

《智能信息处理》课程作业

## 概念格理论与方法及其研究展望

陈梓怡

作业	分数[20]
得分	

2021 年 11 月 29 日

# 概念格理论与方法及其研究展望

陈梓怡

(大连海事大学 信息科学技术学院, 辽宁 大连 116026)

**摘 要** 概念格是根据对象和属性之间的二元关系建立起来的概念层次结构, 又由于概念格能生成 Hasse 示图, 这使它生动、简明地体现出各概念之间的关系. 由于其可视性强, 又与其他的理论不断结合, 而被广泛地应用于软件工程、知识发现等领域. 随着理论的发展, 近年来概念格在概念分析展示、构造关联等方面的优势越来越明显, 与其它学科理论(认知计算、粒计算等)的融合发展也越来越广. 本文从概念格理论的起源定义、研究内容、发展等方面总结了概念格的研究进展, 还提出了未来的一些发展方向.

**关键词** 概念格; 形式概念分析; 知识发现; 规则提取;

**中图法分类号** 8624 **DOI 号** 24861479086539-753

## Concept lattice theory and method and its research prospect

Chen Ziyi

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China)

**Abstract** Concept lattice is a concept hierarchy based on the binary relationship between objects and attributes. Since concept lattice can generate Hasse diagrams, it can vividly and concisely reflect the relationship between concepts. Because of its visibility and continuous combination with other theories, it is widely used in software engineering, knowledge discovery and other fields. With the development of theory, concept lattice has more and more obvious advantages in concept analysis, demonstration, construction and correlation in recent years, and its integration with other theories (cognitive computing, granular computing, etc.) has become more and more extensive. This paper summarizes the research progress of concept lattices from the aspects of origin, definition, research content and development, and puts forward some development directions in the future.

**Key words** Concept lattice; Formal concept analysis; Knowledge discovery; Rule extraction;

## 1 引言

概念格, 也称为 Cralois 格, 又叫做形式概念分析, 由 Wille R 于 1982 年首先提出, 它提供了一种支持数据分析的有效工具. 概念格的每个节点是一个形式概念, 由两部分组成: 外延, 即概念所覆盖的实例; 内涵, 即概念的描述, 该概念覆盖实例的共同特征. 另外, 概念格通过 Hasse 图生动和简洁地体现了这些概念之间的泛化和特化关系. 因此, 概念

格被认为是进行数据分析的有力工具. 从数据集中(概念格中称为形式背景)中生成概念格的过程实质上是一种概念聚类过程; 然而, 概念格可以用于许多机器学习的任务. 目前, 已经有了一些建造概念格的算法, 并且概念格在信息检索、数字图书馆、软件工程和知识发现等方面得到应用.

## 2 概念格基础理论

### 2.1 形式背景和概念格

形式背景 (formal context) 可以表示为三元组  $T=(O, D, R)$ , 其中  $O$  是事例 (对象) 集合,  $D$  是描述符 (属性) 集合,  $R$  是  $O$  和  $D$  之间的一个二元关系, 则存在唯一的一个偏序集与之对应, 并且这个偏序集产生一种格结构, 这种由背景  $(O, D, R)$  所诱导的格  $L$  称为概念格。格  $L$  中的每个节点是一个序偶 (称为概念), 记为  $(X, Y)$ ,  $\downarrow X \in P(O)$

为概念的  $Y \in P(D)$  外延; 称为概念的内涵。每一个序偶关于关系  $R$  是完备的, 即有性质:

$$1) X = \{x \in O \mid \forall y \in Y, xRy\}$$

$$2) Y = \{y \in D \mid \forall x \in X, xRy\}$$

在概念格节点间能够建立起一种偏序关系。具体地, 给定概念  $H_1=(X_1, Y_1)$  和  $H_2=(X_2, Y_2)$ , 则  $H_1 < H_2 \iff Y_1 \subset Y_2$ , 领先次序意味着  $H_1$  是  $H_2$  的父节点或称直接泛化。根据偏序关系可生成格的 Hasse 图: 如果  $H_1 < H_2$  并且不存在另一个元素  $H_3$  使得  $H_1 < H_3 < H_2$ , 则从  $H_1$  到  $H_2$  就存在一条边。

定义 1: 设  $U$  是对象的集合,  $M$  是属性的集合,  $I$  是  $U$  与  $M$  间的关系, 则称三元组  $K=(U, M, I)$  为一个形式背景。  $(u, m) \in I$  表示对象  $u$  具有属性  $m$ 。背景可以用一个对象集合、属性集合以及他们之间的二元关系建立起来的表格来表示, 它的每行表示某一对象, 每列则表示某一属性。

定义 2: 设  $K=(U, M, I)$  为形式背景, 对象集合  $U=\{u_1, u_2, \dots, u_m\}$ , 属性集合  $M=\{m_1, m_2, \dots, m_n\}$ 。设  $m \in M$ , 具有属性  $m$  的对象集合。定义  $g(m)=\{u \in U \mid (u, m) \in I\}$  设  $u \in U$ , 具有属性  $m$  的对象集合。定义  $f(u)=\{m \in M \mid (u, m) \in I\}$

定义 3: 设  $K=(U, M, I)$  为形式背景, 若  $A \subseteq U$ ,  $B \subseteq M$ , 那么存在  $f(A)=\{m \in M \mid u \in A, (u, m) \in I\}$  和  $g(B)=\{u \in U \mid m \in B, (u, m) \in I\}$  如果集合  $A, B$  满足  $f(A)=B, g(B)=A$ 。则称由  $(A, B)$  二元组被称为概念。  $A$  是概念  $(A, B)$  的外延,  $B$  是概念  $(A, B)$  的内涵。

定义 4: 设  $K=(U, M, I)$  为形式背景, 存在两个概念  $(A_1, B_1)$  和  $(A_2, B_2)$  满足

$$(A_1, B_1) \leq (A_2, B_2) \iff A_1 \subseteq A_2 (\iff B_1 \supseteq B_2)$$

则称  $(A_1, B_1)$  为  $(A_2, B_2)$  的子概念,  $(A_2, B_2)$  为  $(A_1, B_1)$  的超概念。由形式背景  $(U, M, I)$  中所有概念根

据它们之间的层次关系有序组成的集合, 常称为  $(U, M, I)$  的概念格。

概念格的形式背景通常是由如表 1 所示的二维数表来表示, 横向维表示属性, 纵向维表示对象, 第  $i$  行  $j$  列的数为 1 表示存在该属性, 为 0 表示不存在该属性。由此可知该形式背景中包含五个对象, 四个属性, 可形成八个概念, 与其相对应的外延与内涵如表 2 所示。由概念间的继承关系可得出如图 1 所示的 Hasse 图。

	G	A	B	C	D
1	0	1	1	1	0
2	1	0	0	0	1
3	1	0	1	1	1
4	0	1	1	1	0
5	1	0	0	0	0

表 2.2.1 形式背景示例

编号	外延	内涵
0	{1,2,3,4,5}	{}
1	{3,5}	A
2	{3,4}	B
3	{1,2,4}	C
4	{3}	A,B
5	{4}	B,C
6	{1,2}	C,D
7	{}	A,B,C,D

表 2.2.2: 所生成的概念

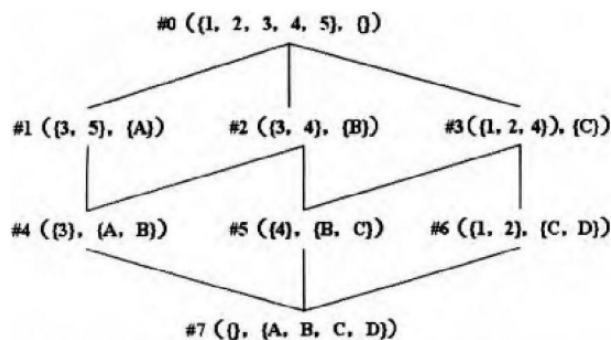


图 2.2.1 与表 1 相对的 Hasse 图

### 2.2 概念格基本原理

概念格的每一个结点都是一个形式概念, 并由内涵和外延两部分组成。外延可以理解为属于概念所有对象的集合, 即概念所覆盖的实例。内涵是全部对象所拥有的共同特征或属性, 即概念的描述。

外延和内涵是概念的两个方面, 他们相互制约又相互联系, 并呈现反比关系, 内涵越多, 外延越

少;反之,内涵越少,外延越多。这说明,人们在数据挖掘和分析时,可以利用这种反变规律,通过减少概念内涵来增加概念外延,或者增加概念内涵来减少概念外延。

由于概念格的结点反映出概念外延和内涵的有机统结点间关系体现出了典型的例化和泛化关系。概念格可以通过 Hasse 图简洁生动体现这概念之间的例化和泛化关系,表明一种概念的层次结构,从本质上体现出概念的共同属性(项目、特征)和实体(记录、对象、交易)之间的关系,因而概念格作为一种科学有力的数据挖掘方法,适合于作为规则发现的基础性数据结构。

### 2.3 优势和特征

概念格是一种核心数据结构,概念格模型根据属性和对象的二元关系进行概念层次结构的构建,并能直接反映出对象和属性的泛化和例化关系,以此更好地在概念层次上建立数据依赖及因果关系模型。概念格数据集的产生必须依托形式背景。常规的数据分析方法通常都存在一个缺陷,在模块化的区域内不能产生一个有效地识别类和对象的指导方案,这些数据分析方法往往会大规模减少给定数据信息,只为获得极少量的“重要参数”。而概念格技术依托数据集中对象和共同属性之间的二元关系,建立概念层次结构,结合格代数理论方法对数据进行分析处理,从而保证了数据信息的最大分解,并最大程度保留了数据之间的各类特殊关系。依托形式背景建立的概念格几乎包含了所有的数据细节,不会因人为因素减少和消除给定数据信息的复杂性,因此概念格在国际数据挖掘和分析领域运用广泛,备受关注。

结合概念格技术的概念定义、优劣对比.运用范围等信息,可以总结归纳出概念格的特征,具体表现在以下三方面:

①形式概念决定了形式背景下具有共同属性的一组最大的对象集合:

②概念格直观明了显示了一种对象和属性二元关系的层次分类;

③概念格中格与子格的关系充分显示出此具体属性是某种抽象属性的不同实例。

## 3 概念格研究内容

概念格理论与方法是形式概念分析研究中的基本内容,该研究已取得一系列的重要成果,主要集

中在概念格模型推广,概念格构造,概念格约简,基于概念格的规则提取,概念知识空间,概念格的粒计算方法等内容。

### 3.1 概念格构造

概念格的构造过程实际上是概念聚类的过程,是应用形式概念分析的前提。通常,概念格的大小是在指数量级上的,而且要处理的数据又多数是海量的,概念格构造算法的研究始终是形式概念分析中的一个主要问题。概念格的构造算法可分为两类:批处理算法和增量算法。使用批处理算法构造概念格要完成两项任务:一是要生成所有的格节点,即所有概念的集合;二是要建立这些格节点间的直接前驱/直接后继关系。按照这两项任务完成次序的不同,我们可以将批处理算法分为任务分割生成模型和任务交叉生成模型。任务分割生成模型是首先生成全部的概念集合,然后再找出这些概念之间的直接前驱/直接后继关系;任务交叉生成模型是在生成概念的过程中同时确定概念之间的关系。增量算法的基本思想是将当前要插入的对象与格中所有的概念求交,根据交的结果进行不同的操作。典型的算法有: Godin 算法、Capineto 算法、AddAtom 算法和 AddIntent 算法。

### 3.2 概念格约简

概念格的约简能够有效地提高概念格的堆护效率,使形式背景中所蕴含的知识易于发现,简化知识的表示方式。约简概念格实际上是在保持对象集不变的条件下,如何求得最小的属性集的过程。目前对于概念格约简理论,国外研究的理论与方法大都是以从约简行与列的角度来约简属性和对象两概念为基础,而国内的研究主要是以张文修等提出的理论为基础,给出概念格属性约简的判定定理、引入形式背景的可辨识属性矩阵,并依此为基础求得属性约简的方法,有的文献在此基础上引入了决策形式背景,进一步研究解决了决策形式背景下概念格属性约简方法。

### 3.3 概念格规则提取

概念格上的规则提取具有广泛的应用前景。规则挖掘是近年来数据挖掘的研究课题,每个概念格节点本质上就是一个最大项目集,为关联规则挖掘提供了平台,体现了概念之间的包含与分类关系,更加易于理解和表示。由于规则本身是由内涵间的关系来描述的,而表现的却是外延之间的包含与被包含关系,正是由于概念节点统一了内涵与外延之

间的关系,基于概念格的分类规则的提取在知识发现等方面有着广泛的应用。目前,对于概念格上分类规则的研究主要集中在优化概念格的构建和求解算法上。

### 3.4 概念知识空间

知识空间理论已成功应用于教育领域,它为构建知识评价体系提供了一种科学的途径。形式概念分析作为知识发现的一种有力工具已被广泛用于信息检索、知识评价等领域。

知识空间理论(简称"KST")是由美国数学心理学家于1983年首先提出的一种数学理论。其通过分析学生对不同水平的一系列有关问题解答情况来确定学生在不同知识中的认知水平。基于教育学和心理学等理论,KST建立了一套数学理论来反映教育规律,为教育评价提供了一种有效的科学方法,也是一种测试学生知识水平和构建学生知识结构的理论。1990年Koppen和Doignon研究了基于专家问询生成知识空间的方法。1993年Dowling在有限知识空间中提出了一种构建知识基的方法,同时研究了根据有限状态族生成知识空间的另一种不同方法。

### 3.5 概念格的粒计算方法

形式概念分析的粒计算方法,是指借助粒计算处理问题的基本思想,分析形式背景中对象与属性之间的依赖关系。粒计算解决问题的基本思想通常包括粒的形成、粒的转化粒的语义解释等。

不同于概念格改胜算法的研究思路,形式概念分析的粒计算方法着重概念个数 $ILI$ 的优化,即概念结果的优化。需要指出的是,一方面,形式概念分析的粒计算方法改变了传统的计算所有概念及其层次结构的研究思路;另一方面,它以具体问题背景为导向优化概念所得结果。简而言之,传统的研究思路是先构建形式背景的所有概念及层次结构,再来发掘它的具体应用;形式概念分析的粒计算方法反过来考虑这件事,即它根据具体问题背景先确定需要哪些概念及结构,甚至有时候这部分概念不包含在传统研究的范围之内,再快速有效找出这些目标概念及结构。

## 4 概念格存在的问题与展望

### 4.1 概念格约简存在的问题

概念约简能保持对概念信息的需求不变,涉及

更少的概念节点,理论意义非常明确。作为一个有前景的研究方向,仍需进一步探讨制约其发展的关键问题,例如保持二元关系不变的概念约简:一个形式背景对应的概念约简通常是不唯一的,那么如何快速获取概念个数最少的概念约简值得考虑。其次,概念约简一般不再是一个格(而是偏序集),那么是否存在某个偏序集能形成格结构?再者,既然概念约简并不改变原始的形式背景,而概念约简又只是全体概念的一部分,那么是否存在完备的概念约简(按照某种方式能生成剩余的概念)也是一个重要课题。决策形式背景上的概念约简如何实施及其语义如何解释,这些都是当前面临的困难问题。

### 4.2 概念格应用存在的问题

第一,基于概念格的 $\text{本体研究}$ :该方向已取得一些初步的研究成果,可适应时间序列模式和多维空间结构融合的本体研究,但这还远远不够,如在多粒度多维度动态数据等复杂环境下如何实现本体快速构建与有效融合仍需进一步研究。

第二,基于概念格的 $\text{认知计算}$ :概念格自身的形成过程在某种程度上体现知识认知的规律,也较适合模拟人脑的认知过程[7]。但是,人脑的认知过程不是简单机械地重复一个过程,它还包含一定的不确定性。为了更好地模拟人脑的认知过程,需要将认知的修正功能及认知遗忘和认知重现等复杂因素考虑在内,这是一个有挑战性的研究课题。

总之,传统的研究方向已较成熟,像概念格模型推广、概念格构造、概念格约简和基于概念格的规则提取,这些研究方向已逐渐开始转入深层次问题的探讨。新兴的研究方向才刚刚起步,像概念知识空间、概念格的粒计算方法和概念格应用中的认知计算等研究工作,目前尚处于理论和方法的完善阶段,有些问题甚至都还没有明确的定论。现阶段需要突破本文指出的各类关键问题,这些方面一旦取得实质性的突破,将会极大促进人工智能、认知科学等领域的进一步发展。

## 结语

概念格理论经过几十年的发展,取得一系列成果,本文从概念格理论的定义、研究内容、发展方向等方面总结了概念格的研究进展,希望能使更多读者关注和了解概念格理论,共同努力推进概念格理论与技术科学发展。另外,概念格理论与方法的研究不再只局限于形式概念分析的理论框架下进

行探讨,它发展到今天已经与其它理论深度交叉融合,如粗糙集、模糊集、粒计算、三支决策及认知计算等,从而可更好地解决机器学习、人工智能、认知科学等领域中存在的科学问题。通过近几年来概念格在国内外发展的现状统计和分析表明,概念格以其独有的特性不断地赢得众多学者的关注,其应用范围也不断地拓展.从概念格理论研究历史中得出,将其他学科与概念格理论相结合会创造出新思想、新方法。

## 5 参考文献

- [1]周飞燕,金林鹏,董军.卷积神经网络研究综述[J].计算机学报,2017,40(06):1229-1251.
- [2]奚雪峰,周国栋.面向自然语言处理的深度学习研究[J].自动化学报,2016,42(10):1445-1465.DOI:10.16383/j.aas.2016.c150682.
- [3]卢宏涛,张秦川.深度卷积神经网络在计算机视觉中的应用研究综述[J].数据采集与处理,2016,31(01):1-17.DOI:10.16337/j.1004-9037.2016.01.001.
- [4]刘建伟,刘媛,罗雄麟.深度学习研究进展[J].计算机应用研究,2014,31(07):1921-1930+1942.
- [5]WILLE R.Restructuring lattice theory:An approach based on hierarchies of concepts[C].Ordered Sets.Dordrecht:Reidel, 1982