文章编号:1672-3961(2008)02-0112-05

关系积理论及属性约简算法

焦吉成[1,2], 高学东[1], 王元璞[2], 赵传领[2]
(1. 北京科技大学管理学院, 北京 100083;
2. 济南钢铁集团总公司技术中心, 山东 济南 250101)

摘要:属性约简是粗糙集RS(rough set)理论的重要研究内容.决策表的最小属性约简是NP-hard问题.本文基于集合理论,提出了关系积概念,把决策表的属性约简过程转化为关系积的运算,充分利用关系积的相关性质,提高了关系积属性约简算法的效率.

关键词:集合;关系积;属性;粗糙集
中图分类号:TP182 **文献标志码**:A

Study of the attribute union theory and attribute reduction algorithm

JIAO Ji-cheng[1,2], GAO Xue-dong[1], WANG Yuan-pu[2], ZHAO Chuan-ling[2]
(1. Management School, University of Science and Technology Beijing, Beijing 100083, China;
2. Jinan Iron and Steel Group Corporation, Jinan 250101, China)

Abstract: Attribute reduction in rough set is the key content of rough set theory. It is a NP-hard problem to get the minimal attribute sets. The attribute union was presented based on the set theory, the attribute reduced procedure was translated to find the attribute union, and the reduced efficiency was improved.

Key words: set theory; attribute union; attribute; rough set;

0 引言

粗糙集理论的主要思想是在保持分类能力不变的前提下,通过属性约简,导出问题的决策或分类规则.应用粗糙集理论处理不确定性问题的最显著特点是不需提供问题所需处理的数据集合之外的任何先验信息.属性集的约简(attribute reduction)是粗集理论中关键的问题之一.所谓约简是属性集的子集,它与原属性集具有同样的分辨能力.约简反映了一个信息系统的本质信息,求解一个信息系统的全部约简或计算出最佳约简都是NP难题.当数据量很大时,应用粗糙集理论算法十分耗时,甚至不可行.

现有的属性约简算法分为3类:

(1) Pawlak约简算法[1]

这种方法按照约简的定义进行求解,需要对条件属性集的幂集中的所有元素进行考察,每次考察都需对决策表进行扫描,该算法的理论指导意义大于其实际应用效果,由于其计算速度慢,且不易于计算机实现,故实际应用的局限性较大.

(2) Skowron[2]算法(也称可辩识矩阵和逻辑运算的约简算法)

收稿日期:2007-01-29
基金项目:中国博士后科学基金资助项目(2005038319)
作者简介:焦吉成(1968-),男,山西省广灵县人,北京科技大学博士研究生,高级工程师,主要从事数据挖掘、计算机仿真研究.
E-mail: jim_jiao0805@sina.com

该算法是 Skowron 教授于 1992 年提出.首先根据信息系统构造一个相关的可辩识矩阵;其次,利用可辩识矩阵中的非空元素构造区分函数(析取式);最后,把析取式转化为合取式,求解区分函数,每一个合取子式对应一种约简,该算法可获得信息系统的所有约简.该算法实际上是将对属性组合情况的搜索演变成逻辑公式的化简,从而简化了问题.其缺点在于:① 对于大规模的信息系统,该算法需存储一个较大的区分矩阵,占用了大量的计算机内存;② 区分函数的求解是一个组合问题,会出现组合爆炸问题,计算过程中数据溢出现象严重. 因此,该算法在处理海量信息系统的约简问题上不是非常有效的.

(3) 各种启发式算法

根据属性重要度、信息熵或可辩识矩阵中属性出现次数等启发信息来寻求信息系统的约简. 如文献【3】把区分矩阵中属性出现次数作为启发信息;文献【4】是基于信息熵的遗传顺序约简算法,文献【5】把属性重要度作为启发式信息等等. 这类算法主要优点是采用多项式时间进行求解,且可以对大规模数据集进行处理.启发式算法的缺点在于利用这类算法所求得的约简不能保证是最小属性约简,有些算法所求得的约简甚至是不完备的.

1 粗糙集及关系积的基本概念

下面首先给出粗糙集理论的基本概念:

(1) 知识表达系统又称为信息系统,可表示为:$S=\langle U,A,V,f\rangle$.其中 U 为对象的非空有限集合;A 为属性的非空有限集合;V 为属性的值域集;f 为信息函数,$f:U\times A\rightarrow V$,如果 $A=C\cup D$,$C\cap D=\Phi$,C 为条件属性集,D 为决策属性集,则把信息系统 $S=\langle U,A,V,f\rangle$称为决策系统,用 $S=\langle U,C\cup\{d\}\rangle$或 $S=\langle U,C\cup D\rangle$来表示,其中 d 为单一的决策属性.从数据库的角度来看,决策系统就是一张表,其中 U 是记录集合,A 是字段集合,每一个对象对应一条纪录,这样,决策系统又可称为决策表.

(2) 在决策系统 $S=\langle U,C\cup\{d\}\rangle$中,对于 $B\subseteq C$,则 B 在 U 上的不可分辨关系定义为 $IND(B)=\{(x,y)\mid(x,y)\in U^2,\forall b\in B(b(x)=b(y))\}$,$IND(B)$把 U 划分为 K 个等价类 $X_1,X_2,\cdots X_k$,$U\mid IND(B)$表示等价关系 $IND(B)$的所有等价类组成的等价类簇,即有 $U\mid IND(B):\{X_1,X_2,\cdots,X_k\}$.

(3) 对于对象集 $X\subseteq U$,属性集 $B\subseteq A$,X 的下近似 $B_-(x)$和上近似 $B^-(x)$分别定义为 $B_-(x)=\bigcup\{Y_i\mid(Y_i\in U\mid IND(B)\wedge Y_i\subseteq X)\}$和 $B^-(x)=\bigcup\{Y_i\mid(Y_i\in U\mid IND(B)\wedge Y_i\cap X\neq\Phi)\}$.

(4) 在信息系统 $S=\langle U,C\cup\{d\}\rangle$中,对属性集 C 的子集 B,属性 $a\in B\subseteq C$ 是 B 中必要的,当且仅当 $IND(B)\neq IND(B-\{a\})$,否则,属性 a 在 B 中是冗余或可省略的.属性集 B 的约简是一个集合 $B'\subseteq B$,当且仅当满足:① B 是独立的;② $IND(B')=IND(B)$.属性集 $B\subseteq C$ 的所有约简簇的交集称为属性集 B 的核,记为 $CORE(B)$,有 $CORE(B)=\bigcap RED(B)$.

为了实现属性约简,获得最小的属性集(规则集),仍从集合划分的角度出必,提出如下 3 个关键定义:

定义 1 在决策系统 $S=\langle U,C\cup\{d\}\rangle$中,设 $P(B_1)$和 $P(B_2)$分别为$B_1\subseteq C$ 和$B_2\subseteq C$ 对 U 导出的等价类,则由 $B_{12}=B_1\cap B_2$ 所导出的等价类 $P(B_1\cap B_2)$称为 $P(B_1)$和 $P(B_2)$的关系积,记为 $P(B_1\cdot B_2)$,如图 1 所示.

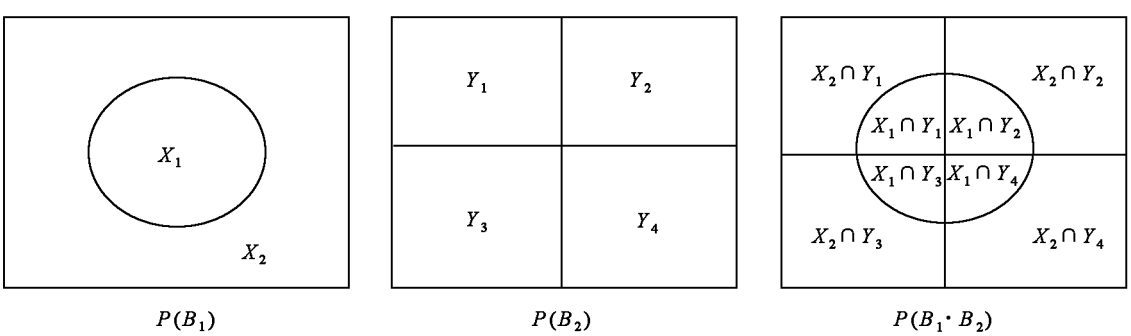

图 1 关系积概念图

Fig.1 Concept of attribute union

关系积概念的实质是两种划分对集合的联合划分,与集合的交不同.集合的交取的是两个集合的公共部分.而关系积只是对集合的一次重新划分.

定义 2　在决策系统 $S=\langle U, C\cup\{d\}\rangle$中,条件属性 $C=\{B_1, B_2, \cdots B_m\}$, $P(B_1\cdot B_2)$称为二元关系积, $P(B_1\cdot B_2\cdot B_3)$称为三元关系积, $P(B_1\cdot B_2\cdot\cdots\cdot B_m)$称为 m 元关系积.

定义 3　在决策系统 $S=\langle U, C\cup\{d\}\rangle$中,令 Q = 决策属性集 = $\{d\}$,条件属性 $C=\{B_1, B_2, \cdots, B_m\}$,如果 $POS_c(Q)=U$,则称论域 U 是 C 上相对于 Q 一致的.也即决策表是确定的决策表,决策表中不包含不一致的信息(样本).

对于不确定的决策表,根据决策表产生不确定性的原因,对决策表进行相关的处理.限于篇幅,转换的有关方法请参见相关文献,这里不展开讨论.

根据以上概念,下面给出以下几个相关定理:

定理 1　关系积运算满足交换率、结合率、分配率等集合运算.

证明　根据集合的有关性质,定理显然成立.

定理 1 可以降低属性组合的计算数量和利用次级关系积生成新关系积,如:现有二元关系积 $P(B_1\cdot B_2)$和 $P(B_2\cdot B_3)$,则可以通过两者的关系积运算,生成 $P(B_1\cdot B_2\cdot B_3)$,而不必重新计算 $P(B_1\cdot B_2)$与 $P(B_3)$的关系积.

定理 2　如果某元关系积的任一元素是决策属性集的子集,则可把该元素删除.

证明　不失一般性,任取 i 元关系积,设 $S(i,j)=P(B_1\cdot B_2\cdot\cdots\cdot B_i)(j)$为 i 元关系积的第 j 个元素(子集),如果 $S(i,j)\cap P(d)=S(i,j)$,由于 $S(i,j)=\bigcup_{k=1}^{m}\bar{s}(i,j,k)$,因此 $S(i,j)$的任何幂子集仍然属于同一决策属性子集,对该元素(子集)的进一步划分已没有必要,可以不考虑该子集.所以可以把 i 元关系积中该元素删除.

推论 1　由于单元素子集一定是决策属性集的子集,所以某元关系积的单元素子集可以直接删除.

定理 3　如果某元关系积通过定理 2 运算后为空集,则该元关系积就是一个属性约简.

证明　根据定理 2 可知,删除的子集是决策属性的子集,如关系积为空,则可知该元关系积的子集一定全部包含在决策属性集中,即: $POS_c(Q)=U$,则该元关系积就是一个属性约简.

定理 4　定理 3 是决策表最小属性约简的必要条件,而不是充分条件.

证明　最小约简是所有约简中属性组合数目最小的属性组合,是决策表约简的一个特例.因此,定理 3 只是最小属性约简的必要条件,而不是充分条件.

2　关系积约简算法

Pawlak 约简算法是从全部的条件属性集开始,逐步去掉对决策属性不必要的条件属性获得最小约简属性的,可以称为一种自顶向下的约简算法.这种方法每去掉一个属性,需要对决策表重新进行组合,生成新的决策表,根据新的决策表判断是否获得了最小约简.因此,需要反复地对决策表进行操作,占用大量的计算资源,也降低了算法的效率.事实上,完全可以根据决策表提供的各属性等价关系及其它们的关系积对属性进行约简,从而获得所有最小约简属性.由于高阶关系积可由次阶关系积或次阶关系积与一阶关系积的集合运算生成,不需要对决策表重新进行组织和扫描,把对表的扫描转化为集合的的运算,提高了算法的效率.利用定理 1 和 2 及推论可以大幅度地减少关系积运算次数.

算法　URedAttrBUU(upgrade reducing attribute based on union)

输入:决策系统 $S=\langle U, C\cup\{d\}\rangle$, C 为条件属性集, $\{d\}$为单一决策属性集.

输出:条件属性集 C 的约简及核.

算法步骤:

Step 0　检验决策表是否是确定的决策表,如不是,退出.

Step 1　对条件属性集中的所有属性和决策属性,计算一元关系积,即: $P(B_i)(i=1,2,\cdots,m)$和 $P(d)$,令 $CacNo=1$; $C=\{\}$;

Step 2 检查一元关系积的子集是否包含在决策属性集中,如成立,则可在一元关系积中删除该子集;

Step 3 根据定理 3 检验是否获得属性的最小约简,如果获得最小约简,则转 Step 6;

Step 4 $CacNo = CacNo + 1$;

Step 5 while($CacNo$〈〉属性集数目) do

Begin

//对属性集中所有属性,计算 CacNo 元关系积.

for $i = 1$ to Card($P(B_{CacNo-1})$) do

for $j = i$ to Card ($P(B_1)$ do

$P(B_{CacNo}) = P(B_{CacNo-1})(i) \cap P(B_1(j)$;

//根据定理 2 和推论 1,检查 $CacNo$ 元关系积的子集是否包含在决策属性集中,如成立,则可在 $CacNo$ 元关系积中删除该子集.

for $i = 1$ to Card($P(B_{CacNo})$) do

If (Card($P(B_{CacNo}(i))) = 1$)or($P(B_{CacNo}(i)) \cap P(d) = P(B_{CacNo}(i))$ then

$P(B_{CacNo}) = P(B_{CacNo})\text{-}P(B_{CacNo}(i))$;

//检查是否获得约简集,如是,则记录下约简集属性,寻找过程结束.

for $i = 1$ to Card($P(B_{CacNo})$) do

If(Card($P(B_{CacNo}(i))) = 0$) then 找到一个最小约简, $C = C + B_{CacNo}(i)$;

If (Card(C)〈〉0) then goto Step 6 else $CacNo = CacNo + 1$;

End;

Step 6 输出 $CacNo$ 元关系积的约简属性.

Step 7 输出核 $CORE_Q(D) = \bigcap B_{CacNo}(i)$

其中,Card 函数用于计算各划分的集合数量.由于本算法采用自底向上搜索策略,如果在 K 元关系积上找到一个属性约简,则没有必要搜索 K 元关系积以上的约简.对同元关系积进行全部计算后,才检验是否找到最小约简,所以本算法可以找到所有的最小属性约简集.

3 算法示例

结合文献【2】提供的一个关于气象信息的决策表(表 1),说明以上约简算法.

表 1 气象信息的决策表

Table 1 Decision table of weather information

U	条件属性				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

令 Q = 决策属性集 = $\{d\}$, C = 条件属性全集 = $\{a_1, a_2, a_3, a_4\}$,则

$$P(C) = \{(1),(2),(3),(4),(5),(6),(7),(8),(9),(10),(11),(12),(13),(14)\};$$

$$P(Q) = \{(1,2,6,8,14),(3,4,5,7,9,10,11,12,13)\};$$

$$POS_C(Q) = U.$$

各条件属性的等价关系如下:

$$P(a_1) = \{(1,2,8,9,11),(3,7,12,13),(4,5,6,10,14)\};$$

$$P(a_2) = \{(1,2,3,13),(4,8,10,11,12,14),(5,6,7,9)\};$$

$$P(a_3) = \{(1,2,3,4,8,12,14),(5,6,7,9,10,11,13)\};$$

$$P(a_4) = \{(1,3,4,5,8,9,10,13),(2,6,7,11,12,14)\}.$$

根据定理 2 和定理 3 知:$P(a_1) = \{(1,2,8,9,11),(4,5,6,10,14)\}$, $P(a_2)$, $P(a_3)$和 $P(a_4)$不变.

检查:因为各属性关系积元素不为空,故一阶关系积不能构成最小约简.

下面计算二阶关系积:

$$P(a_1 \cdot a_2) = \{(1,2),(8,11),(9),(3,13),(12),(7),(4,10,14),(5,6)\};$$

$$P(a_1 \cdot a_3) = \{(1,2,8),(9,11),(3,12),(7,13),(4,14),(5,6,10)\};$$

$$P(a_1 \cdot a_4) = \{(1,8,9),(2,11),(3,13),(7,12),(4,5,10),(6,14)\};$$

$$P(a_2 \cdot a_3) = \{(1,2,3),(13),(4,8,12,14),(10,11),(5,6,7,9)\};$$

$$P(a_2 \cdot a_4) = \{(1,3,13),(2),(4,8,10),(11,12,14),(5,9),(6,7)\};$$

$$P(a_3 \cdot a_4) = \{(1,3,4,8),(2,12,14),(5,9,10,13),(6,7,11)\}.$$

根据定理 2 和推论 1 及定理 3,分别对各关系积进行约简,化简后的二元关系积为

$$P(a_1 \cdot a_2) = \{(8,11),(4,10,14),(5,6)\};$$

$$P(a_1 \cdot a_3) = \{(4,14),(5,6,10)\};$$

$$P(a_1 \cdot a_4) = \{(1,8,9),(2,11)\};$$

$$P(a_2 \cdot a_3) = \{(1,2,3),(4,8,12,14),(5,6,7,9)\};$$

$$P(a_2 \cdot a_4) = \{(1,3,13),(4,8,10),(11,12,14),(6,7)\};$$

$$P(a_3 \cdot a_4) = \{(1,3,4,8),(2,12,14),(6,7,11)\}.$$

由于所有二阶关系积元素不为空,也不构成最小约简.

同理计算三元关系积及其化简,化简后的三元关系积如下:

$$P(a_1 \cdot a_2 \cdot a_3) = \{(4,14),(5,6)\};$$

$$P(a_1 \cdot a_2 \cdot a_4) = \{\};$$

$$P(a_1 \cdot a_4 \cdot a_3) = \{\};$$

$$P(a_4 \cdot a_2 \cdot a_3) = \{\{1,3\},\{4,8\},\{12,14\},\{6,7\}\}.$$

由于 $P(a_1 \cdot a_2 \cdot a_4) = \{\}$, $P(a_1 \cdot a_4 \cdot a_3) = \{\}$,故属性 a_1, a_2, a_4 和 a_1, a_3, a_4 构成了决策表的最小约简集,其核为

$$CORE_Q(P) = \bigcap RED_Q(P) = \{a_1, a_2, a_4\} \bigcap \{a_1, a_3, a_4\} = \{a_1, a_4\}.$$

4　结束语

决策表的属性约简计算过程需对决策表不断的重组和扫描,计算过程对计算机的资源占用量大.本文从关系积的概念出发,把对决策表的扫描转化为集合的运算,只对决策表扫描一次,就可生成所有的最小约简属性集,避免了对决策表的频繁操作.但在生成各阶关系积的时候,仍存在属性组合的爆炸问题,本文充分利用关系积的一些特殊性质,降低了算法的复杂性,获得了比较满意的求解速度.当然,可以根据关系积的特殊性质,提出一些启发式算法,从而降低关系积运算的次数,这是本文下一步的研究重点.

(下转第 126 页)

(上接第116页)

参考文献:

[1] 王国胤.Rough集理论与知识获取[M].西安:西安交通科技大学出版社,2001.
WANG Guo-yin. Rough set theory and knowledge acquisition[M]. Xi'an: Xi'an Jiaotong University Press, 2001.

[2] SKOWRON A, POLKOWSKI L. Synthesis of decision systems from data tables[M]// LIN T Y, CERCONE N. Rough sets and data mining: analysis for imprecise data. Boston: Kluwer Academic Publishers, 1997: 259-300.

[3] HU K Y, DIAO L L, LU Y C, et al. A heuristic optimal reduct algorithm[C]// Appeard in Proceedings of 2nd International Conference on Intelligent Data Engineering and Automated Learning. Hongkong:[s.n.], 2000: 13-15.

[4] DOMINIK S, JAKUB W. Order based genetic algorithms for the search of approximate entropy reducts[C]// WANG G. RSFDGrC 2003, LNAI 2639. Berlin, Heidelberg: Springer-Verlag, 2003: 308-311.

[5] 李克文,吴孟达,张雄明.约简的一种启发式算法[J].计算机工程与科学,2004,26(1):92-94.
LI Ke-wen, WU Meng-da, ZHANG Xiong-ming. A heuristic algorithm for reduction[J]. Computer Engineering & Science, 2004, 26(1):92-94.

(编辑:许力琴)