

《智能信息处理》课程作业

## 形式概念分析

郭爱彬

作业	分数[20]
得分	

2020 年 11 月 11 日

---

# 形式概念分析

郭爱彬

(大连海事大学 信息科学技术学院 辽宁 大连 116026)

**摘 要:** 形式概念分析是一种支持数据分析的有效工具,作为一种用形式化手段表示和处理概念及其依赖关系、概念层次结构的数学理论。形式概念分析的每个节点是一个形式概念,由两部分组成:外延,内涵。另外,形式概念分析通过 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,形式概念分析被认为是进行数据分析的有力工具。从数据集中(形式概念分析中称为形式背景)中生成形式概念分析的过程实质上是一种概念聚类过程;然而,形式概念分析可以用于许多机器学习的任务。目前,被广泛用以解决诸如数据挖掘、数据分析、信息检索、软件工程、人工智能及本体论等领域的各种知识发现问题,引起了数学、计算机科学、图书情报学、医学等诸多学科的普遍关注。

**关键词:** 形式概念分析; 形式背景; 概念格; 形式概念

## Review of formal concept analysis

Guo Aibin

(Dalian Maritime University, Computer Science and

Technology, Liaoning, Dalian, 116026, China)

**Abstract:** Formal concept analysis is an effective tool to support data analysis, as a mathematical theory that uses formal means to express and process concepts, their dependencies, and concept hierarchy. Each node of formal concept analysis is a formal concept, which consists of two parts: extension and connotation. In addition, formal concept analysis vividly and concisely reflects the generalization and specialization relationship between these concepts through the Hasse diagram. Therefore, formal concept analysis is considered a powerful tool for data analysis. The process of generating formal concept analysis from the data set (called formal context in formal concept analysis) is essentially a concept clustering process; however, formal concept analysis can be used for many machine learning tasks. At present, it is widely used to solve various knowledge discovery problems in fields such as data mining, data analysis, information retrieval, software engineering, artificial intelligence and ontology, etc., which has given rise to many disciplines such as mathematics, computer science, library and information science, and medicine. General concern.

**Keywords:** Formal concept analysis; Formal background; Concept lattice

---

# 1 引言

形式概念分析 (Formal Concept Analysis, 简称 FCA) 由 Wille R 于 1982 年提出,其目的是对格理论进行重构,以促进格理论研究者和格理论的潜在使用者之间更好的交流<sup>[1]</sup>。形式概念分析提供了一种较好的层次化对象的分析方法,他能够识别那些具有共同属性的一组对象的组合。一方面,形式概念分析已具备较完善的理论基础,在应用软件工程,特别是软件维护活动中时,具有形式化方面的表达能力,从而具有较强的理论和技术说服力;另一方面,形式对象和形式属性这种二元关系经常出现在软件世界中,对这种二元关系处理的方法也推动形式概念分析技术在软件领域应用的不断发展。概念格主要用于认知计算、机器学习、模式识别、专家系统、决策分析、网页搜索等领域。已逐步发展成一门成熟的知识发现体系,并演变出很多不同的研究范式<sup>[4]</sup>。近年来,概念格应用研究出现一些新领域,比如认知概念学习,规则提取,三支决策等等。本文主要是对形式概念分析的各个方面进行详细介绍。

## 2 基本概念

### 2.1 形式背景

近年来,随着互联网、大数据和计算机技术的飞速发展,概念格作为数据挖掘和分析的重要方法和工具,在概念分析、概念展示、概念构造、概念关联等方面的优势愈发凸显,概念格与其他学科理论。形式概念是现实世界中各种概念的抽象,通过概念外延与内涵之间的关系形式化地刻画抽象概念<sup>[3]</sup>。

在形式概念分析中,数据是用形式背景表示的。形式概念分析是 Wille 提出的一种从形式背景进行数据分析和规则提取的强有力工具,形式概念分析建立在数学基础之上,对组成本体的概念、属性以及关系等用形式化的语境表述出来,然后根据语境,构造出概念格(concept lattice),即本体,从而清楚地表达出本体的结构。这种本体构建的过程是半自动的,在概念的形成阶段,需要领域专家的参与,识别出领域内的对象、属性,构建其间的关系,在概念生成之后,可以构造语境,然后利用概念格的生成算法 CLCA,自动产生本体。形式概念分析强调以人的认知为中心,提供了一种与传统的、统计的数据分析和知识表示完全不同的方法,成为了人工智能学科的重要研究对象,在机器学习、数据挖掘、信息检索等领域得到了广泛的应用。

### 2.2 概念格

概念格作为形式概念分析中核心的数据结构,它的每个结点都是一个形式概念,本质上描述了对对象和属性之间的联系,表明了概念之间的泛化与例化关系,其相应的哈斯图则实现了对数据的可视化。

在模糊概念格方面,国外已有了一些研究成果。例如 Wolff(1998)提出了一种基于模糊信息的表示法,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(1994)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。

形式概念与形式背景是形式概念分析的两个基本柱石。以下将给出形式概念的相关定义。

## 2.3 定义

**定义 2.3.1** 一个形式背景  $K=(G, M, R)$  是由两个集合  $G$  和  $M$  以及  $G$  与  $M$  间的关系  $R$  组成。 $G$  的元素称为对象， $M$  的元素称为属性。 $(g, m) \in R$  或  $gRm$  表示对象  $g$  具有属性  $m$ 。

**定义 2.3.2** 设  $A$  是对象集合  $G$  的一个子集，我们定义  $f(A)=\{m \in M \mid \exists g \in A, gRm\}$  ( $A$  中对象共同属性的集合)。相应地设  $B$  是属性集合  $M$  的一个子集，我们定义  $g(B)=\{g \in G \mid \exists m \in B, gRm\}$  (具有  $B$  中所有属性的对象的集合)。

**定义 2.3.3** 背景  $(G, M, R)$  上的一个形式概念是二元组  $(A, B)$ ，其中  $A \subseteq G, B \subseteq M$ ，而且满足  $f(A)=B, g(B)=A$ 。我们称  $A$  是概念  $(A, B)$  的外延， $B$  是概念  $(A, B)$  的内涵。

**命题 2.1** 如果  $(G, M, R)$  是一个形式背景， $A, A_1, A_2 \subseteq G$  是对象的集合， $B, B_1, B_2 \subseteq M$  是属性的集合，则有以下的一些性质：

- (1)  $A_1 \subseteq A_2, f(A_2) \subseteq f(A_1)$
- (2)  $B_1 \subseteq B_2 \Rightarrow g(B_2) \subseteq g(B_1)$
- (3)  $A \subseteq g(f(A))$
- (4)  $B \subseteq f(g(B))$
- (5)  $f(A)=f(g(f(A)))$
- (6)  $g(B)=g(f(g(B)))$
- (7)  $A \subseteq g(B) \hat{=} B \subseteq f(A) \hat{=} A * B \subseteq R$

**定义 2.3.4** 若  $C_1=(A_1, B_1), C_2=(A_2, B_2)$  是某个背景上的两个概念，而且  $A_1 \subseteq A_2$  (等价于  $B_2 \subseteq B_1$ )，则我们称  $C_1$  是  $C_2$  的子概念(也称为广义子概念)， $C_2$  是  $C_1$  的超概念(也称为广义超概念)，并记作  $C_1 < C_2$ ，关系  $<$  称为是概念的“层次序”，简称“序”。 $(G, M, R)$  的所有概念用这种序组成的集合用  $C(G, M, R)$  表示，称它为背景  $(G, M, R)$  上的概念格。

**定义 2.3.5**  $C_1=(A_1, B_1), C_2=(A_2, B_2)$  是某个背景上的两个概念， $C_1 < C_2$ 。如果  $C_1$  不存在某个子结点  $C_3=(A_3, B_3)$ ，满足  $A_3 \subseteq A_2$ ，则称  $C_1$  是  $C_2$  的直接父结点(直接父概念)， $C_2$  是  $C_1$  的直接子结点(直接子概念)。

由以上定义可知，概念格中概念的外延集合和内涵集合之间存在对偶关系，一个概念格可看作是相互联系的两个概念格。

**定义 2.3.6** 概念格的 Hasse 图解由格结点的偏序关系构成。假定格中两结点  $C_1$  和  $C_2$ ，果  $C_1 < C_2$ ，且不存在任一结点  $C_3$ ，使  $C_1 < C_3 < C_2$ ，则从  $C_1$  到  $C_2$  有边存在。Hasse 图解揭示了概念之间的特定关系，是数据分析和知识获取的有效工具。

## 3 形式背景

### 3.1 形式背景

形式概念分析首先要建立形式背景。形式背景被定义为一个三元组，公式为  $K=(G, M, I)$ ，其中  $G$  为所有对象集合， $M$  为所有属性的集合， $I \subseteq G \times M$  为  $G$  和  $M$  中元素之间的二元关系集合[6]。该三元组可以表示为二维表。在下面表 1 所示的形式背景中，关于对象集合  $G=\{g_1, g_2, g_3, g_4, g_5\}$ ，属性集合  $M=\{a, b, c, d\}$ ，二元关系  $I$  为确定性关系。实际上，形式背景一般都不是直接存在的，需要从数据

源中提取，从而就需要对数据源进行分析，采取不同的策略和算法来提取形式背景。

表 1 形式背景实例

	a	b	c	d
1	X	X		X
2	X		X	
3		X	X	
4	X	X		X
5	X			

3.2 约简形式背景

形式背景的约减包括聚类（行约减）和关联（列约减），行约简则是将属性相同的对象进行合并，列约简是对对象相同的属性进行合并。具体约简过程见表 2，表 3。

表 2 某办公用品销售公司的业务活动

	设备	计算机	复印机	打字机	专用设备
咨询	X	X	X	X	X
制作方案	X	X			
装配和安装	X	X	X	X	X
指导		X	X	X	X
培训与研讨		X			
原件和附件	X	X	X	X	X
维修	X	X	X	X	X
服务和契约		X	X	X	

由于咨询，装配和安装，原件和附件，维修都具有相同的属性，并且具有复印机的对象都具有打字机这个属性，所以将上表净化后得到新的背景如表 3 所示。

表 3 净化后的形式背景

	设备	计算机	复印机 打字机	专用设备
咨询、装配 和安装、原 件和附件、 维修	X	X	X	X
制作方案	X	X		
指导		X	X	X
培训与研讨		X		
服务和契约		X	X	

## 4 形式概念分析方法的思路解析

形式概念分析方法只是一种数学分析方法,运用到具体的分析过程中,可以通过以下环节来实现在具体问题的应用。如图 1 所示。

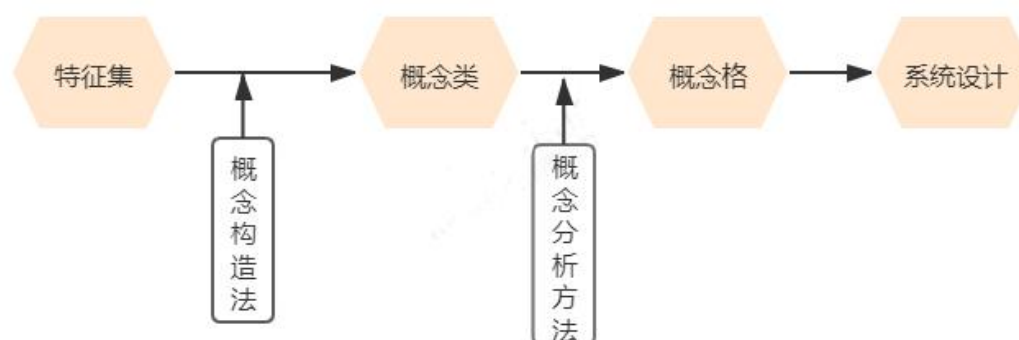


图 1 形式概念分析的具体思路

根据图 1, 特征集的构造即是通过需求分析后所要实现的各项功能的一个系统分析总结, 从而抽象表达出各种具体特征集合。同时, 将各特征集合运用形式概念分析方法, 构造出一些具体的概念, 并将这些概念抽象地表达为具体的类的形式。最后, 通过分析概念类之间的各项联系, 引入概念格的形式, 由具体的构造方法构造出系统的分析概念格图。最终, 根据这些分析, 运用到系统设计的各个阶段, 提高开发效率。

## 5 结束语

德国教授提出的概念格理论是一种基于概念由外延和内涵两部分所组成的思想单元这一哲学理解提出的。它是知识的一种表现模型, 依据知识体在内涵和外延上的依赖或因果关系, 建立概念层次结构。概念格的图体现了一种概念层次

---

结构，实现了对数据的可视化。因此，概念格作为一种具有极大潜力和有效的数据挖掘工具，备受人工智能工作者的广泛关注。概念格理论经过几十年的发展，如今已被广泛运用于软件工程、知识工程、智能控制等领域。本文从概念格理论的起源、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<sup>[2]</sup>。

然而这仍是一个年轻并在高速发展的领域。现在对概念格的研究还有许多有意义的方面，比如概念格规则提取或属性约简的启发式算法；寻找快速的模糊概念格的建格算法；概念格的规则提取问题；基于概念格的数据挖掘模型的实现等等这些都是我们以后重点研究的方向。

## 参考文献

- [1] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts[C]//International Conference on Formal Concept Analysis. Springer, Berlin, Heidelberg, 2009: 314-339.
- [2] 降惠.概念格理论研究进展与发展综述[J].办公自动化,2019,24(09):18-21+28.
- [3] Haiyun Xu,Chao Wang,Kun Dong,Zenghui Yue.Identification and Prediction of Interdisciplinary Research Topics: A Study Based on the Concept Lattice Theory[J].Journal of Data and Information Science,2019,4(01):60-88.
- [4] Jifu Zhang,Xujun Zhao,Sulan Zhang,Shu Yin,Xiao Qin. Interrelation analysis of celestial spectra data using constrained frequent pattern trees[J]. Knowledge-Based Systems,2013,41.