

# 一种模糊概念格模型及其渐进式构造算法

刘宗田 强 宇 周 文 李 旭 黄美丽

(上海大学计算机工程与科学学院 上海 200072)

**摘 要** 传统形式概念分析方法无法处理现实中模糊和不确定信息,因此,对模糊概念格及其信息表示的研究具有重要意义.文中提出了一种模糊概念格模型,提出了模糊形式背景中属性隶属度值的窗口截取方法,定义了模糊概念的模糊参数 $\sigma$ 和 $\lambda$ ,给出了模糊概念格渐进式构造算法,推导出了模糊参数 $\sigma$ 和 $\lambda$ 的渐进式计算公式.模糊参数 $\sigma$ 和 $\lambda$ 分别体现了概念外延对于属性的隶属度的均值和发散程度.在模糊概念格渐进式构造算法中引入两个中间参数以实现模糊参数的渐进式计算.最后,进行了算法性能评估实验,结果表明模糊概念格的这种渐进式构造算法在时间上和空间上都具有良好的性能.

**关键词** 形式概念分析;模糊概念格模型;渐进式构造算法;模糊参数

中图法分类号 TP18

## A Fuzzy Concept Lattice Model and Its Incremental Construction Algorithm

LIU Zong-Tian QIANG Yu ZHOU Wen LI Xu HUANG Mei-Li

(School of Computer Engineering and Science, Shanghai University, Shanghai 200072)

**Abstract** Classical formal concept analysis can not deal with the vague and uncertain information in practice. So the research on fuzzy concept lattice is an important task. This paper proposes a fuzzy concept lattice model, and suggests a method in which a select window is adopted for cutting the membership degrees in fuzzy formal context, two fuzzy parameters,  $\sigma$  and  $\lambda$ , are defined. Then this paper presents a new incremental algorithm to incrementally construct the fuzzy concept lattice by inserting new object one by one, and deduces several formulas about incrementally computing the fuzzy parameters,  $\sigma$  and  $\lambda$ . The two fuzzy parameters embody the average and the diffused degree of the membership degrees in a fuzzy concept respectively. In the incremental construct algorithm of fuzzy concept lattice, two interim parameters are introduced to carry out the incremental computation of the two fuzzy parameters. Experimental results on artificially generated datasets show that the construction algorithm has excellent performance on the time-spatial complexity.

**Keywords** formal concept analysis; fuzzy concept lattice model; incremental construction algorithm; fuzzy parameter

## 1 引 言

形式概念分析是研究知识表示的领域,它的核

心数据结构——概念格是对概念以及概念之间关系的描述,在一定程度上是对客观世界的一种高度简化的描述形式. Wille 基于这种简化,系统地研究了概念的有序性质、格代数性质以及概念格与形式背

收稿日期:2006-01-04;修改稿收到日期:2006-09-03. 本课题得到国家自然科学基金(60275022,60575035)资助. 刘宗田,男,1946年生,教授,博士生导师,主要研究领域为人工智能、软件工程等. E-mail: ztliu@shu.edu.cn. 强宇,女,1971年生,博士研究生,讲师,主要研究方向为数据挖掘、人工智能. 周文,女,1979年生,博士研究生,研究方向为人工智能. 李旭,男,1981年生,硕士研究生,主要研究方向为软件工程、数据挖掘. 黄美丽,女,1978年生,博士研究生,主要研究方向为人工智能.

景的同构性质,开创了形式概念分析这一新的研究领域<sup>[1]</sup>.目前大量研究集中在概念格的数学性质方面和利用概念格作为数学工具,在知识管理<sup>[2]</sup>、机器学习<sup>[3]</sup>、信息检索<sup>[4]</sup>和软件工程<sup>[5]</sup>等诸多领域的应用方面.

在国内外关于形式概念分析的已有研究中,绝大多数是针对标准形式背景(或称为单值背景)的.实际上,信息往往是模糊的和不确定,因而,研究模糊概念格及其相关信息表示具有重要的实用意义.

## 2 相关研究工作

对于模糊、不确定信息上的形式概念分析研究,国外已有了一些研究成果.

Wolff 提出了一种基于模糊信息的表示法<sup>[6]</sup>,将传统的形式背景中的属性用模糊语言变量值表示,依据标度分类形式背景的对象,构造基于标度的格.

Burusco 和 Fuentes 讨论了 L-模糊概念集合的格结构,并给出了一个计算这种格的方法<sup>[7]</sup>.理论上,这种 L-模糊概念格定义并不限定格 L、对象集 X 和属性集 Y 是有限的,但实际上,根据这种模型,计算概念格的方法不仅要求对象集 X 和属性集 Y 是有限的,而且要求格 L 也必须是有限的,因而  $L^{G \times Y}$  也是有限的.这也就是说,它不能处理连续隶属度值的情况.

Girard 和 Ralambondrainy 提出了基于模糊量词集的方法,通过在数据类型定义中插入模糊量词集以处理模糊信息,并研究了由模糊量词描述的形式背景的概念格的构造<sup>[8-9]</sup>.实际上,精确信息与模糊信息常并存于一个多值背景中,此方法仅将模糊属性加上标识,处理时仍将多值背景转换为单值背景来计算格,有一定的局限性.

在本文中,我们给出了一种能处理连续隶属度值的形式背景的模糊概念格模型,该模型简明、直观,支持规则提取、概念聚类、属性约简等运算.

## 3 一种模糊概念格模型

**定义 1.** 模糊形式背景.一个模糊形式背景表示为  $K=(U, A, I)$ ,其中,  $U$  为对象集,  $A$  为属性集,映射  $I$  称为隶属度函数,它满足:

$$I: U \times A \rightarrow [0, 1], \text{或记作 } I(o, d) = m,$$

其中  $o \in U, d \in A, m \in [0, 1]$ .

$I: U \times A \rightarrow [0, 1]$  意味着这样定义的模糊形式背景中隶属度的取值范围是 0 到 1 的连续区间.

**定义 2.** 窗口.对于模糊形式背景中的每个属性,选取两个阈值  $\theta_d$  和  $\psi_d$ ,满足  $0 \leq \theta_d < \psi_d \leq 1$ .  $\theta_d$  和  $\psi_d$  构成窗口,  $\theta_d$  和  $\psi_d$  分别称为窗口的下沿和上沿.

该阈值既可以根据领域背景知识来确定,也可以由用户根据应用的目的而指定.

**定义 3.** 映射  $f$  和  $g$  在模糊形式背景  $K=(U, A, I)$  中,  $O \in P(U), D \in P(A)$  ( $P$  是幂集符号),在  $P(U)$  和  $P(A)$  间可定义两个映射  $f$  和  $g$ :

$$f(O) = \{d \mid \forall o \in O, \theta_d \leq I(o, d) \leq \psi_d\};$$

$$g(D) = \{o \mid \forall d \in D, \theta_d \leq I(o, d) \leq \psi_d\}.$$

**定义 4.** 模糊参数  $\sigma$ .对于对象集  $O \in P(U)$  和属性集  $D \in P(A)$ ,其中,  $D = f(O), o \in O, d \in D$ ,  $|O|$  和  $|D|$  分别是集合  $O$  和集合  $D$  的势,如果  $|O| \neq 0$  和  $|D| \neq 0$ ,则

$$\sigma_d = \frac{1}{|O|} \sum_{o \in O} I(o, d) \quad (1)$$

$$\sigma = \sum_{d \in D} (\sigma_d / d) \quad (2)$$

注意,式(2)中  $\sum$  是模糊集合中的符号.当需要指明具体对象集和属性集  $(O, D)$  时,模糊参数  $\sigma_d$  和  $\sigma$  分别记作  $\sigma_d(O, D)$  和  $\sigma(O, D)$ .

**定义 5.** 模糊参数  $\lambda$ .对于对象集  $O \in P(U)$  和属性集  $D \in P(A)$ ,其中,  $D = f(O), o \in O, d \in D$ ,  $|O|$  和  $|D|$  分别是集合  $O$  和集合  $D$  的势,如果  $|O| \neq 0$  和  $|D| \neq 0$ ,则

$$\lambda_d = \sqrt{\frac{\sum_{o \in O} (I(o, d) - \sigma_d)^2}{|O|}} \quad (3)$$

$$\lambda = \frac{1}{|D|} \sum_{d \in D} \lambda_d \quad (4)$$

当需要指明具体的对象集和属性集  $(O, D)$  时,模糊参数  $\lambda_d$  和  $\lambda$  分别记作  $\lambda_d(O, D)$  和  $\lambda(O, D)$ .

**定义 6.** 模糊概念.如果对象集  $O \in P(U)$  和属性集  $D \in P(A)$  满足  $O = g(D)$  和  $D = f(O)$ ,则  $C=(O, D, \sigma, \lambda)$  被称为模糊形式背景  $K$  的一个模糊概念.  $O$  和  $D$  分别称为模糊概念  $C$  的外延和内涵,  $\sigma$  和  $\lambda$  分别依据式(2)和式(4)计算.

在模糊概念中,  $\sigma$  表示这个概念的外延对应于每个属性的平均隶属度,体现了这个概念具有各个属性的程度;  $\lambda$  是概念外延中各对象对于各个属性的隶属度偏离平均值的平均程度,它体现了这个概念的发散程度.这两个参数对于基于模糊概念格的

模糊规则提取、概念聚类 etc 应用有重要的作用. 在不引起混淆的情况下, 模糊概念可以简单表示为  $C = (O, D)$ .

**定义 7.** 模糊概念格.  $K$  的所有模糊概念之集记为  $CS(K)$ .  $CS(K)$  上的结构是通过泛化例化关系产生的, 其定义为: 如果  $O_1 \subseteq O_2$ , 则  $(O_1, D_1) \leq (O_2, D_2)$ . 通过此关系得到的有序集  $\underline{CS}(K) = (CS(K), \leq)$  是一个格, 称作模糊形式背景  $K$  的模糊概念格, 本文在不引起混淆情况下有时简称模糊格或格.

### 4 模糊概念格的构造原理与算法

基于传统形式背景的概念格构造算法主要分为两类: 批处理算法<sup>[10]</sup>和渐进式构造算法<sup>[11-12]</sup>. 此外, 在处理分布式形式背景时还可以采用并行构造算法<sup>[13-14]</sup>; 利用两个形式背景的同构性质, 还可以采用代换式构造算法<sup>[15]</sup>. 实验证明了, 相对于批处理算法, 渐进式算法具有更多的优越性.

本文在渐进式格生成算法思想的基础上, 提出了一种在模糊形式背景上基于对象的渐进式模糊格构造算法, 推导出模糊参数  $\sigma$  和  $\lambda$  的渐进式计算公式. 为了能够用渐进式方法计算模糊参数  $\sigma$  和  $\lambda$ , 在格的构造算法中, 分别引进中间参数  $k_d$  和  $h_d$ .

$$k_d = \sum_{o \in O} I(o, d) \tag{5}$$

$$h_d = \sum_{o \in O} (I(o, d))^2 \tag{6}$$

由中间参数  $k_d, h_d$ , 式(1)和式(4)可以改写为式(7)和式(8):

$$\sigma_d = \frac{1}{|O|} k_d \tag{7}$$

$$\lambda = \frac{1}{|D|} \sum_{d \in D} \lambda_d = \frac{1}{|D| \sqrt{|O|}} \sum_{d \in D} \sqrt{h_d - 2k_d \sigma_d + \sum_{o \in O} \sigma_d^2} \tag{8}$$

算法:

1. 在模糊形式背景中对应每个属性  $d$  确定阈值  $\theta_d$  和  $\phi_d$ .
2. 概念格  $L$  初始化为空.
3. 从模糊形式背景中取出一个对象  $x^*$ , 形成二元组  $(x^*, f(\{x^*\}))$ , 如果格  $L$  中没有节点  $C(O, D)$  使得  $f(\{x^*\}) \subseteq D$ , 将  $(x^*, f(\{x^*\}))$  加入  $L$ , 计算并保存这个新节点的  $\sigma$  和  $\lambda$  的中间结果, 即对于每个  $d \in f(\{x^*\})$ ,  $k_d := I(x^*, d), h_d := (I(x^*, d))^2$ .

4. 扫描  $L$  中的所有节点, 找出所有满足  $D \subseteq f(\{x^*\})$  的格节点  $C = (O, D)$ , 这些格节点  $C$  为更新点. 将每个更新节

点  $C$  更新为  $C = (OU\{x^*\}, D)$ , 边不更新. 更新并保存  $C$  的  $\sigma$  和  $\lambda$  的中间结果, 即对于每个  $d \in D$ ):

$$K_d := K_d + I(x^*, d),$$

$$h_d := h_d + (I(x^*, d))^2.$$

5. 如果格节点  $C$  满足  $D \cap f(\{x^*\})$  不等于  $L$  中任意格节点内涵, 在这样的节点中取外延最大的一个, 定义为产生子节点, 将每个产生子点  $C_{pro}$  与  $x^*$  生成新生节点  $C_{new} = \{OU\{x^*\}, D \cap f(\{x^*\})\}$ . 连接新生节点到它的子节点和父节点. 更新并保存新生节点的  $\sigma$  和  $\lambda$  的中间结果, 即对于每个  $d \in (D \cap f(\{x^*\}))$ :

$$K_d(C_{new}) := K_d(C_{pro}) + I(x^*, d),$$

$$h_d(C_{new}) := h_d(C_{pro}) + (I(x^*, d))^2.$$

6. 当模糊形式背景为空 (也就是模糊形式背景中的对象集合  $U$  为空) 时, 转步 7, 否则转入步 3 继续.

7. 搜索所有没有子节点的节点, 如果这样的节点多于一个, 生成底节点  $(\Phi, A)$ , 增加这些节点到底节点的边.

8. 对  $L$  中除顶和底以外的每个节点用式(7), (8)算得  $\sigma$  和  $\lambda$ .

下面给出一个例子, 模糊形式背景  $K = (U, A, I)$  和阈值  $\theta_d$  和  $\phi_d$  如表 1 所示.

应用渐进式模糊概念格构造算法生成的模糊概念格及各节点的  $\sigma$  和  $\lambda$  值如图 1 所示.

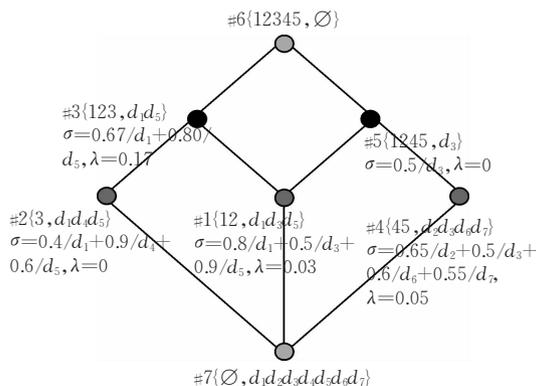


图 1 生成的模糊概念格

表 1 模糊形式背景和阈值

	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$
1	0.9	0	0.5	0.1	0.9	0.1	0.1
2	0.7	0.1	0.5	0.1	0.9	0	0
3	0.4	0.1	0	0.9	0.6	0.3	0.2
4	0	0.7	0.5	0.1	0.1	0.7	0.6
5	1.0	0.6	0.5	0.1	0.1	0.5	0.5
$\theta_d$	0.4	0.3	0.4	0.2	0.5	0.2	0.3
$\phi_d$	0.9	0.8	0.7	0.9	0.9	0.8	0.7

### 5 实验与分析

采用 JAVA 语言实现了该模糊概念格渐进式构造算法, 实验数据采用 100 组随机生成的样本数

据,为了提高数据真实性,对随机生成的每 10 组样本数据取均值,作为测试点.

本算法的空间复杂度由图 2 和图 3 体现. 图 2 测试属性数不变的情况下,对象数与格规模的关系. 其中,横坐标为对象数,初始情况取对象数为 20 个,依次递增 20 个对象数,到对象数为 200 个时为止;纵坐标为生成格的节点数.

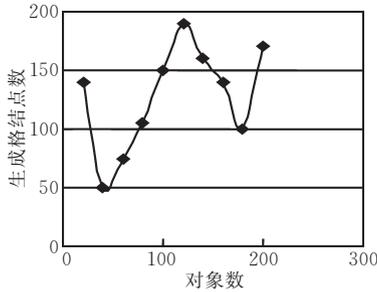


图 2 对象数与格规模的关系

图 3 测试在对象数不变的情况下,属性数与格规模的关系. 横坐标为属性数,初始属性数为 4 个,依次递增 2 个属性数,到属性数为 20 个时为止;纵坐标为格中节点数.

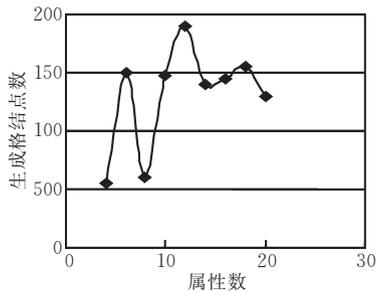


图 3 属性数与格规模的关系

算法的时间复杂性由图 4 和图 5 体现. 图 4 表现在对象数为 20 个时,属性数与运行时间的关系曲线. 图 5 表现属性数为 5 个时,对象数与运行时间的关系曲线.

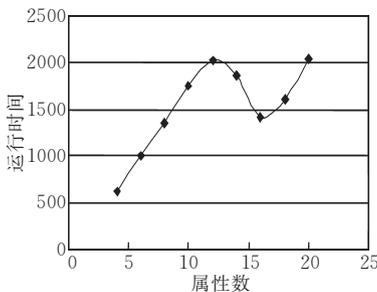


图 4 属性数与运行时间的关系

图 5 的曲线相对于图 4 的曲线而言振幅更大,可见,属性数改变算法运行时间平稳上升偶有波动.

而对对象数改变时,算法运行时间波动较大.

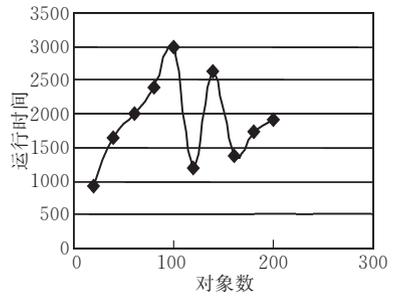


图 5 对象数与运行时间的关系

### 6 结束语

本文定义了一种模糊概念格模型,给出了该模型上的渐进式构造算法. 与多值背景的单值转换法相比,该模型生成格的规模较小,更实用. 在另外的文章中,我们还将介绍基于这种模型的模糊概念格的并行生成算法、模糊关联规则提取算法和模糊概念聚类方法. 今后将研究大型模糊概念格的分布式存储、模糊概念格模型的扩展、模糊分类技术和模糊本体的生成方法等.

### 参 考 文 献

- [1] Wille R. Concept lattices and conceptual knowledge systems. *Computers and Mathematics with Application*, 1992, 23: 493-522
- [2] Krohn U, Davies N J, Weeks R. Concept lattices for knowledge management. *BT Technology Journal*, 1999, 17(4): 108-116
- [3] Kuznetsov S O. Machine learning on the basis of formal concept analysis. *Automation and Remote Control*, 2001, 62(10): 1543-1564
- [4] Carpineto C, Romano G. Information retrieval through hybrid navigation of lattice representations. *International Journal of Human-Computer Studies*, 1996, 45: 553-578
- [5] Godin R, Mili H, Mineau G W, Missaoui R, Arfi A, Chau T-T. Design of class hierarchies based on concept (Galois) lattices. *Theory and Application of Object Systems*, 1998, 4(2): 117-134
- [6] Wolff K E. Conceptual interpretation of fuzzy theory//*Proceedings of the 6th European Congress on Intelligent Techniques and Soft Computing*. Aachen, 1998, 1: 555-562
- [7] Burusco A, Fuentes R. The study of L-fuzzy concept lattices. *Mathware & Soft Computing*, 1994, 3: 209-218
- [8] Girard R, Ralambondrainy H. Conceptual classification from imprecise data//*Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based System*. Granada, Spain, 1996, 1: 247-252

- [9] Girard R, Ralambondrainy H. Conceptual classification from structured and fuzzy data//Proceedings of the 6th IEEE International Conference on Fuzzy System. Barcelona, Spain, 1997: 135-142
- [10] Bordat J P. Calcul pratique du treillis de galois d'une correspondance. Math. Et Sci. Humaines, 24eme année, 1986, 96: 31-47
- [11] Godin R, Missaoui R, Alaoui H. Incremental concept formation algorithms based on Galois (concept) lattices. Computational Intelligence, 1995, 11(2): 246-267
- [12] Xie Zhi-Peng, Liu Zong-Tian. A fast incremental algorithm for building concept lattice. Chinese Journal of Computers, 2002, 25(5): 490-495(in Chinese)  
(谢志鹏, 刘宗田. 概念格的快速渐进式构造算法. 计算机学报, 2002, 25(5): 490-495)
- [13] Njiwoua P, Mephu N E. A parallel algorithm to build concept lattice//Proceedings of the 4th Groningen International Information Technical Conference for Students. Groningen, 1997: 103-107
- [14] Liu Zong-Tian, Li Lian-Sheng, Zhang Qing. Research on a union algorithm of multiple concept lattices//Proceedings of the 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Springer, 2003, 5: 533-540
- [15] Shen Xia-Jiong, Liu Zong-Tian et al. Isomorphic generating of concept lattices//Proceedings of the IEEE International Conference on Granular Computing. Piscataway, 2005, 1: 249-252



**LIU Zong-Tian**, born in 1946, professor, Ph. D. supervisor. His research interests include software engineering and artificial intelligence.

**QIANG Yu**, born in 1971, Ph. D. candidate, lecturer. Her current research interests include datamining and arti-

cial intelligence.

**ZHOU Wen**, born in 1979, Ph. D. candidate. Her current research interests include datamining and artificial intelligence.

**LI Xu**, born in 1981, M. S. candidate. His current research interests include software engineering and datamining.

**HUANG Mei-Li**, born in 1978, Ph. D. candidate. Her current research interests include datamining and artificial intelligence.