《智能信息处理》课程考试

**形式概念分析**

郭佑彰

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 考核 | 到课[10] | 作业[20] | 考试[70] | 课程成绩[100] |
| 得分 |  |  |  |  |

2020年11月12日

形式概念分析

郭佑彰

（大连海事大学 信息科学技术学院, 大连 116026）

摘 要 形式概念分析是由Wille教授于1982年首先提出的，它提供了一种支持数据分析的有效工具。形式概念包括外延和内涵两部分，它本质上描述了形式对象与形式属性之间的关系。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。从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。概念格可以用于完成许多机器学习的任务，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也有很广泛的应用。

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 Youzhang

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

**Abstract** Formal concept analysis was first proposed by Professor wille in 1982. It provides an effective tool to support data analysis. The concept of form includes denotation and connotation, which essentially describes the relationship between formal objects and formal attributes. Concept lattice vividly and concisely reflects the generalization and specialization relationship between these concepts through Hasse diagram. The process of generating concept lattice from formal context is essentially a concept clustering process. Concept lattice can be used to complete many machine learning tasks, and it is also widely used in information retrieval, digital library, software engineering and knowledge discovery.

**Key words** Formal concept analysis; Formal background; Concept lattice

# **1 形式概念**

## 1.1 概念的含义及表示

在英文中，概念（concept）一词，其词根词缀为共同（con-）以及拿（-cept-），译为概念、观念、思想，常常指被广泛接受的思想或概念。概念（concept）一词源于拉丁语conceptus,其过去分词为concipere。该词的词根词缀意思为共同（com-）以及截取（capere）。而在中文中，概念又由两个字概和念构成。概，是古代一种量具用词，表示用作对古代量具‘斛’的满量状态做出校准。原定义为：量米粟时，使用木板在斗斛上刮平，使其处于一定范围以内，不至于过满，表示对事物做出限定，使其不超出范围。现代字义为：处于一定范围内，如：大概、概念、概括。念，则是令心，心之力，心通思，自然地思维，常思。

而在现代的定义中，概念具有两个基本特征，也就是概念的内涵和外延。因为概念的外延是从其内涵中演绎出来的，所以外延中的所有元素都具有共同的本质属性。概念的内涵就是指这个概念的含义，即该概念所反映的事物对象所特有的属性。概念的内涵为：任意标识或范畴位置，任意标识在范畴结构中的位置认知。或表达成：概念{S/合{正/反}}：合{正/反}是范畴的结构式，其中的正、反、合是三个固定的范畴结构位置，S是任意标识，S可以放到不同的位置上以获得不同的范畴意义而被定义认知为一个有某种抽象意义的概念。概念的外延就是指这个概念所反映的事物对象的范围，即具有概念所反映的属性的事物或对象。概念的外延为：内涵性质和外延种类。概念的外延包括：内涵性质{性质∨范畴}，外延种类{个体∨种类}。概念分为“性质、范畴、个体、种类”共四大基本类型。集合论中某集合的任何元素都可以而且只能属于这四类概念及其他的集合（不包括这个某集合本身）。

自然概念则是把人们认为很自然的事物用概念的形式表现出来。跟概念一样，自然概念也是由对象集以及其对应的属性集构成。自然概念是概念一个子集。自然概念的表示方法与概念的表示方法相同。

概念即对象集以及对象集所具有的属性集的总和。概念的表示方法有三种形式：表达式法、二维表法、图示法，如图1所示。其中，对象集为概念的外延，而对象集所具有的属性集为概念的内涵。在使用二维表表示概念时，每一个对象及其属性（每一行）都为一个概念。

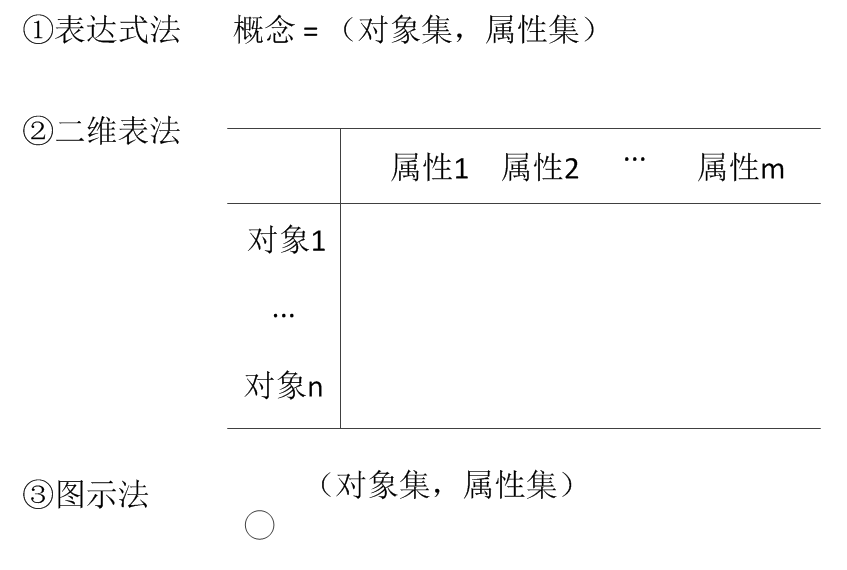


图1 概念的三种表示方法

## 1.2 形式概念的定义

形式概念分析(Formal Concept Analysis，FCA)这一概念是在1982年由R.Wille提出的。形式概念分析是应用数学的一个分支，是信息处理、知识处理的一种理论。形式概念即形式对象集以及形式对象集所具有的形式属性集的总和。其定义为：

定义1[1] 设X为形式对象集U的一个子集，B为形式属性集A的一个子集，有二元关系R⊆U×A，若X={x|x∈U，∀a∈B，xRa}，B={a|a∈A，∀x∈X，xRa}，则二元组（X，B）被称为形式概念。

举例如下：（{人，狗}，{哺乳，嗅觉}）。在该形式概念中，人和狗都具有嗅觉，且都是哺乳动物；反过来说，具有嗅觉的是人和狗，是哺乳动物的也是人和狗。

## 1.3 形式背景

单个的形式对象集以及形式对象集所具有的形式属性集构成了形式概念，形式概念的作用是构建自然概念的层次连通结构。而自然概念则是把人们认为很自然的事物用概念的形式表现出来。人脑对客观世界的认知可以构成许许多多的形式概念机，这就决定着单一的形式概念不足以描述自然界万事万物之间的关系，当不同的形势概念集相互联系时，人们就需要引入新的概念来描述它，将人脑中的形势概念集置于一张二维表中，就构成了形式背景（formal context），如图2所示。

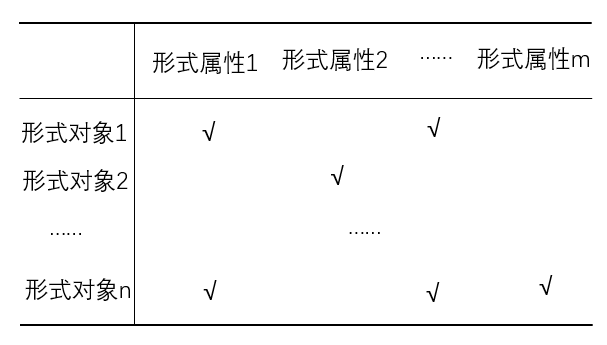


图2 形式背景

在实际应用过程中，一些复杂的形式背景可能会出现问题，比如对象或者属性的冗余。例如冗余信息可能表现在两个或多个不同的对象拥有相同的属性，人们在实际应用中把具有相同属性的不同对象合并来进行约简，称为“约简行”；同理，有时候不同的对象之间虽然具有不同的属性，但是属性与属性之间却存在着异构的现象，人们把这样的属性进行合并来进行约简，称为“约简列”；有时候多个属性之间能相互关联，某一个或多个决策属性可以由一个或者多个其他条件属性推出，人们通过“关联规则抽取”来进行约简。通过行或列的约简人们就可以进行信息的合并，从而消除形式背景之间的冗余。

# 2 概念格

## 2.1 概念格的概念

格（lattice）的意义是任两个元素的上确界和下确界都存在的偏序集。完备格为任一子集的上确界和下确界存在的偏序集，其特点是只有一个最高点，且只有一个最低点，且图中任何两点连通。概念格是元素为概念的完备格。概念格，也称为Cralois格，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，每一个形式概念都是由外延和内涵两部分组成。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。从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。目前，概念格已被广泛应用于机器学习、模式识别、专家系统、数据挖掘、信息检索等领域[2]。

## 2.2 概念格的构建

概念格的表示形式是Hasse图，概念格的构建的基础是形式背景，形式背景描述了多个形式概念之间的关系，单个形式概念描述了形式对象以及形式对象所具有的形式属性之间的关系。所以，概念格的构建必须明确不同形式对象以及不同形式对象所具有的形式属性。

概念格的构建包含以下几个步骤：生成形式背景，约简形式背景，生成单值形式背景，确定父子关系，绘制Hasse图，补充各形式概念的上确界和下确界，最后获得概念格。

某公司正常运转需要一些设备，如计算机、打印机、空调、电灯以及专用设备。这些设备所对应的“对象-属性”二元关系如下所示。

计算机 = （咨询，方案，安装，指导，培训，配件，维修，服务）

打印机 = （咨询，安装，指导，配件，维修，服务）

空调 = （咨询，安装，指导，配件，维修，服务）

电灯 = （方案，安装，配件，维修）

专用设备 = （咨询，安装，指导，配件，维修）

这些形式概念组成的形式背景如图3所示。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 计算机 | 打印机 | 空调 | 电灯 | 专用设备 |
| 咨询 | Y | Y | Y |  | Y |
| 方案 | Y |  |  | Y |  |
| 安装 | Y | Y | Y | Y | Y |
| 指导 | Y | Y | Y |  | Y |
| 培训 | Y |  |  |  |  |
| 配件 | Y | Y | Y | Y | Y |
| 维修 | Y | Y | Y | Y | Y |
| 服务 | Y | Y | Y |  |  |

图3 形式背景

可以看到，在图3中存在不同行不同列的值相同的情况，这就产生了冗余。在形式背景部分已经提到了信息合并，消除冗余的方法，即约简行、约简列以及关联规则抽取。利用以上概念可以将具有相同属性的打印机、空调合并，将计算机、打印机、空调的共同属性指导以及咨询合并，把所有对象都有的配件、安装以及维修合并，得到约简后的形式背景，如图4所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 计算机 | 打印机/空调 | 电灯 | 专用设备 |
| 方案 | Y |  | Y |  |
| 指导/咨询 | Y | Y |  | Y |
| 培训 | Y |  |  |  |
| 维修/配件/安装 | Y | Y | Y | Y |
| 服务 | Y | Y |  |  |

图5 最终的形式背景

之后，根据最终的形式背景可以通过基于父子关系的渐进式构建，生成Hasse图，得到概念格，如图6所示。

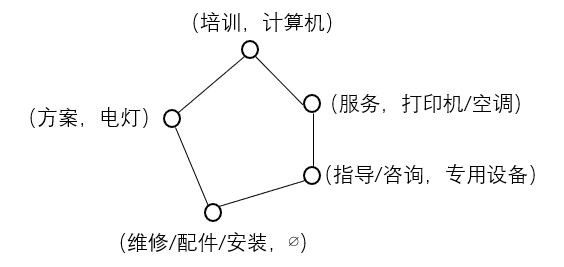


图6 概念格

## 2.3 结论分析

根据以上例子可以得知，Hasse图由于自身的特点可以清晰的表示概念格，Hasse图的每一个结点都是对象，在图中省略了自反性、省略了箭头，使图像简洁，箭头默认朝下则是基于父子关系的渐进式构建。通过观察Hasse图，人们可以更好地了解不同形式对象之间属性的继承关系，使得人们可以快速得到更多有用的信息。对于人们已知的知识，经过概念格的形式进行推演，可以更好地了解不同形式概念之间的关系，帮助人们对系统有更好的认识，从而发现潜在的知识。

# 1 总结

由Wille R于1982年首先提出的形式概念分析提供了一种支持数据分析的有效工具。形式概念包括外延和内涵两部分，它本质上描述了形式对象与形式属性之间的关系。形式背景则以二维表的形式描述了不同形式对象以及他们的形式属性之间的关系。在此基础上的概念格则是以Hasse图的形式，描述了不同形式对象之间的父子关系。概念格结构模型来源于形式概念分析理论，是形式概念分析理论中的核心数据分析工具。

概念格体现了一种概念层次结构,实现了对数据的可视化。因此,概念格受到了人们的广泛关注。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、人工智能等领域。本文从概念一词的含义到形式概念的定义及表示，再到形式背景，再到格的解释以及概念格的概念以及应用等方面总结了概念格的研究进展。当然，概念格仍是一个高速发展的领域，对于粗糙模糊概念格的研究、基于概念格的数据挖掘模型的实现等等都是以后的重点研究方向。

参 考 文 献

[1] 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.

[2] 张云中. 基于形式概念分析的领域本体构建方法研究[D]. 吉林大学, 吉林，2009.

[3] 李金海, 魏玲, 张卓,等. 概念格理论与方法及其研究展望[J]. 模式识别与人工智能, 2020(7).

[4] 朵琳, 杨丙. 一种基于用户兴趣概念格的推荐评分预测方法[J]. 小型微型计算机系统, 2020(10).