

文章编号:1001-9081(2010)04-1086-03

一种新的决策表核属性计算方法

冯 林^{1,2}

(1. 四川师范大学 计算机科学学院,成都 610101;
2. 四川师范大学 可视化计算与虚拟现实四川省重点实验室,成都 610068)
(scfengyc@126.com)

摘要:属性约简是粗糙集理论研究的一个核心问题,而核属性的确定往往是决策表中属性约简的基础。结合决策表的树型结构表示,给出了决策表中正域和非正域的计算方法,并从核属性的定义出发,计算树型决策表中正域和非正域相对于属性全集正域和非正域的变化,提出了一种计算决策表中核属性的方法。对其时间和空间复杂度的分析,以及对一个气象决策表例子的实验结果,证明了这些方法的有效性。

关键词:粗糙集;决策表;属性约简;核属性

中图分类号: TP18 文献标志码:A

New algorithm for computing attribute core in decision tables

FENG Lin^{1,2}

(1. College of Computer Science, Sichuan Normal University, Chengdu Sichuan 610101, China;
2. Sichuan Key Laboratory of Visualization Computing and Virtual Reality, Sichuan Normal University, Chengdu Sichuan 610068, China)

Abstract: Attribute reduction is one of the key problems in rough set theory, and determination of core attribute is the basis to solve the problem of attribute reduction. First, by using the tree structure knowledge expression in the decision table, an approach for computing positive and negative regions was introduced. Next, according to changes of positive and negative regions in tree structure decision table relative to the conditional attribute set, an algorithm for computing core attributes was developed. An efficiency of the proposed methods was illustrated by time and space complexities analysis and experimental result in a weather decision table.

Key words: rough set; decision table; attribute reduction; core attribute

0 引言

粗糙集理论由波兰逻辑学家 Pawlak 教授于 1982 年提出。由于粗糙集理论能够定量分析处理不严密、不确定或不完全的信息和知识,因此,作为一种具有极大潜力和有效的智能信息处理技术受到了广泛关注。它已在特征选择^[1]、决策规则生成^[2]、数据融合^[3]、故障诊断^[4]等方面取得了较成功地应用。

决策表(也称决策信息系统)是粗糙集理论的知识表达工具,决策表中的属性约简是粗糙集理论及其应用研究的热点问题之一,而决策表中属性核的计算对解决属性约简这一核心问题具有极其重要的意义。它能有效缩小属性约简算法在属性空间的搜索范围,降低属性约简算法的复杂度。因此,如何高效地确定一个决策表中的属性核显得非常关键。在计算核属性诸多方法中,使用较广泛的是文献[5]中提出的利用分明矩阵来确定核属性,它首先建立决策表中的分明矩阵,然后指出分明矩阵中单元素属性即为决策表的核属性;文献[6]中指出了文献[5]方法的错误,并定义了一个新的分明矩阵来求解核属性;文献[7]中对文献[5]的方法进行了改进,改进后的算法主要优点是不需要建立分明矩阵的中间步骤,因此,降低了文献[5]方法的空间、时间复杂度。

通常决策表都是以表结构来表示和存储数据的,树型结构是决策信息系统的另外一种表示方式,目前已有一些研究

人员在粗糙集理论的研究中引入了树型结构^[8-9]。然而,对于树结构表示下的属性核计算的研究却没有得到广泛关注,本文基于决策表的树型结构给出了一种计算决策表中核属性的算法。根据对算法的复杂度分析及实验结果,证明了本文算法的高效性。

1 核属性判定方法

为了方便叙述,首先引入与本文相关的粗糙集理论的基本概念^[10]。

定义 1 决策表或决策信息系统。一个决策表 $S = (U, R, V, f)$, 其中 U 是对象的集合,也称为论域, $R = C \cup D$ 是有限非空属性集合, $C \cap D = \emptyset$, $D \neq \emptyset$, 子集 C 和 D 分别称为条件属性集和决策属性集, $V = \bigcup_{r \in R} V_r$ 是属性值的集合, V_r 表示属性 r 的值域, $f: U \times R \rightarrow V$ 是一个信息函数,它指定 U 中每个对象 x 的属性值。

文献[10]指出:一个多决策属性的决策表可以容易地转换为多个单决策属性的决策表,因此,本文讨论的均为单决策属性决策表。

定义 2 不分明关系。给定 $S = (U, R, V, f)$, $B \subseteq R$, 不分明关系 $IND(B)$ 定义如下:

$$IND(B) = \{(x, y) \mid (x, y) \in U \times U; \forall_{b \in B} \forall_{x \in U} \forall_{y \in U} (f(x, b) = f(y, b))\}$$

定义3 条件分类和决策分类。给定 $S = (U, C \cup D, V, f)$, 设 $U/IND(C)$ 和 $U/IND(D)$ 分别为论域 U 在属性集 C 和 D 上形成的划分, 条件分类定义为 E_i ($E_i \in U/IND(C)$), 决策分类定义为 X_j ($X_j \in U/IND(D)$), 其中, $|X|$ 为集合 X 的势。

定义4 相对正域。设 U 为一个论域, P, Q 为定义在 U 上的两个等价关系簇, Q 的 P 正域 $POS_P(Q)$ 定义如下:

$$POS_P(Q) = \bigcup_{x \in U/Q} P_-(X)$$

定义5 近似分类质量。给定 $S = (U, C \cup D, V, f)$, $B \subseteq A, B$ 对决策 D 的近似分类质量 $\gamma_B(D)$ 定义为:

$$\gamma_B(D) = |POS_B(D)| / |U|$$

定义6 属性核。设 P, Q 是定义在 U 上的两个等价关系簇, 若 $POS_P(Q) = POS_{P \setminus \{r\}}(Q)$, 则称属性 r 为 P 中相对于 Q 不必要的, 否则称 r 为 P 中相对于 Q 必要。 P 中所有相对于 Q 必要的属性组成的集合称为 P 的 Q 核。

根据以上定义, 给出核属性的判定定理。

定理1 给定 $S = (U, C \cup D, V, f)$, $a \in C$, 如果 a 是核属性, 删去 a 有以下两种情况:

- 1) 至少存在属于正域但不属于同一条件分类的两个对象产生冲突(根据定义6, 容易得证);
- 2) 至少存在一个原本在决策表正域中的对象, 一个原本在决策表非正域上的对象, 它们在条件属性 $C \setminus \{a\}$ 上的取值一样。

证明

因为 a 是核属性, 删去 a 后, 根据定义6, 有 $POS_c(D) \neq POS_{C \setminus \{a\}}(D)$ 成立。也就是说, 条件属性集 C 对决策 D 的正域必将发生改变, 那么, 一定存在对象 $x \in POS_c(D)$, 而 $x \notin POS_{C \setminus \{a\}}(D)$ 。

另一方面, 由于 $U/IND(C)$ 是 $U/IND(C \setminus \{a\})$ 的细分, 因此, 至少存在一对对象 $y \in U$, 有 $(x, y) \in IND(C \setminus \{a\})$ 成立, 即结论2) 成立。证毕。

定理2 给定 $S = (U, C \cup D, V, f)$, $a \in C$, 如果 a 是核属性, 则 $\gamma_c(D) > \gamma_{C \setminus \{a\}}(D)$ 。

证明 在 S 中, 如果 a 是核属性, 由定理1知:

$$POS_c(D) \supseteq POS_{C \setminus \{a\}}(D)$$

即有:

$$|POS_c(D)| > |POS_{C \setminus \{a\}}(D)|$$

成立。根据定义5得:

$$\gamma_c(D) > \gamma_{C \setminus \{a\}}(D)$$

推论1 给定 $S = (U, C \cup D, V, f)$, 设 $POS_c(D)$ 是 S 的正域, $NPOS_c(D)$ 是 S 的非正域, 属性 a 是核属性的充要条件是: $\exists x \in POS_c(D) \wedge x \in E_i$ 满足:

$$\exists y \in NPOS_c(D) \wedge (y \in E_i)$$

或者:

$$\exists y \in POS_c(D) \wedge (y \in E_i) \wedge (D(x) \neq D(y))$$

其中: $E_i \in U/IND(C \setminus \{a\})$ ($i = 1, \dots, n$), n 为条件分类的个数。

决策表 S 的树表示是通过 S 的所有条件属性 C 和决策属性 D 的取值来决定的, 该树型结构的高度为 $(|C| + |D|)$, 且所有对象都在树型结构的同一层。

根据推论1可知: 如果要判定某个属性是否是核属性, 首先要得到信息系统 S 的正域和非正域, 由此, 我们给出树表示下决策表中正域和非正域的计算方法。

1.1 核属性计算方法描述

算法1 树表示下决策表中正域和非正域的计算方法。

输入: $S = (U, C \cup D, V, f)$ 。

输出: S 的正域 $POS_c(D)$ 和非正域 $NPOS_c(D)$ 。

Step1 $POS_c(D) = \emptyset, NPOS_c(D) = \emptyset, Tree = NULL$;

Step2 For $\forall x_i \in U (i = 1, 2, \dots, |U|)$ Do

If $Tree = NULL$ Then

InsertToTree(Tree, x_i) End If

Else { $j = 1$;

While ($j \leq |C| + 1$) Do

根据 x_i 在 C_j 的取值寻找 x_i 在 $Tree$ 中的分支

$j = j + 1$;

Loop

}

Loop

Step3 比较对象在 C_m 和 D 上的取值,

1) 如果对于同一 C_m 有多个分支, 且这些分支对应的对象集合为 $\{U_{i1}, U_{i2}, \dots, U_{ik}\}$ (k 为分支数目), 则 $NPOS_c(D) = NPOS_c(D) \cup U_{il}$ ($l = 1, 2, \dots, k$);

2) 如果对于同一 C_m 只有一条单支, 且这条单支对应的对象集合为 U_j , 则 $POS_c(D) = POS_c(D) \cup U_j$ 。

Step4 Stop。

在得到了 S 的正域和非正域后, 下面给出 S 在树表示的属性核的算法。

算法2 S 的属性核计算方法。

输入: 由算法1得到的正域 $POS_c(D)$ 和非正域 $NPOS_c(D)$ 。

输出: S 的属性核 $Core$ 。

Step1 $Core = \emptyset, i = 1$;

Step2 For $\forall a_i \in C (i = 1, 2, \dots, |C|)$ Do

$Tree = NULL$; //Tree 中不含属性 a_i

For $\forall x \in POS_c(D)$ Do

If $Tree$ 中有某个叶子节点 y 满足 $C \setminus \{a_i\}$

$(x) = C \setminus \{a_i\}(y) \wedge D(x) = D(y)$ Then

$Core = Core \cup \{a_i\}; i = i + 1$; 转 Step2;

Else InsertToTree(Tree, x);

End If

Loop

For $\forall x \in NPOS_c(D)$ Do

If $Tree$ 中有某个叶子节点 y 满足 $C \setminus \{a_i\}$

$(x) = C \setminus \{a_i\}(y)$ Then

$Core = Core \cup \{a_i\}; i = i + 1$; 转 Step2;

End If

Loop

$i = i + 1$;

Loop

Step3 Stop

1.2 算法分析

1.2.1 时间复杂度分析

对于本文计算核属性方法,算法1的时间复杂度为 $O(|C|+|U|)$,算法2的时间复杂度为 $O(|C|^2+|U|)$,因而总的时间复杂度为 $O(|C|^2+|U|)$;文献[7]中的算法时间复杂度为 $O(|C|^2+|U|\log|U|)$;文献[5],由于要用分明矩阵来计算属性核,建立分明矩阵的时间复杂度为 $O(|C|+|U|^2)$,再由分明矩阵计算核属性的时间复杂度为 $O(|C|+|U|^2)$,因此整个算法的时间复杂度为 $O(|C||U|^2)$ 。因此,本文算法的时间复杂度低于文献[5]与文献[7]中算法的时间复杂度。

1.2.2 空间复杂度分析

对于本文计算核属性方法,由于算法存在着一个建树的过程,因而其空间复杂度较大,其空间复杂度为 $O(\vartheta|C|+|U|)$ (ϑ 为一个正的常数)。文献[7]中的空间复杂度 $O(|U|)$,而文献[5]中用分明矩阵来计算属性核算法的空间时间复杂度 $O(|C|+|U|^2)$ 。因此,本文算法的空间复杂度高于文献[7]中算法空间复杂度,低于文献[5]中算法的空间复杂度。

表1 关于气象例子的决策表

序号	条件属性集 C				决策属性 D
	Outlook(a_1)	Temperature(a_2)	Humidity(a_3)	Windy(a_4)	
1	Sunny	Hot	High	False	N
2	Sunny	Hot	High	True	N
3	Overcast	Hot	High	False	P
4	Rain	Mild	High	False	P
5	Rain	Cool	Normal	False	P
6	Rain	Cool	Normal	True	N
7	Overcast	Cool	Normal	True	P
8	Sunny	Mild	High	False	N
9	Sunny	Cool	Normal	False	P
10	Rain	Mild	Normal	False	P
11	Sunny	Mild	Normal	True	P
12	Overcast	Mild	High	True	P
13	Overcast	Hot	Normal	False	P
14	Rain	Mild	High	True	N

表2 各方法所用时间

属性核方法	所用时间/s
本文方法	0.016
文献[5]中方法	0.027
文献[6]中方法	0.029
文献[7]中方法	0.018

3 结语

属性核的确定是粗糙集理论研究的一个重要内容。本文从核属性的定义出发,结合决策表中树型结构的表示,给出了树表示下求属性核的有效方法。虽然本文的工作只在一个小数据集上进行了实验,但文中方法可以扩充到海量数据集中,因此,本文方法对基于粗糙集理论的海量数据挖掘也提供了一种有益的尝试。

致谢:感谢重庆邮电大学胡峰副教授为本文提供的大力支持与帮助!

参考文献:

- [1] 王加阳,薛双盈.一种快速广义动态约简算法[J].计算机工程,2008,34(21):56-58.

2 实验测试

为了验证上述方法的有效性,给出一个表1所示的气象决策表例子。

首先,根据算法1及算法2,可求得表1属性核 $Core = \{a_1, a_4\}$;而文献[5-7]中方法得到的属性核 $Core = \{a_1, a_4\}$,说明本文方法与一些粗糙集的经典方法,计算结果完全一致。

为了验证本文方法的高效性,对表1中的数据进行实验,实验的硬件环境是:CPU:Intel Core Duo 2.2 GHz,内存2 GB,Windows XP 操作系统。

具体测试过程如下:

- 1)用本文方法计算表1中的属性核并记录其运行时间;
- 2)用文献[5-7]中方法计算表1的属性核,并分别记录其运行时间,结果见表2。

从表2测试结果来看,对比经典求属性核方法,本文方法具有较高的运行效率。

- [2] 张雪英,刘凤玉, KRAUSE J. 粗糙集分类算法的近似决策规则和规则匹配方法[J]. 计算机科学, 2005, 32(6): 129-132.
- [3] 徐捷,徐从富,耿卫东,等. 基于粗糙集理论的动态目标识别及跟踪[J]. 电子学报, 2002, 30(4): 605-607.
- [4] 肖迪,胡寿松. 实域粗糙集理论及属性约简[J]. 自动化学报, 2007, 33(3): 253-258.
- [5] HU X H, CERCONE N. Learning in relational database: A rough set approach [J]. International Journal of Computational Intelligence, 1995, 11(2): 323-338.
- [6] 叶东毅,陈昭炯. 一个新的差别矩阵及其求核方法[J]. 电子学报, 2002, 30(7): 1086-1088.
- [7] 赵军,王国胤,吴中福,等. 一种高效的属性核计算方法[J]. 小型微型计算机系统, 2003, 24(11): 1950-1953.
- [8] 苗夺谦,王珏. 基于粗糙集的多变量决策树的构造方法[J]. 软件学报, 1997, 8(6): 425-431.
- [9] 梁道雷,黄国兴,金健. 一种多变量决策树方法研究[J]. 计算机科学, 2008, 35(1): 211-212.
- [10] 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.