基于粗糙集理论的概念格属性约简及算法*)

吴 强[1,2] 周 文[1] 刘宗田[1] 陈慧琼

(上海大学计算机工程与科学学院 上海 200072)[1] (绍兴文理学院计算机系 绍兴 312000)[2]

摘 要 粗糙集理论在数据分析中善于解决约简冗余属性与对象、寻求最小属性集等问题,而约简概念格是形式概念知识表示中解决复杂性的重要途径。用粗糙集的方法重新认识形式概念及概念格,并把二者有机地结合起来,为概念格的约简提供了一个新的思路和方法。本文就这些问题做了一些基本的研究。

关键词 概念格,粗糙集理论,约简

Intent Reduction of Concept Lattices and its Algorithm Based on Rough Set Theory

WU Qiang[1,2] ZHOU Wen[1] LIU Zong-Tian[2] CHEN Hui-Qiong

(School of Computer Engineering and Science, Shanghai University, Shanghai 20072)[1]

(Department of computer science, Shaoxing College of Art and Science, Shaoxing 312000)[2]

Abstract Rough set theory has advantage in solving the problems of the reduction of objects and intents, seeking minimum intent set, and etc in data analysis. In the same time, the reduction of concept lattice plays an important role to reduce the complexity in knowledge representation based on formal concept analysis. From the point of view of rough set, some new methods of concept lattice reduction are presented which take the advantage of both rough set and formal concept analysis. The aim of this paper is to produce some new methods of concept lattice reduction by the combination of rough set theory and formal concept analysis.

Keywords Formal concept analysis, Rough set theory, Reduction

形式概念分析是一个从对象数据表里自然聚类抽取信息的数据分析方法。从数据集中(形式概念分析中称为'背景')生成概念格的过程实质上是一个概念聚类的过程。这种聚类在传统意义上被自然地解释为概念且具有一个子概念-超概念的层次排序,即概念格。基本粗糙集理论认为"概念"即是对象的集合,"知识"是将对象进行分类的能力。假定我们对论域里的对象或实例有必备的知识(这些知识可以认为是对对象的内涵或属性的某种刻画),通过这些知识能够将对象划分到不同的类别。如果两个对象具有相同的信息即根据已有的信息无法将它们区分开来,则它们是不可区分的、等价的关系。不可区分关系是粗糙集理论的最基本的概念,它使得知识具有了清晰的数学定义以便于用集合运算处理。粗糙集理论在数据分析中善于解决的基本问题包括发现属性间的依赖关系、约简冗余属性与对象、寻求最小属性子集以及生成决策规则等。在数据挖掘中,粗糙集主要被用来分类、聚类和提取关联规则。分类是将对象分配到已知的一个先前类,聚类则是在事先不知道类的情况下将对象聚成类,它可以看作是分类的一种形式。粗糙集理论主要是应用于分类,最近一些研究者已经尝试用它来聚类[1]。

形式概念分析和粗糙集理论是两种不同的数学方法,但是利用等价关系的聚类我们可以将二者有机地结合起来。

1 概念格的属性约简

一般说来概念格是源自形式背景的。一个形式背景由集合 G、M 以及它们之间的关系 I 组成,记为 $C=(G,M,I)$。G 的元素称为对象,M 的元素称为属性。即对 $g\in G, m\in M$ 有 gIm 或有 $(g,m)\in I$。形式背景中的一个形式概念是一个对 (A,B),其中 $A\subseteq G, B\subseteq M$,满足 $A^*=B$,且 $B^*=A$,$A^*:=\{m\in M\mid \forall g\in A \quad gIm\}$ $B^*:=\{g\in G\mid \forall m\in B \quad gIm\}$。A、B 分别称为形式概念 (A,B) 的外延(extent)和内涵(intent)。如果 (A_1,B_1) 和 (A_2,B_2) 是一个形式背景的两个形式概念,$A_1\subseteq A_2$(等同于 $B_2\subseteq B_1$)记为 $(A_1,B_1)\leqslant(A_2,B_2)$。关系 $\leqslant$ 被称为层次序或简单序。按此方式排序的所有形式概念的集合被称为形式背景的概念格。对 $x\in G, a\in M$ 有 $x^*=\{x\}^*$,$a^*=\{a\}^*$。

定理 1[2] 若 $A_1,A_2\subseteq G$ 和 $B_1,B_2\subseteq M$ 分别是背景 $C=(G,M,I)$ 上的对象集合与属性集合,则有:

$$(1)\begin{matrix}A_1\subseteq A_2\Rightarrow A_2^*\subseteq A_1^*\\ B_1\subseteq B_2\Rightarrow B_2^*\subseteq B_1^*\end{matrix}$$

$$(2)\begin{matrix}A_1\subseteq A_1^{**}\\ B_1\subseteq B_1^{**}\end{matrix}$$

$$(3)\begin{matrix}A_1^*=A_1^{***}\\ B_1^*=B_1^{***}\end{matrix}$$

对概念格来说,属性约简大致可以分为绝对属性约简和相对属性约简两类。

定义 1 对于一个给定的概念 (G_1,M_1),如果属性集合 M_2 满足下面两个条件:

$$M_2^*=M_1^*=G_1 \tag{1}$$

$$M_3^*\supset M_2^*=G_1 \tag{2}$$

对任意的 $M_3\subset M_2$,则它被称为概念的一个内涵约简或称绝对约简[6]。

定义 2 在概念 $C=(A,B=\{B_1,B_2,\cdots,B_n\}$ 中,对其中

*)本文受国家自然科学基金项目"分布式概念格数学模型及算法研究"(编号:60275022)资助。吴 强 副教授,博士生,从事人工智能,知识发现与知识表示等方面的研究;周 文 博士生,研究方向为人工智能;刘宗田 教授,博导,从事人工智能、软件工程等方面的研究;陈慧琼 讲师,从事人工智能、软件工程等方面的研究。

的某个内涵 B_i，如果存在 $C\leqslant C_I$，$B_{np}\subseteq C_I$，使得 $\exists B_s\subset\{B_{np}\cup B-\{B_i\}\}$ （$B_s\neq\varnothing$且 $B_i=\cup P, P\in B$）则称 B_i 为该概念的相对冗余内涵。对概念 $C=(A,B)$，$B'=B-\{C$ 的所有相对冗余内涵$\}$，称(A,B')为概念 C 的相对约简（概念），把以相对约简形式表示的概念格称为相对约简格[7]。

从以上定义可以看出，无论是绝对约简还是相对约简，概念本身依然存在概念格结构不应该改变。

2 基于粗糙集理论的概念格与属性约简

在粗糙集理论下，对于构成概念的任意集合我们可以认为其是粗糙集合。对于形式背景 $C=(G,M,I)$ 上的每个非空子集 $A\subseteq M$ 定义一个如下等价关系 $R_A=\{(x_i,x_j)\in G\times G: a_k(x_i)=a_k(x_j), \forall a_k\in A\}$。因为 R_A 是一个 G 上的等价关系，其分 G 为不相交的子集簇，记为 $G/R_A=\{[x_i]_A: x\in G\}$，$[x_i]_A$，表示由 A 决定的 x_i 等价类，$[x_i]_A=\{x_j\in G: (x_i,x_j)\in R_A\}$。对称的用同样方法可以对属性进行定义 $B\subseteq G, R_B=\{(a_i,a_j)\in M\times M: x_l(a_i)=x_l(a_j), \forall x_l\in B\}$以及$[y_i]_B$。

容易看出，R_A 即为粗糙集中的不可分辨关系 $IND(A)$[9]，即有 $IND(A)=R_A$。

定理 2 在一个给定的背景上，概念格的每一个结点是粗糙集在此背景属性和对象的分类并集。即对背景上的概念 C 上的概念(X,Y)有 $X=\bigcup\limits_{x\in X}[x]_{R_M}$，$Y=\bigcup\limits_{y\in Y}[y]_{R_G}$。

证明：显然，$\varnothing$，G 是在分割簇中。对 $x\in X$，因为 R_M 是自反的，有 $x\in[x]_{R_M}$，即 $x\in\bigcup\limits_{x\in X}[x]_{R_M}$。因此 $X\subseteq\bigcup\limits_{x\in X}[x]_{R_M}$。反之，因为 X 是概念的对象集，那么 $X=(\bigcup\limits_{x\in X}x)^{**}=(\bigcap\limits_{x\in X}x^*)^*=(\bigcap\limits_{x\in X}[x]_{R_M}^*)^*=(\bigcup\limits_{x\in X}[x]_{R_M})^{**}\supseteq\bigcup\limits_{x\in X}[x]_{R_M}$，故 $X=\bigcup\limits_{x\in X}[x]_{R_M}$。同理可证属性成立。

因此，对属性分类能力的约简即可完成概念格的约简。粗糙集的约简方法就可以应用到概念格的约简上来。

给定一个背景 $C=(C,M,R)$，属性约简是一个属性集的最小集 $B\subseteq M$，满足 $IND_C(B)=IND_C(M)$。也就是说一个约简是保持域的分类的属性的最小集。这个集保持整个属性集所完成的分类能力。

假设 C 具有 n 个对象。C 的不可分辨矩阵是一个 $n\times n$ 得对称阵，其元素如下 $C_{ij}=\{b\in B\mid b(x_i)\neq b(x_j)$，对 $i,j=1,\cdots,n$。其每个元素由 x_i 与 x_j 不同的属性集构成。由粗糙集定义 m 个布尔变量 $b_1^*,\cdots,b_m^*$ 的不可分辨函数

$$f_c(b_1^*,\cdots,b_m^*)=\wedge\{\vee C_{ij}^*\mid 1\leqslant j\leqslant i\leqslant n, c_{ij}\neq\varnothing\}$$

这里 $C_{ij}^*=\{b^*\mid b\in c_{ij}\}$. f_c。的所有蕴含式决定了背景的属性的所有约简集。

表 1 一个形式背景

	a	b	c	d	e
1		×			
2		×		×	×
3	×		×	×	
4	×		×	×	
5		×	×	×	×
6			×	×	

例 1 由表 1 给出的形式背景下的可辨识函数为：

$$f_d(a,b,c,d,e)=(d\vee e)(a\vee b\vee c\vee d)(c\vee d\vee e)(b\vee c\vee d)(a\vee b\vee c\vee e)c(b\vee c\vee e)(a\vee b\vee e)(a\vee b)(b\vee e)$$
$$=(d\vee e)c(a\vee b)(b\vee e)=abcd\vee acde\vee bcde\vee abce$$

背景上的对象分类[1]、[2]、[3]=[4]、[5]、[6]，属性分类[a]、[b]、[c]、[d]、[e]。其约简前和约简后约简集{a,b,c,d}上的背景、概念格如表 1、表 2 及图 1、图 2 所示。

表 2 约简后的背景

	a	b	c	d
1		×		
2		×		×
3	×		×	×
4	×		×	×
5		×	×	×
6			×	×

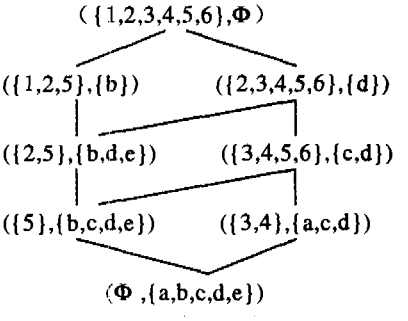

图 1 约简前的概念格

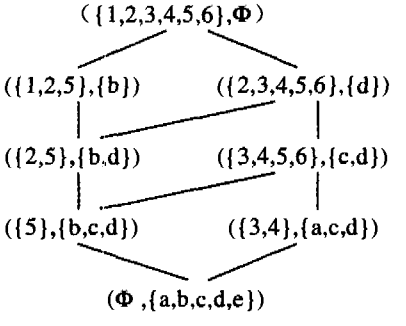

图 2 约简后的概念格

定理 3 由分辨函数决定的属性约简所构成的背景上的概念格是源背景的子图。

证明：对于背景上的一个形式概念(X,Y)，由定理 1 可得 $Y^*=\bigcup\limits_{x\in X}[x]$，$X=Y^*$，$Y=X^*$。因为约简不改变属性的分类能力，可以推出约简后的形式概念 Y^* 不变，而由定理 1 $(Y^*)^*\subseteq Y$ 即 $X^*=Y^{**}=\bigcup\limits_{y\in Y}[y]\subseteq Y$ 均为约简前对应 X^* 的子集，也即约简后的概念格的节点集是约简前节点的子集，故为子图。

例 2 一个背景的概念格及其约简

表 3

	a	b	c	d	e
1	×		×		×
2		×		×	
3		×	×	×	
4	×	×	×		
5		×	×	×	
6			×		×

其可辨识函数为

$$f_d(a,b,c,d,e)=(a\vee b\vee c\vee d\vee e)(a\vee b\vee d\vee e)(a\vee b\vee d\vee e)a$$
$$c(a\vee c\vee d)c(b\vee c\vee d\vee e)$$
$$(a\vee d)(b\vee d\vee e)$$
$$(a\vee d)(a\vee b\vee e)$$

$$(b \vee d \vee e)$$
$$= ac(b \vee d \vee e)$$
$$= abc \vee acd \vee ace$$

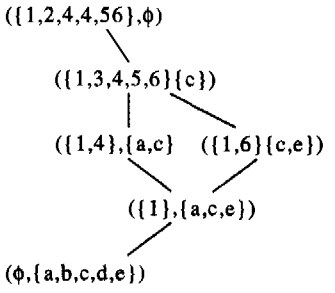

图 3

属性有三个约简集，原背景下概念格和三个约简集下的概念格分别如下：

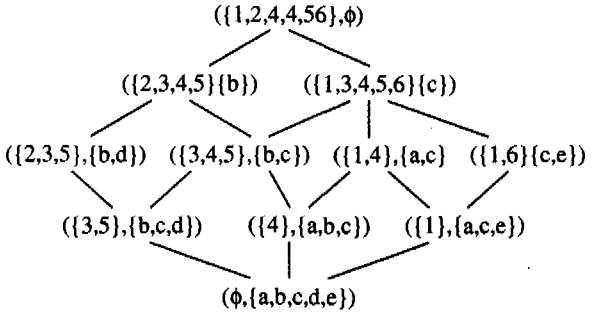

图 4

约简集$\{a,b,c\}$下的概念格如图 5。

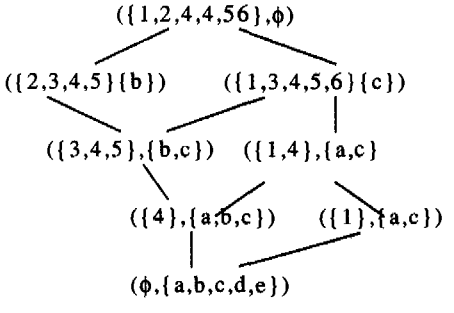

图 5

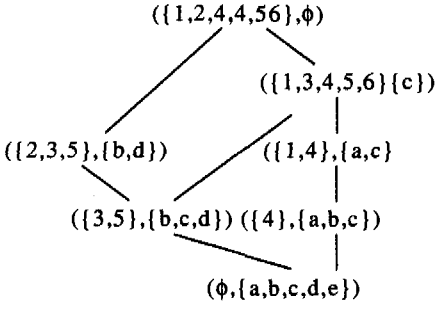

图 6

约简集{a,c,d}下的概念格如图 6。约简集{a,c,e}下的概念格如图 3。

定理 4 由可辨识函数决定的属性约简集构成的背景上的概念格与原背景同构需满足以下条件：(1)约简集不改变对象对属性的拥有值。(2)对形式概念中的对象集 $A_i \neq A_j$ 有 $A_i^* \neq A_j^*$ $(i \neq j)$。

证明：反证法：(1)若改变了对象对属性的拥有值，即改变了 xIy 关系。虽然属性仍然保持了对 x 的分类能力，但由概念及概念格定义知其对应关系改变，故格发生变化。(2)若 $A_i \neq A_j$，有 $A_i^* = A_j^*$ $(i \neq j)$。不妨设 $A_i \subset A_j$ 则由概念格定理 1 立即得出 $A_j^* \subset A_i^*$，同理对 $A_i \supset A_j$ 亦成立。故矛盾。

3 基于粗糙集理论的概念格约简算法

根据以上理论，我们可以得到基于粗糙集理论的概念格约简算法。

输入：概念格背景数组
输出：概念格内涵约简集
Begin
约简集为空；
For i：=1 to n−1(n 为对象的总个数)
{ For j：=1 to n
  { 比较对象的每个属性值；
    If 对象 x_i，x_j 在属性 k 的值不等 Then 相应属性间取 ∨ 运算；
    EndIf
    与下个 j 值的属性集合取 ∧ 运算；
  }
}
化简得到析取范式，每个小项即为约简集；
检查每个约简集下对象的取值情况；
If 约简集是的某个对象属性值为空 Then
  删去该集合；
Else
    对所有形式概念的对象集两两比较
      If $A_i^* = A_j^*$ $(i \neq j)$或 $A_i = \varnothing$ Then
      删去该集合；
      EndIf
EndIf
If 还有约简集 Then
约简集即为原背景的约简；
Else 无约简；
EndIf
End

结束语 本文研究的基于粗糙集理论的概念格约简实际上是绝对属性约简。我们对绝对属性约简的定义进行了修正，强调了这种约简存在的条件。由于析取范式求解所存在的复杂性，这种方法在求解过程上仍显繁琐，可以进一步加以改进。这也是有待进一步解决的问题。

参 考 文 献

1 Magnani M. Technical report on Rough Set Theory for Knowlege Discovery in Data Bases. July 2003
2 Ganter B, Wille R. Formal Concept Analysis, Mathematical Foundations, Springer, Berlin, 1999
3 Yao Y Y. A comparative study of formal concept analysis and rough set theory in data analysis, Rough Sets and Current Trends in Computing. In: Proceedings of 3rd International Conference, RSCTC'04, 2004
4 Yao Y Y. Concept lattices in rough set theory. In: Proceedings of 23rd International Meeting of the North American Fuzzy Information Processing Society, 2004
5 Saquer J, Deogun J S. Concept approximation based on rough sets and similarity measures. Int. J. Appl. Math. Comput. Sci. 2001, 11,(3):655～674
6 谢志鹏，刘宗田. 概念格节点的内涵缩减及其计算. 计算机工程，2001,3:9～10,39
7 张意德，简程，赵文兵，等. 相对约简格及其构造. 计算机工程与应用，2002,6:196～197,239
8 Pawlak Z. Slowinski R Rough set approach tO multi—attribute decision analysis [J]. European Journal of Operational Research, 1994,72(2):443～459
9 Pawlak Z. Rough Sets [M]. Norwell, Netherlands: Kluwer Academic Publishers, 1991