2010年 2月 第 2期
Vol. 31 No. 2 2010

一种无决策属性的信息系统的属性约简算法

朱颢东[1,2], 钟 勇[1,2]

[1](中国科学院 成都计算机应用研究所, 四川 成都 610041)
[2](中国科学院 研究生院, 北京 100039)
E-mail: zhuhaodong80@163.com

摘 要: 经典属性约简及其延伸算法是基于有决策属性的信息系统的属性约简算法, 它们对无决策属性的信息系统的属性约简无能为力. 为此, 本文以粗集理论为基础, 对无决策属性的信息系统从集合论的论域划分方面进行研究, 提出了一种适用于无决策属性的信息系统的启发式属性约简算法. 该算法在一定程度上能够解决无决策属性的信息系统属性约简问题, 进一步扩展了粗集理论的应用范围. 实例表明该算法是有效可行的.

关 键 词: 属性约简; 决策属性; 信息系统; 集合论
中图分类号: TP301 文献标识码: A 文 章 编 号: 1000-1220(2010)02-0360-03

Attribution Reduction Algorithm on Information Systems without Decision Attributes

ZHU Hao-dong[1,2], ZHONG Yong[2]

[1] (Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu Sichuan 610041, China)
[2] (The Graduate School of the Chinese Academy of Sciences, Beijing 100039, China)

Abstract: The classical attribute reduction algorithm and its extended algorithms base on information systems with decision attributes and can not be applied to attribute reduction of no decision attributes information systems. So, based on rough set theory, this paper studied no decision attributes information systems in domain division of set theory and presented a heuristic attribute reduction algorithm. To a certain extent, the algorithm can resolve the attribute reduction problem of no decision attributes information systems and extend application of Rough Set Theory. The analysis of the realistic example shows that the algorithm is effective and feasible.

Key words: attribute reduction; decision attribute; information system; set theory

1 引 言

粗集理论是波兰数学家 Pawlak于 1982年提出的研究不确定和不精确知识的数学工具[1,2], 已被成功应用于人工智能, 数据挖掘, 模式识别与智能信息处理等领域, 引起了国际学术界的关注[3,4]. 由于属性约简是 RoughSet理论的重要研究内容之一, 许多研究者都致力于信息系统的属性约简算法研究. 目前人们已经提出了多种属性简算法, 如基于正域的属性约简算法, 基于区分矩阵的属性约简算法和基于信息熵的属性约简算法等. 不过, 这些属性约简算法都是基于有决策属性的信息系统. 无决策属性的信息系统也是一类重要的数据集, 如在聚类分析中使用的信息系统就没有决策属性. 对于无决策属性的信息系统, 如不进行属性约简, 其高维稀疏特性会使得一些与其相关的算法性能急剧下降, 不仅需要花费很长的时间, 而且算法结果也很难令人满意. 为了解决这个问题, 最有效的方法就是通过属性约简来进行降维. 目前, 在无决策属性的信息系统的属性约简方面人们往往使用的是数据挖掘中的无监督特征选择方法, 基于粗集理论的无决策属性的信息系统的属性约简算法寥寥无几. 在这种情况下, 本文以粗集属性约简理论为基础, 对无决策属性的信息系统从集合论的论域划分方面进行研究, 从而提出了一种适用于无决策属性的信息系统的启发式属性约简算法. 实例分析表明该算法是可行的.

2 粗集理论的相关基本概念

粗糙集理论从新的视角出发对知识进行了定义, 它将知识定义为不可区分关系的一个族集, 这使得知识具有了一个清晰的数学意义, 并可用数学方法进行处理.

定义 1. 信息系统可以表示为 $S=\langle U, R, V, f\rangle$, U为对象集合, $R=C\cup D$是属性集合, 其中 C为条件属性集, D为决策属性集, V是属性值的集合, V_r表示属性 r的值域, $f: U\times R\to V$是一个映射函数, 它指定 U中每一个对象 x的属性值. 信息系统也可用二维表来表示, 称之为决策表, 其中行代表对象 x_i, 列代表属性 r, $r(x_i)$表示第 i个对象在属性 r上的取值[5]. 如果在信息系统中 $D=\varnothing$, 则此时的信息系统称为无决策属性的信息系统, 表示为 $S=\langle U, C, V, f\rangle$.

定义 2. 对于每个属性子集 $B\subseteq C$, 定义一个不可分辨二元关系 (不分明关系) Ind(B): $Ind(B)=\{(x, y)\mid (x, y)\in U\times U$,

收稿日期: 2008-10-20 基金项目: 四川省科技计划项目 (2008GZ0003)资助; 四川省科技厅科技攻关项目 (07GG006-014)资助; 中国科学院人才培养计划项目 ("西部之光")资助. 作者简介: 朱颢东, 男, 1980年生, 博士, 研究方向为软件过程技术与方法; 钟 勇, 男, 1966年生, 博士生导师, 研究员, 高级咨询师, 研究方向软件过程技术与方法.

$\forall b \in B(f(b, X) = f(b, Y))\}$. $Ind(B)$是等价关系, 由这种等价关系导出的对 U的划分记为 $U/Ind(B)$.

定义 **3** 无决策属性的信息系统 $S=(U, C, V, f)$的属性约简问题是: 求 $R \subseteq C$, 使得

$$J = \min|R| \text{ 且 } U/ind(R) = U/ind(C)$$

3 基于集合划分的属性重要性

对于无决策属性的信息系统 $S=(U, C, V, f)$, 设已选择的属性子集 $R \subseteq C$和待选择的条件属性 $c \in C-R$产生的划分为:

$$\pi_R = U/ind(R) = \{X_1, X_2, \cdots, X_r\}$$
$$\pi_c = U/c = \{Y_1, Y_2, \cdots, Y_t\}$$

那么, 两个划分的积划分 π (包含空集)[6]:

$$\pi = \pi_R \pi_c = U/ind(R \cup c) = \begin{pmatrix} E_{11} & E_{12} & \cdots & E_{1t} \\ E_{21} & E_{22} & \cdots & E_{2t} \\ \vdots & \vdots & \vdots & \vdots \\ E_{r1} & E_{r2} & \cdots & E_{rt} \end{pmatrix} \quad (1)$$

其中: $E_{ij} = X_i \cap Y_j$, $i = 1, 2, \cdots, r$; $j = 1, 2, \cdots, t$, 满足 $X_i = \bigcup_{j=1}^{t} E_{ij}$, $Y_j = \bigcup_{i=1}^{r} E_{ij}$.

增加 R集的等价关系, 即加细 R集的划分, 则式(1)中不为空的元素越多, 说明对 R集的加细越多, R集的秩 $rank$(划分的块数)增加越多.

对于信息系统 S中待选择的条件属性 $c \in C-core(C)$, 首先令 $R = core(C)$, $U/ind(R) = \{X_1, X_2, \cdots, X_r\}$, $U/c = \{Y_1, Y_2, \cdots, Y_t\}$, 因此可获得如式(1)所示的 E阵, 对 E阵进行以下化简和定义:

$$E_{ij} = X_i \cap Y_j \Rightarrow E1_{ij} = \begin{cases} 0, & E_{ij} = \varnothing \\ 1, & E_{ij} \neq \varnothing \end{cases},$$
$$\Rightarrow E2_i = \sum_{j=1}^{t} E1_{ij}$$
$$\Rightarrow E3_i = \begin{cases} 0, & E2_i = 1 \\ E2_i, & E2_i > 1 \end{cases}.$$

条件属性集的划分达到论域上的最大划分, 当 $D=\varnothing$时获取系统的约简集, 实际就是在待选择的属性集中, 找出尽可能最多地细分核属性(或已选择的属性子集)的条件属性, 使获得的约简集属性数最少且同样达到最大划分的结果[6]. 如果以划分块为基本信息粒, 则可建立相对于核属性(或当前已选择的属性子集 R)的最大差异度作为属性重要性的启发式规则. 最大差异度以待选择的属性划分与已选择的属性集 R的划分的积划分可否使新的 R集秩产生最大增加(贪婪算法)为计算值.

定义 **4** 属性 c关于 R的最大差异度属性重要性为:

$$sig(c) = \sum_{i=1}^{r} E3_i \quad (2)$$

或在 E阵的基础上, 以函数关系嵌入定义, 属性 c关于 R的最大差异度属性重要性为:

$$sig(c) = \sum_{i=1}^{r} g\left(\sum_{j=1}^{t} f(E_{ij})\right) \quad (3)$$

其中 $f(E_{ij}) = \begin{cases} 0, & E_{ij} = \varnothing \\ 1, & E_{ij} \neq \varnothing \end{cases}$, $g(*) = \begin{cases} 0, & \sum f(E_{ij}) = 1 \\ \sum f(E_{ij}), & \text{其他} \end{cases}$

从定义 4可以看出, 属性 c与 R集的划分差异越大, 重要性越高, 越应首先选取.

4 启发式属性约简算法

算法思想: 算法开始先令 $R=\varnothing$, 这样在确定 R集后, 对 $C-R$中的每一个属性求取差异度属性重要性, 然后根据差异度的大小选取属性, 进行启发式约简. 当 R集选择了一个属性后其不可区分关系发生变化, 因论域上由 R集决定的划分下只有一个对象的划分块不能再被细分, 则可从论域中保留出去, 这样对后续属性的评价减少了干扰, 启发式约简得到优化. 算法用伪码表示如下:

输入: 信息系统 $S=\{U, C, V, f\}$, $C=\{c_1, c_2, \cdots, c_m\}$
输出: 属性约简集 $red(C)$
Step **1**. $R=\varnothing$;
Step **2** 计算 S的 $ind(C)$;
Step **3** 求 S的核属性, R=核属性集;
Step **4** $C=C-R$, IF $C=\varnothing$ THEN $red(C)=R$输出 $red(C)$ STOP; ELSE转 Step5
Step **5** IF $R=\varnothing$ THEN
$R=c_r(c_r=\{c_i | rank(c_i) = max(rank(c_i)), c_i \in C\})$;
Step **6** 计算 $ind(R)$, $U=U-x(x \in U, [x]_R=\{x\})$;
Step **7** 在 U上, 计算 E阵和化简, 依据(3)式计算 $sig(c_i)$, $c_i \in C$;
Step **8** $c_{max}=\{c_i | max\ sig(c_i), c_i \in C\}$, $R=R \cup c_{max}$;
Step **9** IF $ind(R)=ind(C)$ THEN $red(C)=R$输出 $red(C)$ STOP; ELSE Step10
Step **10** $C=C-c_{max}$; 转 Step6

与决策表属性约简比较, 无决策分类时的属性约简, 是以每一个对象为一类的约简所以要求更多的属性加细划分.

该算法的时间复杂性分为两部分, 一是计算系统积划分, 二是属性重要性的定义. 两个计算都包括求等价关系的交. 因两个等价关系交运算的时间复杂性为 $O(|U|^2)$, 若对各个条件属性求解, 最坏情况下时间复杂性为 $O(|C||U|^2)$, 这在一定程度上能够解决无决策属性的信息系统属性约简问题, 进一步扩展粗集理论的应用范围.

5 算法示例

对表 1(见下页)所示的无决策属性 CTR数据库, 依据本文算法进行属性约简计算.

Step **1**-Step **5**. 得: $core(C)=\{c_1, c_2, c_4\}$, $R=core(C)$;
Step **6** 计算 $U/ind(R)$, $U=U-x([x]_R=\{x\})$;
Step **7**. 在 U上, 计算 E阵和化简, 依(3)式计算 $sig(c_i)$, $i=3, 5, 6, 7, 8, 9$, 结果排列为:

c_3	c_5	c_6	c_7	c_8	c_9
5	6	1	2	4	3

Step **8**-Step **10**. $R=R \cup c_6$, 但 R不满足约简集条件, 转 Step6

Step 6 计算 $U/ind(R)$, $U=U-X([x]_R=\{x\})$;

Step 7 在 U上, 依据(3)式计算 $sig(c_i)$, $i=3,5,7,8,9$ 结果排列为:

c_3	c_5	c_7	c_8	c_9
2	5	1	4	3

Step 8-Step 10 $R=R\cup c_7$, R仍不满足约简集条件, 再转 Step 6

Step 6 计算 $U/ind(R)$, $U=U-X([x]_R=\{x\})$;

Step 7 在 U上, 依据(3)式计算 $sig(c_i)$, $i=3,5,8,9$ 结果如下:

c_3	c_5	c_8	c_9
1	3	2	3

Step 8 $c_{max}=c_5$, $R=R\cup c_{max}$;

Step 9 $ind(R)=ind(C)$, $red(C)=R=\{c_1, c_2, c_3, c_4, c_6, c_7\}$ 输出 $red(C)$, STOP.

表 1 无决策属性的 CTR数据库

Table 1 CTR dataset without decision attributes

U	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
1	0	1	1	1	1	1	1	1	0
2	0	1	0	1	1	1	1	0	0
3	0	0	1	1	1	1	1	0	1
4	0	1	0	1	1	0	0	0	0
5	0	1	0	0	1	0	0	1	2
6	0	1	0	1	1	0	1	0	2
7	1	0	0	0	0	1	2	0	1
8	0	0	0	0	0	1	2	0	0
9	0	0	0	0	0	1	0	1	0
10	1	0	0	1	0	1	2	0	1
11	1	0	0	1	1	0	0	0	0
12	0	0	0	0	1	0	0	0	0
13	1	0	1	1	0	1	1	0	0
14	1	0	0	0	0	0	2	0	0
15	0	0	1	1	1	0	1	0	0
16	0	1	0	1	1	0	1	1	0
17	0	0	0	1	1	0	1	1	0
18	1	0	0	1	0	1	0	0	0
19	0	0	0	1	0	1	0	0	0

在最后一轮的属性重要性计算中, 两个属性 c_3 和 c_8 具有相同重要性计算值, 都满足约简需要, 所以实际上得到两个约简结果, 另一个约简集为 $red(C)=\{c_1, c_2, c_4, c_6, c_7, c_8\}$.

6 结 论

本文以粗集属性约简理论为基础, 对无决策属性的信息系统从集合论的论域划分进行研究, 从而得出无决策属性的信息系统在属性特征描述下, 可达到论域的最大划分. 因此在无决策属性的信息系统属性约简中, 对于最小属性子集的选择, 可以使用已选择的属性与待选择的属性相互差异度来进行, 差异度越大, 则属性重要性越高, 那么该属性就越可能被选择, 并且在属性选择的动态过程中, 减少论域中最小划分对属性重要性评价的影响后, 可快速获得约简集. 以此为基础提出了一个适用于无决策属性的信息系统得启发式属性约简算法, 在一定程度上能够解决无决策属性的信息系统属性约简问题, 进一步扩展粗集理论的应用范围.

References:

[1] Pawlak Z. Rough sets[J]. International Journal of Computer & Information Sciences, 1982, 11: 341-356.

[2] Pawlak Z, Grzymala-Busse J, Slowinski R, et al. Rough sets[J]. Communications of the ACM, 1995, 38(11): 89-95.

[3] Pawlak Zdzislaw. Rough classification[J]. International Journal of Human-Computer Studies, 1999, 51(2): 369-383.

[4] Liu Qing-zhen, Cai Jin-ding, Wang Shao-fang. Fault diagnosis of power electronic circuits based on rough set neural network system[J]. Electric Power Automation Equipment, April 2004, 24(4): 45-48.

[5] Wang Guo-yin. Rough set theory and knowledge acquisition[M]. Xi'an: Xi'an Jiaotong University Press, 2001.

[6] Ni Zi-wei, Cai Jing-qiu. Discrete mathematics[M]. Beijing: Science Press, 2002, 10.

附中文参考文献:

[5] 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.

[6] 倪子伟, 蔡经球. 离散数学[M]. 北京: 科学出版社, 2002, 10.