《智能信息处理》课程作业

基于形式概念的研究

陈佳慧

作业	分数[20]
得分	

2020 年 11 月 12 日

基于形式概念的研究

陈佳慧 [1)]

[1)] (大连海事大学 信息科学技术学院, 大连 201279)

摘　要　形式概念分析（FCA），它认为它的核心是“概念”，是思维的一个形式单位，了解到 FCA 的使用就是把人类已经知道的知识或者常识通过此技术手段发现和挖掘新知识，从而对原先已经认识的世界知识进行利用然后改造世界。利用抽象代数的格理论和映射理论的数学原理，开发了一套 FCA 的工具和算法，这有助于我们分析和表示任何上下文，作为它的范围和意图之间的关系。从范围和意图的子集中提取的概念可以以格的形式组织，给出一个包含层次结构，称为概念格，可以用于各种领域，进行数据挖掘。本文主要叙述了 FCA 的相关理论基础，及概念格的算法等，本文主要讨论了形式概念的相关理论基础，内容；对于形式概念的处理工具进行了简单说明；最后，讨论了构建概念格的基本算法，讨论并且总结了未来形式概念分析的发展趋势。

关键词　FCA；概念格；形式背景

中图法分类号　TP36　　DOI 号

Research Based on formal concept

CHEN Jia Hui[1)]

[1)](School of Information Science and Technology, Dalian Maritime University, Dalian 201279)

Abstract　Formal concept analysis (FCA) believes that its core is "concept", which is a formal unit of thinking. To know that the use of FCA is to discover and excavate new knowledge through this technical means by human already known knowledge or common sense, so as to make use of the previously known world knowledge and then transform the world. Using the mathematical principles of lattice theory and mapping theory of abstract algebra, a set of FCA tools and algorithms have been developed that help us analyze and represent any context as a relationship between its scope and intent. Concepts extracted from a subset of scope and intent can be organized as lattices, giving an inclusion hierarchy known as concept lattices that can be used in a variety of domains for data mining. This paper mainly describes the relevant theoretical basis of FCA, and the algorithm of concept lattice, etc. This paper mainly discusses the relevant theoretical basis and content of formal concepts. The processing tool of formal concept is explained simply. Finally, the basic algorithm of constructing concept lattice is discussed and the development trend of formal concept analysis in the future is summarized.

Key words　Formal Concept Analysis　；Concept lattice；　Formal context

1　形式概念

1.1　FCA 相关概念

形式概念分析（FCA）最初是一个应用数学的领域[1]，它试图实现一种表达人类概念思维的方法，而且建立在数学秩序理论的基础上。它已经发展成为一种无监督的学习技术，用于发现数据中的概念结构。这些结构以图形方式表示为概念层次结构。这允许分析复杂的结构和发现数据中的依赖关系。在无监督学习领域，基结构和特征选择是非常重要的，并且在这方面已经发展了最新的技术。马

尔可夫决策过程(MDP)用于不确定性下的序列决策。作为一种无监督的学习技术，最近的研究一直是在自动特征向量基础的构建中用于 MDP 的低维问题特定表示的方法。然而，这种方法的一个问题是，奖励敏感和奖励不变的特征向量基似乎都不是令人满意的。为了解决这个问题，Mahadevan 和 Liu 提出了一种以平均报酬为第一基向量的方法，通过平均调整的过渡矩阵扩展奖励函数。强化学习(RL)是无监督学习模型中的另一种方法，它将 Agent 与环境之间的相互作用作为 MDP 来建模，它使大问题中的特征选择更有效。它将非策略收敛时差学习与凸凹鞍点公式相结合利用凸正则化法进行 BLES 特征选择。从概念生成算法及其基本原理出发，分析这些在基础构建和特征选择方面的最新进展有趣的发现。

定义 1[1]. 一个抽象的形式背景定义为 K = (O，M，R) ,一个简单的形式背景如下表 1 所示:

表 1 形式背景例子

	A	B	C	D
1	a1	b1	c1	d1
2	a1	b2	c1	d2
3	a2	b1	c2	d3
4	a3	b3	c1	d4

由对象集 O 和属性集 M 以及 O 与 M 之间的关系 R 组成，且 R⊆O × M。其中 h = (P，Q)为 K 上的任一个二元组，P⊆O，Q⊆M，而且满足如下映射关系:

$f(P) = \{ m \in M \mid \forall o \in P, oRm \}$

$g(Q) = \{ o \in O \mid \forall m \in Q, oRm \}$

如果 f(P) = Q，g(Q) = P，则意味着二元组(P，Q)满足外延内涵的最大扩展性，那么定义 h = (P，Q)为 K 上的一个形式概念,Q 被称作概念 h 的内涵，P 被称作概念 h 的外延。

1.2 概念格定义

概念格（Concept Lattice）是一个以概念为元素的偏序集，其中每个节点是一个概念[2]。概念格结构模型来源于形式概念分析（FCA）理论，是 FCA 中的核心数据分析工具，它本质上描述了对象（样本）与属性（特征）之间的关联。

定义 2[2]. 设 h1 = (P1，Q1) ，h2 = (P2，Q2) 和 h3 = (P，Q)是 K 上的三个形式概念，如果 P1⊆P2(或 Q1⊇Q2)，则称 h2 是 h1 的祖先概念，h1 是 h2 的子孙概念，记为 h1≤h2，其中“≤”表示为概念之间的“层次序”。特别地，若不存在形式概念 h3，有 h1≤h3≤h2 成立,则称概念 h2 是 h1 的直接父概念，h1 是 h2 的直接子概念。K 中的所有形式概念，用这种层次(偏)序组成的集合称为 K 上的概念格，标记为〈L(O，M，R)，≤〉，为描述方便，简记为 L，一个简单的概念格如下图 1 所示。

之前我们提到过形式概念以二元表的形式可以组成形式背景，也说过概念有三种表示方式：表达式法，二维表法，图示法。所以我们可以使用图示法把形式概念进行表示，这样我们就可以构建一个哈斯图，如果哈斯图是完备格，那么这个哈斯图就是概念格。

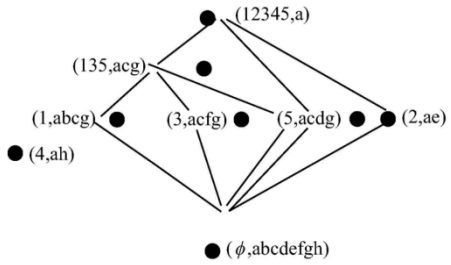

图 1 概念格的表示

概念格理论的主要思想是在形式背景中寻找所有的概念并构造出格结构以此刻画出数据集中对象与属性之间的关系。构造概念格是概念格应用的前提，但构造概念格已被证明是 NP 问题；因此，人们在构造概念格之前希望在保持格结构不变的情况下，尽可能的简化数据。目前，概念格约简的研究包括对象约简、属性约简、纵横向维护和内涵约简等。

FCA 获取一个输入表，指定一组对象及其属性，并在输入数据中查找所有属性的自然簇和对象的所有自然簇，其中自然对象簇是共享公共属性子集的所有对象的集合，自然属性簇与自然对象簇一一对应。各种各样的像缩放、修剪等操作和可视化工具，如各种线图有助于维护给定形式上下文的正式概念格。这些行动有助于从明确表示为形式概念。

2 审查方法

本文的主要目的是对 FCA 的研究进行分析,以及概念格的相关算法，并探讨和总结其发展趋势。我们考虑从谷歌学者和 Mendeley 检索的项目，以反对“正式”上的查询 概念分析”。通过探索这些检索到的项目中的资源，根据使用的各种 FCA 概念、正在开发的工具和对 SE、KR&R、DM 和 CC 的应

用领域进行分类，对这些项目进行了审查。此外，还审查了基于 FCA 概念的各种自由和开放源码软件工具的特点。列出这些自由和开放源码软件工具的主要来源是犹他州普里斯的网站。在形式背景部分我们已经提及过关于形式背景的约简，其中包括：约简行、约简列、关联规则抽取等方法。在概念格中，进行的约简规则和形式背景一样，方法类似、原理相同。

2.1 工具

作为概念知识处理和数据分析的形式化方法，FCA 主要涉及两个任务：从形式上下文生成概念和将生成的概念表示为概念格。在 FCA 领域[3]，以算法和工具的形式对这两个任务提供了强大的支持，并且 FCA 是非常强大的，我们可以利用 FCA 的不断的进行数据挖掘，得到更多有益的信息，进一步完善我们想要建立的系统。

3 概念格

虽然概念格应用起来非常的方便，但是在应用概念格的过程中，概念格的构造效率低下始终是一大难题，人们对此进行了广泛的研究，为此提出了各种不同的构造算法，但只有少数的算法能够同时生成相应的 Hasse 图，这些算法主要可以分为 3 大类：渐进式算法、并行算法和批处理算法。

3.1 渐进式算法

渐进式算法它的基本思想就是：在给定原始形式背景 K=(U，A，R) 所对应的初始概念格 L=(CS(K) ，≤)以及新增对象 x*的情况下，求解形式背景 K*=(U∪ {x*}，A，R) 所应的概念格 L*=(CS(K*)，≤)。对于初始概念格中的每个节点，根据它和新增对象 x*的特征集 f(x*) 之间的关系，格中节点可被分为更新格节点、产生子格节点和不变节点。当插入 x*时就根据节点类型对概念格做不同处理，实现节点和相应边的更新，其中典型的是 Godin 的算法。

虽然渐进式算法是概念格构造的一类重要算法。但大多关注于形式背景中对象或属性增加的情况。而当形式背景的属性减少时，已有的算法则需要重新构造概念格,较为费时，可以对概念格的结构进行一定的修改，同时也对渐进式算法进行改进。

3.2 批处理算法

批处理算法按照生成节点和边的次序不同有两种途径：一种是首先生成全部的概念集合，然后再找出节点间的边；另一种是每次生成少量概念，并将这些概念链接到节点集合中。前者称任务分割生成模型，如 Ganter 算法、Chein 算法等；后者称任务交叉生成模型，如 Bordat 算法。

3.3 并行算法

随着数据规模的不断增大，传统的渐进式和批处理算法在时间、空间复杂性方面的问题越来越突出，主要是因为生成概念格所采用的数据是集中式存储的，而算法是串行的。解决这一问题的有效途径是利用高性能并行计算机和网络并行计算的能力，因此近年来国内外的研究者纷纷将批处理算法的并行性和渐进式算法的高效性相结合提出了概念格的并行算法。

并行，分为两种，一种是时间上的并行，另一种是空间上的并行。它的基本思想是用多个处理器来协同求解同一问题，即将被求解的问题分解成若干个部分，各部分均由一个独立的处理机来并行计算。并行计算系统既可以是专门设计的、含有多个处理器的超级计算机，也可以是以某种方式互连的若干台的独立计算机构成的集群。通过并行计算集群完成数据的处理，再将处理的结果返回给用户，通过并行算法，可以大大提高效率，减轻负担。

3.4 概念格的应用

当前，基于概念格的知识发现体系已日臻完善，并且已经演变出很多研究范式。知识发现领域中，分类规则和关联规则本身就是具有价值的知识。人们在进行规则知识挖掘分析时，概念格内涵集之间的关系可以描述规则知识，非常有利于知识提取。概念格外延集之间包含和近似包含关系。可以充分体现规则知识。概念格的结点又反映了内涵和外延的统一，结点关系体现了概念泛化和例化关系，因而非常适合作为知识发现的基础数据结构。概念格每个结点的内涵本质上就是最大项目集，如同决策树和粗糙集，概念格也成为当前国内外数据分析和知识提取的有效工具之一。

近年来，国内外很多专家学者都致力于通过概念格进行聚类知识、分类知识、离群知识和关联知识等算法研究。面对目前数据海量、模糊、粗糙和不确定等特点，为了更好地利用概念格进行知识提取和知识表示，专家学者对概念格进行了扩展研究，大致可分为：扩展概念格、粗糙概念格、模糊概念格、约束概念格、量化概念格、加权概念格和多维概念格。概念格的扩展将使得概念构造效率更高，更加实用。

4 总结

我们会不断的利用已知的知识或者是现有的成果，不断地使用形式概念形式进行分析，从而不断进步，由二元关系导出来的概念格是一种非常有用的形式化工具，然后从概念格上提取各种各样的信息，进行数据挖掘，无论是先前还是未来，这都是一个高热度的研究领域，研究方向包括：高效的建格算法及剪纸算法；从格式上产生有用的规则等。

参 考 文 献

[1] ZHANG S，GUO P，ZHANG J，et al．A completeness analysis of frequent weighted concept lattices and their algebraic properties[j]. Data and Knowledge Engineering，2012，81/82: 104－117.

[2] 晏力, 刘鹏慧. 基于形式概念分析的属性约简[J]. 西华大学学报(自然科学版), 2012, 04: 10-20

[3] ZHAI YH, LI DY, QUKS. Fuzzy decision implication canonical basis[J]. International Journal of Machine Learning and Cybernetics, 2018, 9(11): 1909-1917.