 1158

7 Wong S K M, Ziarko W. On optimal decision rules in decision tables. Bulletin of Polish Academy of Sciences, 1985, 33: 693~ 696

8 陆汝钤. 人工智能.北京: 科学出版社, 1996
(Lu Ruqian. Artificial Intelligence(in Chinese). Beijing Science Press, 1996)

9 苗夺谦, 王珏. 粗糙集理论中概念与运算的信息表示. 软件学报, 1999,
(Miao Duoqian, Wang Jue. An information representation of concepts and operations in rough set theory. Journal of Software(in Chinese), 1999, 10(2): 113-116)