基于幂图的属性约简搜索式算法

陈玉明[1),2)] 苗夺谦[1),2)]

[1)](同济大学计算机科学与技术系 上海 201804)

[2)](嵌入式系统与服务计算教育部重点实验室 上海 201804)

摘 要 粗糙集理论是一种新的处理不精确、不完全与不一致数据的数学工具.属性约简是粗糙集理论的重要研究内容之一,已有的属性约简算法主要是基于代数表示与信息表示的方法.同一问题在不同的知识表示下,其求解难度是不同的.文中从改变属性约简问题的知识表示入手,提出了该问题的一种新的表示方式——幂图;给出了基于幂图的属性约简搜索式算法,把属性约简计算问题转化为在幂图中的搜索问题.理论分析表明新算法是有效的,为属性约简研究提供了一条新的途径.

关键词 粗糙集;属性约简;幂图;粒计算;知识表示

中图法分类号 TP18 **DOI 号**: 10.3724/SP.J.1016.2009.01486

Searching Algorithm for Attribute Reduction Based on Power Graph

CHEN Yu-Ming[1),2)] MIAO Duo-Qian[1),2)]

[1)](*Department of Computer Science and Technology, Tongji University, Shanghai* 201804)

[2)](*The Key Laboratory of Embedded System and Service Computing of Ministry of Education, Shanghai* 201804)

Abstract Rough set theory is a new mathematical tool to deal with imprecise, incomplete and inconsistent data. Attribute reduction is one of important issues in rough sets. Most existing algorithms are studied under both algebra and information representations. As problem solving under different knowledge representations corresponding to different difficulties, the new knowledge representation, called power graph, is presented in this paper. Searching algorithms based on power graph are also proposed, which can translate computing problem of attribute reduction into searching problem in power graph. The algorithms will provide a new method in attribute reduction and the efficiency of the method has been proved in theoretical analysis.

Keywords rough sets; attribute reduction; power graph; granular computing; knowledge representation

1 引 言

波兰科学家 Pawlak 1982 年提出的粗糙集(rough sets)是一种新的处理不精确、不完全与不一致数据的数学理论[1].近年来该理论在机器学习、数据挖掘及模式识别等多个领域得到了广泛的应用[2-3].在基于粗糙集理论的知识获取研究中,属性约简是其中最核心的组成部分之一,许多学者已对属性约简算法进行了大量的研究[4-11].现有的属性约简研究主要是基于代数表示与信息表示的方法[4].在代数表示下,大体上可分为基于正区域的属性约简算法[5-6]、基于差别矩阵及在此基础上改进的属性约简算法[7-8]等.在信息表示下,主要有基于信

收稿日期:2007-11-05;最终修改稿收到日期:2009-06-10.本课题得到国家自然科学基金(60475019,60775036)、博士学科点专项科研基金(20060247039)资助.**陈玉明**,男,1977 年生,博士研究生,讲师,主要研究方向为粗糙集理论与模式识别等.E-mail: cym0620@163.com.
苗夺谦,男,1964 年生,教授,博士生导师,主要研究领域为粗糙集理论、粒计算、Web 智能与模式识别等.

息熵的属性约简算法[9]、基于互信息的决策表约简算法[10]和基于条件信息熵的决策表约简算法[11]等.但这些算法都不是基于图搜索式的算法.

知识表示方式研究是人工智能研究的中心内容之一.对于传统人工智能问题,任何比较复杂的求解技术都离不开两方面的内容——表示与搜索.知识表示方法很多,有图示法、公式法、结构化方法、陈述式表示和过程式表示等[12].德国数学家 Wille 于1982年首先提出形式概念分析(formal concept analysis)用于概念的发现、排序和显示[13].形式概念分析是一种知识表示方式,国内外学者进行了大量的研究,把形式概念分析应用于属性约简之中[14-15].但形式概念分析理论使知识表示复杂化,不利于问题的求解.因此,有必要研究更简单直观的知识表示方式.

图是一种直观形象的知识表示方式,在此基础上有宽度优先搜索、深度优先搜索和启发式搜索等算法.图搜索的优势在于带有回溯机制,保存了搜索路径,因而可以根据已付出的代价,搜索下一步的最佳路径.属性约简研究包括信息系统的属性约简和决策表的属性约简,本文主要研究信息系统的属性约简.根据属性的子集之间的包含关系,本文提出一种新的知识表示方式——幂图.把该知识表示方式应用于属性约简之中,给出两种基于幂图的属性约简搜索式算法,把属性约简计算问题转化为在幂图中的搜索问题.理论分析表明,本文提出的算法是有效的.

2　基本概念及幂图

定义 1[16].　称 $IS=(U,A,V,f)$ 为信息系统,其中 U 是非空有限集,称为论域;A 是有限属性集;$V=\bigcup_{a\in A}V_a$,V_a 表示属性 a 的值域;$f:U\times A\to V$ 是一个信息函数,即对 $\forall x\in U,a\in A$,有 $f(x,a)\in V_a$.任一属性子集 $B\subseteq A$ 决定了一个二元不可区分关系 $IND(B)$:$IND(B)=\{(x,y)\in U\times U\mid \forall a\in B, f(x,a)=f(y,a)\}$.$U/IND(B)$构成了 U 的一个划分,称其为 U 上的一个知识,其中每个等价类称为一个知识粒.为方便计,有时将 $U/IND(B)$ 简记为 U/B.

特别地,如果 $U/B=\omega=\{[x]_B\mid [x]_B=\{x\},x\in U\}$,称其为恒等关系;如果 $U/B=\delta=\{[x]_B\mid [x]_B=U,x\in U\}$,称其为全域关系.

由于等价关系、属性、知识、划分等概念之间是等价的,因此在以下的论述中,将不再对其进行区分.

定义 2[16].　设 $IS=(U,A,V,f)$ 为信息系统,$U/A=\{X_1,X_2,\cdots,X_m\}$,则 A 的知识粒度定义为

$$GD(A)=\sum_{i=1}^{m}\frac{|X_i|^2}{|U|^2}.$$

性质 1.　$\forall B\subseteq A$,有 $\frac{1}{|U|}\leq GD(B)\leq 1$.

当 U/B 为恒等关系时,即 $U/B=\omega$,则 B 的知识粒度达到最小值 $\frac{|U|}{|U|^2}=\frac{1}{|U|}$;当 U/B 为全域关系时,即 $U/B=\delta$,则 B 的知识粒度达到最大值 $\frac{|U|^2}{|U|^2}=1$.

命题 1.　设 $IS=(U,A,V,f)$ 是一个信息系统,$B,C\subseteq A$,则有

(1) 若 $B\supset C$,则 $GD(B)\leq GD(C)$;

(2) 若 $B\subset C$,则 $GD(B)\geq GD(C)$.

证明.　参见文献[16].

定义 3.　设 $IS=(U,A,V,f)$ 是一个信息系统,$\forall a\in A$,如果 $IND(A-\{a\})=IND(A)$,则称 a 是 A 中不必要的(多余的)属性;否则,称 a 是 A 中必要的属性.

定义 4.　设 $IS=(U,A,V,f)$ 是一个信息系统,属性集 A 中所有必要的属性组成的集合,称为属性集 A 的核,记作 $Core(A)$.

定义 5.　设 $IS=(U,A,V,f)$ 是一个信息系统,$B\subseteq A$ 是一个属性子集,如果满足

(1) $IND(B)=IND(A)$;

(2) 对 $\forall b\in B$,有 $IND(B-\{b\})\neq IND(B)$;

则称 B 是 A 的一个约简.

命题 2.　设 $IS=(U,A,V,f)$ 是一个信息系统,$B,C\subseteq A$ 是两个属性子集,若 $IND(B)=IND(C)$,则 $GD(B)=GD(C)$.

证明.　令 $U/B=\{X_1,X_2,\cdots,X_m\}$,$U/C=\{Y_1,Y_2,\cdots,Y_n\}$,已知 $IND(B)=IND(C)$,所以 $U/B=U/C$,即 $\{X_1,X_2,\cdots,X_m\}=\{Y_1,Y_2,\cdots,Y_n\}$;由知识粒度定义可知,$GD(B)=\sum_{i=1}^{m}\frac{|X_i|^2}{|U|^2}$,$GD(C)=\sum_{j=1}^{n}\frac{|Y_j|^2}{|U|^2}$,而 $\{X_1,X_2,\cdots,X_m\}=\{Y_1,Y_2,\cdots,Y_n\}$,所以 $GD(B)=GD(C)$.　证毕.

注.　该命题的逆未必成立.

命题 3.　设 $IS=(U,A,V,f)$ 是一个信息系统,$B,C\subseteq A$ 是两个属性子集,且 $B\subseteq C$,若 $GD(B)=GD(C)$,则 $IND(B)=IND(C)$.

证明. 反证法. 假设 $IND(B)=IND(C)$ 不成立,即 $IND(B)\neq IND(C)$;由 $B\subseteq C$,则 $IND(B)\supseteq IND(C)$,因为 $IND(B)\neq IND(C)$,所以 $IND(B)\supset IND(C)$;令 $U/B=\{X_1,X_2,\cdots,X_m\}$,$U/C=\{Y_1,Y_2,\cdots,Y_n\}$,由 $IND(B)\supset IND(C)$,则 $\{X_1,X_2,\cdots,X_m\}\supset\{Y_1,Y_2,\cdots,Y_n\}$,所以 $\sum_{i=1}^{m}\frac{|X_i|^2}{|U|^2}>\sum_{j=1}^{n}\frac{|Y_j|^2}{|U|^2}$,由知识粒度定义可知,$GD(B)=\sum_{i=1}^{m}\frac{|X_i|^2}{|U|^2}$,$GD(C)=\sum_{j=1}^{n}\frac{|Y_j|^2}{|U|^2}$,所以 $GD(B)>GD(C)$,这和已知 $GD(B)=GD(C)$ 相矛盾. 故命题得证. 证毕.

命题 4. 设 $IS=(U,A,V,f)$ 是一个信息系统,$\forall a\in A$ 在 A 中是不必要的(多余的)属性,其充分必要条件是 $GD(A-\{a\})=GD(A)$.

证明. 必要性. 设 $\forall a\in A$ 在 A 中是不必要的属性,由定义 3 知 $IND(A-\{a\})=IND(A)$ 成立;由命题 2 可知,$GD(A-\{a\})=GD(A)$.

充分性. 设 $\forall a\in A$,由 $GD(A-\{a\})=GD(A)$ 和命题 3 可知,$IND(A-\{a\})=IND(A)$ 成立,故 $\forall a\in A$ 在 A 中是不必要的属性. 证毕.

推论 1. $\forall a\in A$ 在 A 中是必要的属性,即 $IND(A-\{a\})\neq IND(A)$,其充分必要条件是 $GD(A-\{a\})\neq GD(A)$.

推论 2. $\forall a\in A$,且 $GD(A-\{a\})\neq GD(A)$,则 $Core(A)=\{\bigcup a\}$.

命题 5. 设 $IS=(U,A,V,f)$ 是一个信息系统,$B\subseteq A$ 是 A 的一个约简的充分必要条件为

(1) $GD(B)=GD(A)$.

(2) 对 $\forall b\in B$,有 $GD(B-\{b\})\neq GD(B)$.

证明. 由定义 5、命题 2、命题 3 及推论 1 可以证明. 证明略.

定义 6. 设 $Power(A)$ 为属性集合 A 的幂集,给定有向图 G,G 的顶点为 $Power(A)$ 的元素,G 的边满足条件:$\forall B,C,D\in Power(A)$,若 $Card(B)-1=Card(C)=Card(D)+1$ 且 $(B\cap D)\subset C\subset(B\cup D)$,则存在 D 到 C,C 到 B 的有向边,称此有向图 G 为 A 的幂图.

设属性集合 A 有 m 个元素,则其幂集有 2^m 个元素,根据幂图定义易知 A 的幂图有 2^m 个顶点,对于任意顶点 $V\in Power(A)$,根据 $Card(V)$ 的大小可将幂图分成 $m+1$ 层,设分别为 $L_0,L_1,\cdots,L_m$,具有相同 $Card(V)$ 为同一层,$Card(V)=m$ 的顶点组成第 L_0 层,$Card(V)=m-1$ 的顶点组成第 L_1 层,$\cdots$,$Card(V)=0$ 的顶点组成第 L_m 层.

例 1. 属性集 $A=\{a,b,c\}$,其幂图为图 1 所示.

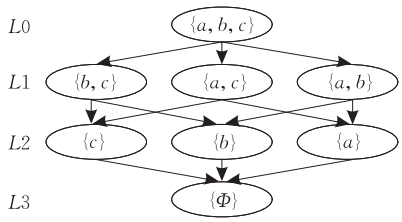

图 1 幂图

定义 7. 设 G 为属性集 A 的幂图,对 G 的顶点结点进行扩充,使顶点结点保存两个量,一个是顶点 $V\in Power(A)$,另一个是其知识粒度 $GD(V)$,称此扩充的幂图为知识粒度幂图.

推论 3. 幂图中结点子集的知识粒度具有反单调性.

证明. 由命题 1 与知识粒度幂图的定义可以推论出来. 证明略.

命题 6. 设 $IS=(U,A,V,f)$ 是一个信息系统,G 为其知识粒度幂图,$K_m\in G$ 为 G 中 L_m 层某结点,且 $GD(K_m)=GD(A)$,如果 K_m 的所有邻接下层 L_{m+1} 层结点的知识粒度都大于 $GD(K_m)$,则 K_m 为该信息系统的约简.

证明. 反证法. 假设 K_m 不是约简,由 $GD(K_m)=GD(A)$ 和命题 3,可知 $IND(K_m)=IND(A)$,所以约简必在 K_m 的真子集中;因为 K_m 的所有邻接下层 L_{m+1} 层结点的知识粒度都大于 $GD(K_m)$,根据 K_m 子集知识粒度的反单调性,知 K_m 真子集的知识粒度都大于 $GD(K_m)$,因为 $GD(K_m)=GD(A)$,所以 K_m 真子集的知识粒度都大于 $GD(A)$,即,$\forall k\in\{K_m$ 真子集$\}$,有 $GD(k)>GD(A)$,由推论 1,可知 $IND(k))\neq IND(A)$,所以约简必不在 K_m 的真子集中,这与前面推出约简必在 K_m 的真子集中相矛盾. 命题得证. 证毕.

命题 1 说明对于 A 的属性子集,随着属性的减少,知识粒度增大,越来越粗;反之,随着属性的增多,知识粒度减小,越来越细. 幂图反映属性的增加与减少关系,知识粒度反映属性子集的划分即知识的粗细程度,知识粒度幂图则把两者结合起来,既反映属性子集划分的粗细程度,又体现其空间拓扑结构. 命题 6 给出了幂图中约简的判定. 所以,可以把属性约简问题转化为在知识粒度幂图中的搜索问题.

3　基于幂图的属性约简搜索式算法

知识表示方法是问题求解所必需的.表示问题是为了进一步解决问题.从问题表示到问题的解决,有一个求解过程,也就是搜索过程.在这一过程中,采用适当的搜索技术,包括各种规则、过程和算法等推理技术,力求找到问题的解答.图搜索技术是一种在图中寻找路径的方法,从初始结点出发到目标结点寻求满足一定要求的路径,可以是最佳路径,也可以是用户要求的最佳路径.求解数据约简问题实质上是个搜索问题.幂图是一种图的知识表示方式,采用这种表示方式,可以非常方便地给出求解数据约简问题的图搜索式算法.属性约简可以从两个方向进行搜索,一个是 Top-down 搜索,另一个是 Bottom-up 搜索.Top-down 方式搜索是从幂图的顶层出发,往下搜索,知识粒度由细到粗,直到找到约简为止,Bottom-up 搜索是从核出发,先求核,然后在幂图中逆向搜索,知识粒度由粗到细,直到找到约简为止.下面给出这两种搜索方式的一般算法.

算法 1.　TSA(Top-down Searching Algorithm of attribute reduction based on power graph).

输入:信息系统 $IS=(U,A,V,f)$

输出:某个约简或者所有约简

1. 建立一个只含有起始节点 A 的搜索图 G,把 A 放在一个叫 OPEN 的未扩展节点表中,计算起始节点 A 的知识粒度 $g=GD(A)$.

2. 建立一个叫 CLOSED 的已扩展节点表,其初始为空表.

3. LOOP:若 OPEN 表是空表,则退出.

4. 选择 OPEN 表上的第一个节点,把它从 OPEN 表移出并放进 CLOSED 表中,称此节点为节点 n.

5. 按幂图扩展节点 n,同时生成后继节点的临时集合 M.把 M 的这些成员作为 n 的后继节点添入图 G 中.

6. 对那些未曾在 G 中出现过的(即未曾在 OPEN 表或 CLOSED 表上出现过的,在和 M 同层的节点中找)M 成员,设置一个通向 n 的指针,计算这些成员的知识粒度,把大于 g 的加进 CLOSED 表(大于 g 的节点,其后继节点不存在约简,不必扩展),其它的加进 OPEN 表.

7. 检测节点 n 是否为约简(检测方法按命题 6,即集合 M 的成员的知识粒度全部大于 g,则 n 为约简).若是,则找到某个约简,输出约简;若不是,则转步 9.

8. 问是否要找下一个约简?若不是,则退出.

9. 按某一任意方式或按某个启发信息重排 OPEN 表.

10. GOLOOP.

以上搜索过程可用如图 2 所示的算法流程图来表示.这个算法一般包括各种各样具体的从上到下求约简的图搜索式算法.此算法生成一个明确的图 G(称为搜索图)和 G 的一个子集 T(称为搜索树),搜索树由步 6 设置的指针来确定.步 9 对 OPEN 表的节点进行排序,以便能够从中选出最好的节点作为步 5 扩展使用.这种排序可以是任意的,属于盲目搜索,也可以是一种启发信息为准则,属于启发式搜索.步 7 检测节点为目标节点时,即找到了一个约简,这时可以继续找下去,直到 OPEN 表为空表,则找到了所有约简.

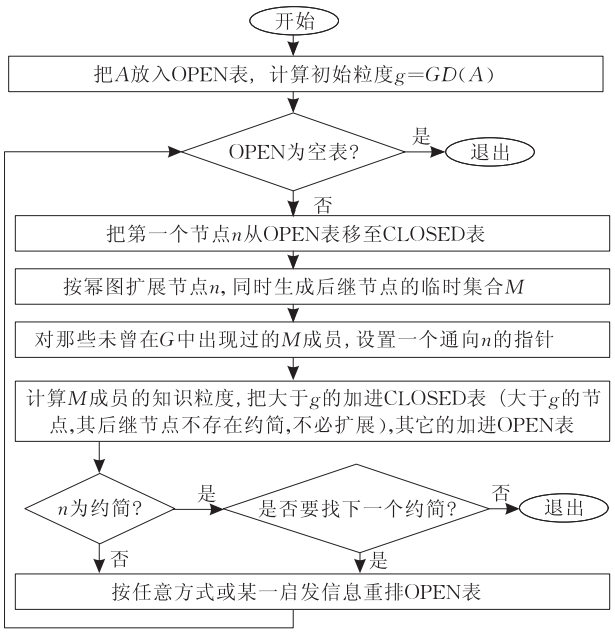

图 2　TSA 算法流程图

算法 2.　BSA(Bottom-up Searching Algorithm of attribute reduction based on power graph).

输入:信息系统 $IS=(U,A,V,f)$

输出:最优约简或者次优约简

1. 计算 $g=GD(A)$.

2. 根据推论 2 求核 $Core(A)$:对 $\forall a\in A$,计算 $GD(A-\{a\})$,所有 $GD(A-\{a\})$ 值不等于 g 的属性构成核 $Core(A)$[$Core(A)$可能为空集].

3. 建立搜索图 G,$Core(A)$为起始节点,把 $Core(A)$放在一个叫 OPEN 的未扩展节点表中.

4. 建立一个叫做 CLOSED 的已扩展节点表,其初始为空表.

5. LOOP:若 OPEN 表是空表,则退出.

6. 选择 OPEN 表上的第一个节点,把它从 OPEN 表移出并放进 CLOSED 表中,称此节点为节点 n.

7. 计算 $GD(n)$.若 $GD(n)=g$,则 n 为最优或者次优约简,输出此约简;若 $GD(n)\neq g$,则转步 9.

8. 问是否要找下一个约简?若不是,则退出;若是,则转步 11.

9. 按幂图逆向扩展节点 n，同时生成后继节点的临时集合 M. 把 M 的这些成员作为 n 的后继节点添入图 G 中.

10. 对那些未曾在 G 中出现过的（即未曾在 OPEN 表或 CLOSED 表上出现过的，在和 M 同层的节点中找）M 成员，设置一个通向 n 的指针，并加进 OPEN 表.

11. 按某一任意方式或按某个启发信息重排 OPEN 表.

12. GOLOOP.

以上搜索过程可用如图 3 所示的算法流程图来表示. 此算法的搜索从核出发，利用启发信息进行向上搜索，能得到约简或次优约简. 这种方式对于约简离核近的数据能快速找到约简.

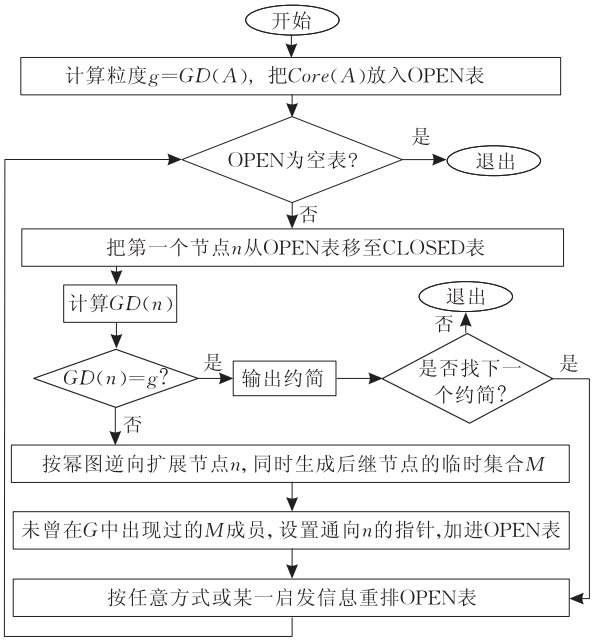

图 3 BSA 算法流程图

按照上述基于幂图的属性约简搜索式算法，根据如何重排 OPEN 表的不同，可以构造出各种各样具体的搜索算法. 按搜索的方向可以分为 Top-down 和 Bottom-up 算法. 按节点扩展的方式分为宽度优先和深度优先. 按启发式信息的有无可以分为盲目搜索和启发式搜索. 而启发式搜索又可以根据启发函数准则的不同，构造出各种各样的具体算法. 图搜索带有回溯机制，保存了搜索路径，因而可以非常方便地控制搜索的方向和进程.

4 示例说明

设信息系统 $IS=(U,A,V,f)$，其中 $U=\{x_1,\cdots,x_5\}$，$A=\{a,b,c,d,e\}$，如表 1 所示.

使用算法 TSA 即 Top-down 方式搜索，假设按宽度搜索，重排 OPEN 表的启发信息为知识粒度最小，即按宽度和采用知识粒度最小这一启发信息扩展幂图，图 4 表示了其扩展过程，括号内为知识粒度值，星号表示不必再扩展，虚线也不必扩展，因为步 6 已对幂图进行了剪枝. 一直搜索下去，直到 OPEN 表空，可以找到所有约简 $\{a,b,e\}$ 和 $\{a,b,c\}$. 比较这两个约简，可以得到最小约简 $\{a,b,c\}$ 和 $\{a,b,e\}$.

表 1 信息系统

U	A				
	a	b	c	d	e
x_1	1	0	2	1	0
x_2	0	0	1	2	1
x_3	2	0	2	1	0
x_4	0	0	2	2	2
x_5	1	1	2	1	0

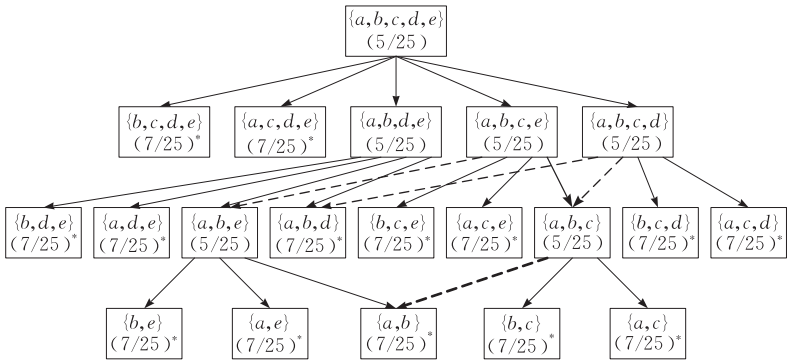

图 4 TSA 算法中幂图的扩展

使用算法 BSA 即 Bottom-up 方式搜索，同样按宽度搜索，重排 OPEN 表的启发信息为知识粒度最小.

首先计算 $GD(A)=\dfrac{(1+1+1+1+1)}{5\times 5}=\dfrac{5}{25}$；然后根据算法步 2 求核 $Core(A)=\{a,b\}$，并计算 $GD(Core(A))=\dfrac{7}{25}$.

从核出发，在幂图中逆序扩展搜索，图 5 表示了其扩展过程. 当节点的知识粒度等于 $GD(A)$ 时，则

找到了一个约简.从图中可以看出,$\{a,b,c\}$为其约简.若继续找下去,可以找到第二个约简$\{a,b,e\}$,再继续找下去还可以找到有一定冗余的约简$\{a,b,c,d\}$和$\{a,b,d,e\}$.这时比较这 4 个约简,也可以得到最小约简$\{a,b,c\}$和$\{a,b,e\}$.

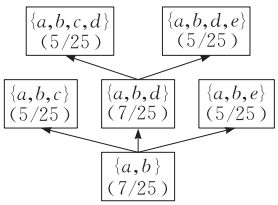

图 5　BSA 算法中幂图的扩展

从示例说明可见,幂图是一种非常方便的知识表示方式,可以根据启发信息的不同构造出许多效率各一的搜索式算法,并且可以随意控制搜索的方向和进程,这就方便我们根据具体的问题构造出效率较高的算法出来.

5　算法复杂度分析

(1)空间复杂度

信息系统占用$|U|\,|A|$空间.求所有约简或者最优约简,采用 TSA 算法或者 BSA,则按宽度搜索,最坏情况下扩展整个空间,幂图总的结点数为$2^{|A|}$个,复杂度为$O(2^{|A|}\,|A|)$,总的空间复杂度为$O(|U|\,|A|+2^{|A|}\,|A|)$.

求某个约简或者次优约简,则按深度搜索,最坏情况下扩展$|A|$层,每层保留一个结点,空间复杂度为$O(|U||A|+|A|^2)$.

(2)时间复杂度

根据知识粒度的定义,可知计算知识粒度的时间复杂度为$O(|A|\,|U|^2)$.求所有约简或者最优约简,采用 TSA 算法或者 BSA,则按宽度搜索,最坏情况下搜索整个空间,幂图总的结点数为$2^{|A|}$个,时间复杂度为$O(2^{|A|}\,|A|\,|U|^2)$,这和一般求所有约简的算法的复杂度相同.同时可以看出求所有约简的时间复杂度呈指数增长,是个 NP 问题,但基于幂图的搜索式算法可以根据启发式信息进行剪枝,这和一般求所有约简的算法在时间效率上是有改进的,搜索式算法在时间效率上的提高,很大程度上取决于重排 OPEN 表的启发式信息.

求某个约简或者次优约简,采用 TSA 算法或者 BSA,则按深度搜索,如果重排 OPEN 表的启发式函数采用属性重要度.那么 TSA 算法由两部分时间组成:求属性重要度和判别是否为约简;BSA 由 3 部分时间组成:求核、求属性重要度、判别是否为约简.求核时要计算$|A|$次知识粒度,每次计算知识粒度的时间复杂度是$O(|A|\,|U|^2)$,故求核的时间复杂度是$O(|U|^2|A|^2)$.求一次属性重要度时间复杂度是$O(|U|^2)$,每层分别求$|A|,|A|-1,\cdots,1$次,那么求属性重要度时间复杂度是$O(|U|^2|A|^2)$.判别是否为约简,最坏情况下搜索$|A|$层,每层判别一次,共判别$|A|$次,每次判别计算知识粒度的时间复杂度是$O(|A|\,|U|^2)$,则判别是否为约简的时间复杂度是$O(|A|\,|U|^2|A|)$.所以,最坏情况下,TSA 算法总的时间复杂度是$O(|U|^2|A|^2+|A|\,|U|^2|A|)$,BSA 算法总的时间复杂度是$O(|U|^2|A|^2+|U|^2|A|^2+|A|\,|U|^2|A|)$,即都为$O(|U|^2|A|^2)$,当$|U|\geq|A|$,则为$O(|U|^2)$,当$|U|<|A|$,则为$O(|A|^2)$.

6　结　语

本文提出一种新的知识表示方式——幂图,并给出两种基于幂图的属性约简搜索式算法,把属性约简的计算问题转化为图搜索式问题,以直观形象的方式展现了属性约简的过程,为属性约简问题的求解提供了一条新的途径.理论分析表明,幂图是一种新的知识表示方式,可以非常方便地构造出许多不同的启发式搜索算法来求解数据约简问题.因此,下一步的工作就是利用幂图这种知识表示,研究出更好的启发式函数.

致　谢　衷心感谢匿名审稿人提出的宝贵建议!

参　考　文　献

[1] Pawlak Z. Rough sets. International Journal of Computer and Information Science, 1982, 11(5): 341-356

[2] Pawlak Z. Rough set approach to multi-attribute decision analysis. European Journal of Operational Research, 1994, 72(3): 443-459

[3] Liu Qing. Rough Sets and Rough Reasoning. Beijing: Science Press, 2001(in Chinese)
(刘清. Rough 集及 Rough 推理. 北京:科学出版社, 2001)

[4] Wang G Y. Rough reduction in algebra view and information view. International Journal of Intelligent Systems, 2003, 18(6): 679-688

[5] Guan J W, Bell D A. Rough computational methods for information systems. Artificial Intelligences, 1998, 105(1-2): 77-103

[6] Liu Shao-Hui, Sheng Qiu-Jian, Wu Bin et al. Research on efficient algorithms for rough set methods. Chinese Journal of Computers, 2003, 26(5): 524-529(in Chinese)
(刘少辉，盛秋戬，吴斌等. Rough集理论高效算法的研究. 计算机学报，2003，26(5)：524-529)

[7] Skowron A, Rauszer C. The discernibility matrices and functions in information systems//Intelligent Decision Support-Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, 1992: 331-362

[8] Wang Jue, Wang Ju. Reduction algorithm based on discernibility matrix the ordered attributes method. Journal of Computer Science and Technology, 2001, 16(6): 489-504

[9] Miao Duo-Qian, Wang Jue. An information representation of the concepts and operations in rough set theory. Journal of Software, 1999, 10(2): 113-116(in Chinese)
(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，1999，10(2)：113-116)

[10] Miao Duo-Qian, Hu Gui-Rung. A heuristic algorithm for reduction of knowledge. Journal of Computer Research & Development, 1999, 36(6): 681-684(in Chinese)
(苗夺谦，胡桂荣. 知识约简的一种启发式算法. 计算机研究与发展，1999，36(6)：681-684)

[11] Wang Guo-Yin, Yu Hong, Yang Da-Chun. Decision table reduction based on conditional information entropy. Chinese Journal of Computers, 2002, 25(7): 759-766(in Chinese)
(王国胤，于洪，杨大春. 基于条件信息熵的决策表约简. 计算机学报，2002，25(7)：759-766)

[12] Cai Zi-Xing, Xu Guang-You. Artificial Intelligence: Principles and Applications. 3rd Edition. Beijing: Tsinghua University Press, 2001(in Chinese)
(蔡自兴，徐光祐. 人工智能及其应用. 第3版. 北京：清华大学出版社，2004)

[13] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts//Rival I eds. Ordered Sets. Reidel: Dordrecht-Boston, 1982: 445-470

[14] Yao Y Y. Concept lattices in rough set theory//Dick S, Kurgan L, Pedrycz W eds. Proceedings of 2004 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS 2004). IEEE Catalog Number: 04TH8736, 2004: 796-801

[15] Zhang W X, Wei L, Qi J J. Attribute reduction in concept lattice based on discernibility matrix//LNAI 3642, 2005: 157-165

[16] Miao Duo-Qian, Fan Shi-dong. The calculation of knowledge granulation and its application. Systems Engineering—Theory & Practice, 2002, 22(1): 48-56(in Chinese)
(苗夺谦，范世栋. 知识的粒度计算及其应用. 系统工程理论与实践，2002，22(1)：48-56)

CHEN Yu-Ming, born in 1977, Ph. D. candidate, lecturer. His research interests include rough set theory and pattern recognition.

MIAO Duo-Qian, born in 1964, professor, Ph. D. supervisor. His research interests include rough set theory, granular computing, web intelligence and pattern recognition.

Background

Rough set theory has been successfully applied to several data analysis tasks in the field of artificial intelligence, such as data mining, knowledge discovery, pattern recognition, decision analysis, process control, image processing and medical diagnosis. Attribute reduction is one of the most important topics of rough set theory. Most existing algorithms are studied under both algebra and information representations. Under algebra representation, reduction algorithms mainly include positive region based algorithm, discernible matrix based algorithm and improved discernible matrix based algorithm. Under information representation, reduction algorithms mainly include information entropy based algorithm, conditional information entropy based algorithm and mutual information based algorithm.

The main object of this paper is to provide a novel knowledge representation for attribute reduction question. As problem solving under different knowledge representations corresponding to different difficulties, power graph which is a graph representation is presented in the paper. Two kinds of searching ways based on power graph are also proposed, which are Top-down and Bottom-up. The search efficiency based on power graph is various according to different heuristic information. How to find more efficient heuristic information under the graph representation is future work.

The research is supported by the National Natural Science Foundation of China under grant Nos. 60475019, 60775036, and the Research Fund for the Doctoral Program of Higher Education of China under grant No. 20060247039.