

粗糙概念格分层建格算法及应用

刘保相 陈焕焕 柳洁冰

(河北联合大学理学院 唐山 063009)

摘 要 粗糙概念格能够反映对象与特征间的确定与不确定关系,具有处理不确定性知识的能力,格的构建在应用过程中具有重要的意义。通过分析粗糙概念格的概念和结构,并结合一般概念格的构建思想,针对决策形式背景,提出一种以决策属性值为切入点的粗糙概念格的分层建格方法,从而丰富了粗糙概念格的构建理论。通过实例验证了该方法简单直观,效果良好。

关键词 粗糙概念格,分层建格,决策属性值,近似,内涵,外延

中图分类号 TP182 文献标识码 A

Layered Construction Algorithm and Application of Rough Concept Lattice

LIU Bao-xiang CHEN Huan-huan LIU Jie-bing

(College of Science, Hebei United University, Tangshan 063009, China)

Abstract Rough concept lattice can reflect the certain and uncertain relationships between objects and attributes, and can deal with uncertainty knowledge. It is an important task to construct concept lattice efficiently in the applications. Analyzing the concept and structure of rough concept lattice, and combining with the construction of the general concept lattice thought, this paper proposed a layered construction algorithm, which is based on the decision-attribute value in decision context as the breakthrough point, and enriches the construction theory. The examples show that this algorithm is simple and intuitive, and gets good effect.

Keywords Rough concept lattice, Layered construction algorithm, Decision-attribute value, Approximation, Intent, Extent

1 引言

概念格由 Wille R 于 1982 年首次提出,是一种数据分析和规则提取的有效工具^[1]。概念格由一系列的格节点形成,每个格节点是一个形式概念,由内涵和外延组成,反映了一类概念层次关系,具有完备性和精确性的特点,广泛应用在信息检索、数字图书馆、知识发现等方面^[2]。

一般概念格是一种研究和处理概念内涵与外延确定性关系的数学方法,对于提取不确定性知识存在较大的局限性。文献[3]将粗集理论应用于概念格,提出了一种新的概念格:粗糙概念格,其采用粗集理论中近似的方法描述概念格的内涵所拥有的外延,具有描述不确定知识的能力。

目前,国内学者对概念格进行了多方面的深入研究^[4,5]。构建方法已比较成熟,可分为两类^[2]:批处理算法(batch algorithm)和渐进式算法(incremental algorithm)。

批处理算法首先根据形式背景生成概念节点,然后将每一个形式概念节点与它的直接父节点和直接子节点进行连接。根据形成概念的层次顺序的不同,可分为自顶向下、自底向上和枚举法 3 类。其中自顶向下的方法是首先确定顶节

点,其次再向下构建格节点,常见的有 Bordat 方法;自底向上的方法与自顶向下的方法相反,如 Chein 方法;枚举方法需要先枚举形式概念节点,再确定其父子关系。渐进式算法的思想需要首先初始化概念格为空,将当前要插入的对象和现有格中所有的格节点作交运算,根据结果的不同采用不同的方法,典型的算法有 Godin, Capincot 和 T. B. H 的算法。

Bordat 算法^[6]是一种一般概念格的自顶向下构造方法,该方法首先构造概念格中最上面的概念,然后形成此概念的所有直接的子节点,再将所有的具有父子关系的节点连线,依次对每个节点重复上面的过程,直至找出底部节点。此方法的优点是简单快捷、易于计算机化,缺点是生成了一些冗余的形式概念节点。

文献[7,8]提出了一种粗糙概念格的构造方法,首先是构造所有的内涵集合,然后对每个内涵求其对应的上近似外延与下近似外延,生成格节点,再构造父子关系,最后增加上下界节点。此方法的缺点是需要节点间寻找父子关系,效率较低。

本文以 Bordat 算法为基础,并对其进行改进,逐层生成格节点,及时合并同层中的冗余节点,改进算法比原有的算法简

到稿日期:2012-06-21 返修日期:2012-09-25 本文受河北省自然科学基金(A2011209046, A2012209030)资助。

刘保相(1957—),男,教授,主要研究方向为模糊集、粗糙集、概念格、人工智能等, E-mail: Liubx5888@126.com; 陈焕焕(1983—),女,硕士,主要研究方向为模糊集、粗糙集、概念格、人工智能; 柳洁冰(1981—),女,硕士,主要研究方向为模糊集、粗糙集、概念格。

单。在此基础上,提出了粗糙概念格的一种分层构造算法。

2 粗糙概念格及其性质

定义 1 设 A, B 为任意两个集合, $P(A)$ 为 A 的幂集, $B = \{b_1, b_2, \dots, b_n\}$, 称 $A * B = \{(a, b) | a \in P(A), b = "b_1 b_2 \dots b_n"$ 为集合 A 和集合 B 的条件决策属性积。

定义 2 设 $(G, C * D, I)$ 为决策背景, 其中: G 为非空对象集, C 为条件属性集, D 为决策属性集。对于 $\forall Y \subseteq C \cup D$, $M = \{x | \forall x \in G, f(x) \cap Y \neq \Phi\}$, $N = \{x | \forall x \in G, f(x) \cap Y = Y\}$, 三元序偶 (M, N, Y) 构成了概念格 L 的任一节点, 则称 L 为由决策背景 $(G, C * D, I)$ 诱导的粗糙概念格。其中: Y 为内涵, 是对概念的描述; M 为上近似外延, N 为下近似外延, 分别表示内涵可能与肯定覆盖的对象集。也就是说, 概念的外延可以由内涵的上下近似集得到。

定义 3 对于决策背景 $(G, C * D, I)$, 可定义算子 f, g 为: $\forall x \in G, f(x) = \{y | \forall y \in C \cup D, xIy\}$, f 是对象 x 具有的所有属性的映射; $\forall y \in C \cup D, g(y) = \{x | \forall x \in G, xIy\}$, g 是属性 y 所覆盖的所有对象的映射。

定理 1 设序偶 (M, N, Y) 为粗糙概念格的任一节点, 其中: $M = \{m_1, m_2, \dots, m_r\}$, $N = \{n_1, n_2, \dots, n_s\}$, $Y = \{y_1, y_2, \dots, y_t\}$, 则:

- (1) $M = g(y_1) \cup g(y_2) \cup \dots \cup g(y_t)$
- (2) $N = g(y_1) \cap g(y_2) \cap \dots \cap g(y_t)$
- (3) $Y \subseteq f(m_1) \cup f(m_2) \cup \dots \cup f(m_r)$

由粗糙概念格的定义和定理可得出下面的性质。

性质 1 在粗糙概念格中, 若格 $A(M_1, N_1, Y_1)$ 和 $B(M_2, N_2, Y_2)$ 的内涵数目相同, 即 $|Y_1| = |Y_2|$, 则格节点 A 和 B 位于同层, 自顶至下的概念内涵逐层减 1。

性质 2 若格节点 $H(M, N, Y)$ 有 n 个子节点, 分别为: $H_1(M_1, N_1, Y_1), H_2(M_2, N_2, Y_2), \dots, H_n(M_n, N_n, Y_n)$, 则 $M = \bigcup_{i=1}^n M_i, N \subseteq N_i, Y = \bigcup_{i=1}^n Y_i$ 。

证明: 因为格节点 H 是 H_i 的前驱,

所以 $Y \supseteq Y_i$, 即 $Y = \bigcup_{i=1}^n Y_i$;

又因为 $M_i = g(y_1) \cup g(y_2) \cup \dots \cup g(y_t)$, $Y = \bigcup_{i=1}^n Y_i$,

所以 $M = \bigcup_{i=1}^n M_i$;

设 $\forall x \in N$, 则 $\forall y \in Y$, 都满足 xRy , 所以 H_n 中任意属性 y 也满足 xRy , 即 $x \in N_i$, 则 $N \subseteq N_i$ 。

3 粗糙概念格的分层建格方法

3.1 建格思想

采用自顶向下的构造方法, 首先构造最上层节点, 然后逐层构造。以决策属性值为切入点, 构造根节点的子节点, 然后对按决策值分类的子节点分别构造其自身的子节点及其各层节点。

3.2 建格方法

假设有决策背景 $(G, C * D, I)$, 其中 G 为对象集, C 为条件属性集, D 为决策属性集, I 为 G 与 $C \cup D$ 间的一个二元关

系, $C = \{c_1, \dots, c_i\}$, 决策属性的 j 个值为: $d_1, \dots, d_j$, 且 C, D, G 均为非空有限集合。

(1) 处理决策背景中的多决策属性, 组合所有的多决策属性值, 将结果作为单决策属性; 否则直接转 (2)。

(2) 构造第一层节点——根节点。将条件属性集合与单决策属性值做笛卡尔乘积, 其结果作为根节点的内涵, 即为: $\{c_1, \dots, c_i, d_1, \dots, d_j\}$, 上近似外延为对象集 G , 下近似外延为空集 (因为每个对象的决策值不可能有多个), 则根节点为: $(G, \Phi, \{c_1, \dots, c_i, d_1, \dots, d_j\})$ 。

(3) 构造第二层格节点——根节点的子节点。以决策属性值为切入点, 构造根节点的子节点。决策属性共有 j 个值, 则共有 j 个子节点。

令 $j = j + 1, j = 0, 1, 2, \dots$

首先找出决策属性值 d_j 覆盖的对象, 即 $g(d_j) = \{g \in G, gId_j\}$;

其次找出 $g(d_j)$ 中每个对象所具有的所有条件属性, 将这些条件属性组合与 d_j 做笛卡尔乘积, 其结果作为此节点的内涵;

最后根据上近似外延与下近似外延的定义, 求出二者, 这样就可得出第 j 个子节点。

(4) 构造第 m 层格节点, $m = 3, 4, \dots$ 。

对 $m-1$ 层决策属性值为 d_j 的每个格节点中的条件属性集合的所有的最大幂集与 d_j 分别做笛卡乘积, 生成子节点的内涵, 然后根据定义求出上近似外延与下近似外延; 直至内涵为第二层节点中条件属性集合的最小幂集与 d_j 的笛卡尔乘积为止。

备注: 在建格过程中, 消去所有上下近似外延为空集的集合, 合并同层格节点中上下近似外延与内涵均相同的格, 将此格与上层节点中的相关格相连接。

(5) 构造末梢节点: 其上下近似外延与内涵均为空集。

3.3 算法分析

假设决策背景中记录数为 n_1 , 决策属性的个数为 n_2 , 每个条件属性的值的个数为 n_3 , 条件属性个数为 n_4 。初始化决策值时, 其时间复杂度为 $O(n_1 \times n_2 \times n_3)$; 生成根节点的子节点时, 以决策属性值为切入点, 可生成其时间复杂度为 $O(n_2 \times n_3)$; 在生成第三层节点以下的各层节点时, 需要用每个条件属性集的最大幂集和决策属性做笛卡尔积, 其总共的时间复杂度为 $O((2^{n_4-1})n_2 \times n_3)$ 。算法的优点是自动生成父子关系, 且能及时合并相同的节点。

4 实例分析

以医疗诊断的决策背景为例, 验证分层建格算法。表 1 是医疗诊断的决策背景 $(G, C * D, I)$, 其中 $G = \{1, 2, 3\}$ 是对象集, $C = \{c_1, c_2, c_3, c_4, c_5\}$ 是条件属性集, c_1, c_2, c_3, c_4, c_5 分别代表发热、咳嗽、头痛、乏力、食欲不振症状; $D = \{d_1, d_2\}$ 为决策属性集, d_1, d_2 分别代表流行性感冒和水痘两种病症; I 表示 G 中对象所具有的 $C \cup D$ 中的属性值。

表1 医疗诊断的决策背景

G	CUD					D
	c1	c2	c3	c4	c5	
1	✓		✓	✓		d1
2	✓	✓			✓	d2
3	✓			✓		d1

(1)生成根节点:1# ($G, \Phi, \{c_1, c_2, c_3, c_4, c_5, d_1, d_2\}$)

(2)生成第二层节点:

对于决策属性 d_1 :覆盖的对象 $g(d_1)=\{1,3\}$

对象1具有的条件属性为: c_1, c_3, c_4

对象3具有的条件属性为: c_1, c_4

所以内涵为: $\{c_1, c_3, c_4, d_1\}$

上近似外延为: $\{1,2,3\}$

下近似外延为: $\{1\}$

因此,2#节点为: $(\{1,2,3\}, \{1\}\{c_1, c_3, c_4, d_1\})$;

对于决策属性 d_2 :覆盖的对象 $g(d_2)=\{2\}$

对象2所具有的条件属性为: c_1, c_2, c_5 ,即内涵为: $\{c_1, c_2,$

$c_5\}$

上近似外延为: $\{1,2,3\}$

下近似外延为: $\{2\}$

所以3#节点: $(\{1,2,3\}, \{2\}\{c_1, c_2, c_5, d_2\})$ 。

(3)根据3.2节第(4)步的方法构造第三层节点:

以生成 d_1 所对应的2#节点的子节点为例:

内涵 $\{c_1, c_3, c_4, d_1\}$ 中的条件属性集为: $\{c_1, c_3, c_4\}$

其所对应的最大幂集有3个,分别是:

$\{c_1, c_3\}, \{c_1, c_4\}, \{c_3, c_4\}$

则可生成3个子节点,其内涵分别为:

$\{c_1, c_3, d_1\}, \{c_1, c_4, d_1\}, \{c_3, c_4, d_1\}$

根据内涵所对应的上下近似外延的概念可生成其分别对

应的外延,则其3个子节点分别为:

4#: $(\{1,2,3\}, \{1\}\{c_1, c_3, d_1\})$

5#: $(\{1,2,3\}, \{1,3\}\{c_1, c_4, d_1\})$

6#: $(\{1,3\}, \{1\}\{c_3, c_4, d_1\})$

同理可生成 d_2 所对应的3#节点的子节点,分别为:

7#: $(\{1,2,3\}, \{2\}\{c_1, c_2, d_2\})$

8#: $(\{1,2,3\}, \{2\}\{c_1, c_5, d_2\})$

9#: $(\{2\}, \{2\}\{c_2, c_5, d_2\})$

这样就生成了第三层节点。

(4)同上步生成第四层节点,合并相同格节点:

10#: $(\{1,2,3\}, \{1,3\}\{c_1, d_1\})$

11#: $(\{1,3\}, \{1\}\{c_3, d_1\})$

12#: $(\{1,3\}, \{1,3\}\{c_4, d_1\})$

13#: $(\{1,2\}, \{2\}\{c_1, d_2\})$

14#: $(\{2\}, \{2\}\{c_2, d_2\})$

15#: $(\{2\}, \{2\}\{c_5, d_2\})$

(5)生成末梢节点:16: (Φ, Φ, Φ)

所以:决策背景所对应的粗糙概念格如图1所示。

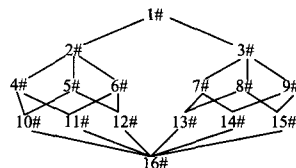


图1 表1决策背景所对应的粗糙概念格

图1中除1#根节点和16#末梢节点外,其余的节点中上近似外延是可能具有内涵中某个条件属性的对象集合,下近似外延是内涵所覆盖的所有对象的集合,即下近似外延中每个对象肯定具有内涵集合中的所有属性。比如节点 $(\{1, 3\}, \{1\}\{c_3, d_1\})$ 中,上近似外延集合 $\{1, 3\}$ 表示所有患有头痛可能患有流行性感冒的对象的集合,下近似外延 $\{1\}$ 表示所有肯定患有头痛可能患有流行性感冒的集合。其中14#格节点与15#格节点的上近似外延与下近似外延相同,已经退化为一般的概念格。所以,粗糙概念格是一般概念格的扩展,而一般概念格是粗糙概念格的特例。

对比分析:

(1)与一般概念格的Bordat方法比较:减少了冗余节点的生成;

(2)建格过程中自动生成了父子关系。

结束语 本文在基于粗糙概念格的概念的基础上,对一般概念格的Bordat方法进行了改进,以决策属性值为出发点,给出了一种粗糙概念格的分层构造方法,丰富了概念格的建格理论,并采用实例进行了验证。这种方法不需要最后构造父子关系,而是在建格的过程中自动形成了父子关系,且对同层中重复生成的节点能够及时处理,算法简单直观,易于实现。

参考文献

- [1] Wille R. Restructuring lattice theory: An approach based on hierarchies of concepts[C] // Rival I, ed. Ordered Sets. Reidel, 1982:415-470
- [2] 胡可云,陆玉昌,石纯一. 概念格及其应用进展[J]. 清华大学学报:自然科学版,2000,40(9):77-81
- [3] 杨海峰,张继福. 一种新的概念格结构:粗糙概念格[J]. 计算机技术与应用进展,2006:212-216
- [4] 沈夏炯,韩道军,刘宗田,等. 概念格构造算法的改进[J]. 计算机工程与应用,2004,24:100-103
- [5] Xie Zhi-peng, Liu Zong-tian. A Fast Incremental Algorithm for Building Concept Lattice[J]. Chinese J. Computers, 2002, 5:490-496
- [6] Godin R. Incremental concept formation algorithm based on Galois (concept) lattices[J]. Computers Intelligence, 1995, 11(2): 246-267
- [7] 杨海峰,张继福. 粗糙概念格及构造算法[J]. 计算机工程与应用, 2007, 43(24):172-175
- [8] 黄加增. 基于粗糙概念格的多决策属性分析[D]. 昆明:昆明理工大学,2008