《智能信息处理》课程作业

图像数据挖掘的形式概念分析

钱 程

作业	分数[20]
得分	

2021 年 11 月 28 日

图像数据挖掘的形式概念分析

钱程

（大连海事大学　信息科学与技术学院　大连　116026）

摘　要　形式概念分析(Formal Concept Analysis，FCA)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 lat-tice),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化的，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。本文简要介绍了研究背景、数据挖掘与概念形成、概念格理论和基于概念格的知识表达与处理等内容。

关键词　形式概念分析；概念格；图像数据挖掘

Formal Concept Analysis Base on Relationship

Qian Cheng

(Department of Information Science Technology, Dalian Maritime University, Dalian　116026)

Abstract　Formal Concept Analysis (FCA) is a powerful tool for data Analysis and rule extraction from Formal context, which is based on mathematics, the concept lat-tice, I. E. Ontology, is constructed according to the context, which can express the structure of ontology clearly. The process of ontology construction is semi-automatic. In the stage of concept formation, domain experts are needed to identify the objects and attributes in the domain and construct the relationships between them, then, the ontology is generated automatically by using CLCA algorithm to generate concept lattice. Formal Concept Analysis, which emphasizes human cognition as the center and provides a completely different method from traditional statistical data analysis and knowledge representation, has become an important research object of artificial intelligence, it is widely used in machine learning, data mining, information retrieval and so on. This paper briefly introduces the research background, data mining and concept formation, concept lattice theory and knowledge representation and processing based on concept lattice.

Keywords　Formal concept analysis;Concept lattice;Image data mining

1 研究目的

随着空间数据获取技术、存储技术的迅速发展，我们已经积累了大量的空间数据，其中大部分是图像数据。由于图像数据能够客观、形象地反映现实世界，在很多领域都得到了很好的应用。在遥感领域，遥感技术正在形成一个多层次、多角度、全方位和全天候的全球立体对地观测网。高、中、低轨道结合，大、中、小卫星协同，粗、细、精分辨率互补。仅地球观测系统EOS-AMI每日获取的遥感图像数据量就以TB级计算。在其他领域，我们也获取了海量的图像数据，并且每天以TB级的数量增加，如医学图像、

L业图像、视频图像等。并且随着网络技术的迅速发展，Internet网络正成为图像数据的一个巨大的存储仓库。在这些图像中必然包含着某些反映实体对象的变化、相互关系以及隐藏在其中的各种模式、演化规律等高层次的知识。然而，这些海量的图像数据的利用效率目前却仍然处于很低的层次，主要还是处于低层次的图像处理、存储技术。

从数字图像处理的研究和发展历程来看，可以分为低、中、高三个水平(级别)，有“数据驱动”和“知识驱动”之分，基于知识的图像处理是当前人们所关心和追求的，以便实现自动化、智能化和实时化。例如，遥感图像数据处理是一个秩亏过程，在理论上无法直接获取定量的结果，只有依赖于知识，才能实现遥感图像解译与提取的自动化和智能化。

2 形式背景

图像数据挖掘的过程可以理解为从图像数据中形成概念的过程，反映的是从图像数据中抽取出不同层次的概念，通过概念的泛化与概化来分析概念之间的关系，从而挖掘出图像数据中潜在的、隐含的规律性的知识。为了实现图像数据挖掘的计算机自动(半自动)化处理，就必须首先建立一个图像数据挖掘的数学的形式化的分析方法和分析体系，形式概念分析理论和商空间理论提供了这样的有效的形式化分析处理方法。

形式概念分析理论，又叫概念格理论，是一种用数学的形式化语言来反映人形成概念的过程的集合理论模型，用来研究特定领域可能存在的概念的几何结构、概念格形式。概念格基于一元关系，体现了概念内涵和外延的统一，反映了对象和特征之间的联系以及概念间的泛化与例化关系，非常适合于发现数据中潜在的概念。形式概念分析理论利用其相应的Hasse图实现了数据概念层次的可视化。在挖掘规则知识的过程中，规则本身是用内涵集之间的关系来描述的，而体现于相应外延集之间的包含(或近似包含)关系。形式概念分析理论的这种反映概念形成的过程与数据挖掘的从数据中产生知识的过程基本上是一致的，因此，概念格非常适合作为规则发现的基础性数据结构，因此可以用来进行知识规则的挖掘，利用形式概念分析理论，能够对数据挖掘的知识规则进行理论上的很好的解释，本文将对此问题进行详细讨论;同时，形式概念分析理论的Hasse图的构建过程其实是一个概念聚类过程，因此，也可以利用形式概念分析理论研究聚类和分类过程。

3 国内外现状

目前，图像数据挖掘，特别是遥感图像数据挖掘的研究还处于理论探讨和初步实验阶段，其理论和方法还都不太成熟，基本上是在其它的相关理论和技术，如一般的图像处理、基于内容的图像检索、一般数据挖掘等的基础上进一步进行扩展而发展起来的，因此，现在的图像数据挖掘技术往往是既具有一般处理技术的特点，又具有知识发现和数据挖掘的性质，也可以说，其知识发现和数据挖掘的性质还不是十分的明显，这也说明了图像数据挖掘理论与技术处于初期发展阶段的特点。也就是说在图像数据挖掘与其他相关理论与方法之间可能还存在一个模糊的界限，但是也不能因为这个原因就否定图像数据挖掘与知识发现。下面对图像数据挖掘与知识发现的国内外研究现状进行分析、解释。

目前有很多研究者对图像数据挖掘正在进行研究，但是，总体而言，图像挖掘的研究还处于初期实验阶段，从不同的角度对图像挖掘进行了理论性的探讨和初步的实验，还缺乏一个完整的切实有效的方法体系。虽然图像数据挖掘的许多相关技术发展已经相当成熟，但是图像数据挖掘本身仍然处于初步实验阶段，有许多问题需要我们进一步研究。同时，由于遥感图像数据本身具有海量数据、模糊性、多光谱(高光谱)等特点，迫切需要研究一些新理论、新方法，并引入到遥感图像数据挖掘中来，需要对图像数据挖掘的总体框架进行进一步修改和完善，同时对图像数据挖掘的关键技术进行进一步的研究和开发。总之，遥感图像数据挖掘是一个非常有前途，同时又是一个难度较大，具有挑战性的研究课题，为了促使图像

数据挖掘的进一步发展，重点应该解决以下问题:

对图像数据挖掘的总体框架和研究思路进一步修改完善，研究一些适合于图像数据挖掘的新理论、新方法;研究切实有效的为图像数据挖掘服务的图像特征的提取和组织存储方法;设计高效的基于内容的图像索引和检索技术，以便于对海量的图像数据集进行快速有效地访问;设计具有较强的语义功能的图像查询语言;研究各种图像知识和图像模式的可视化表达的技术;研究高效的图像知识的存储和管理技术;研究基于图像数据挖掘的知识的图像的智能化处理技术，如基于知识的图像分类、基于知识的图像检索、基于知识的目标识别以及开展图像数据挖掘的相关领域的应用研究工作，如基于图像数据挖掘的全球变化研究、基于图像数据挖掘的土地利用/土地覆盖变化研究等等。

4 数据挖掘与概念形成

4.1 数据挖掘与概念形成

数据挖掘首次出现在 1989 年 8 月举行的第十一届国际联合人工智能学术会议上。数据挖掘被定义为:“从数据中发现隐含的、先前不知道的、潜在有用的信息的非平凡过程”。数据挖掘的过程，即从数据库中发现知识的过程，可以理解为从数据库中形成概念的过程。

概念形成是人脑学习的一个重要特征，从概念形成去探讨人脑学习，从而探讨通过数据库中大量的数据的学习从而产生概念和知识的过程，被认为是一个行之有效的途径。所谓概念，就是在头脑里所形成的反映对象的本质属性的思维形式。把所感知的事物的共同本质特点抽象出来，加以概括，就成为概念，概念都具内涵和外延，并且随着主观、客观世界的发展而变化。如果能够建立一种数学的形式化的数据结构将概念的内涵和外延以及概念与概念之间的不同层次的抽象关系表达出来，将可以对数据挖掘和知识发现的过程进行有效地分析和处理。形式概念分析理论正是提供了这样的一个工具，形式概念分析理论所形成的形式化体系可以很好地利用数学的方法描述概念的形成过程。

4.2 数据挖掘的过程

数据挖掘和知识发现的过程可以根据状态空间理论作一个很好的解释。李德毅教授提出了以发现状态空间理论作为KDD的总体框架发现状态空间是一个三维立体空间，是发现系统实施多种算法的运作空间。在一个二维的平面基底—知识基上逐步抽象，关系数据库可以抽象地看成一个二维通用大表，纵向为属性，横向为元组。根据知识发现任务，在原始的数据库经过查询、选择(或抽样)、统计和压缩等数据聚焦处理后，形成宏元组，是发现状态空间的基底，也可以认为是初始的知识模板。在发现状态空间进行多种知识汇集和发现操作。模板方向，即面向知识模板的操作，是从微观到宏观的发现知识的操作。由一块知识模板上升到抽象级别更高的另一块模板，是提高知识抽象度的操作，是以归纳为核心的知识发现活动。对于空间数据挖掘对应的状态空间，还增加了一个尺度维，在尺度维上表达了空间数据由细到粗多比例尺或多分辨率的几何变换过程。面向尺度的操作是对空间数据由细到粗的计算、变换、概括、综合过程。

数据挖掘也可以看成一个从不同的视角，从低层到高层，不断地抽象，不断地产生抽象层次更高的概念，从而产生知识的过程。

5 概念格

概念格，也称 Galois 格，是形式概念分析理论的基本数据结构。概念格的基本思想就是将每一个概念用一个节点来表示，对概念进行形式化的表达，称之为形式概念。每个形式概念由两部分组成:外延，即概念所覆盖的实例，是概念所包含的对象;内涵，即概念的描述，也就是该概念覆盖实例的共同特征。形式概念分析理论通过数学的形式化语言将概念的内涵和外延表达出来。

概念格可以通过 Bass 图体现这些概念之间的泛化和特化关系，反映数据中所蕴含的概念之间的相互关系。概念格是进行数据分析的一种十分有力的工具。

从数据库中发现知识的过程，可以理解为从数据库中形成概念的过程。概念格理论提供了这样一个反映概念形成的形式化工具，将概念的内涵与外延作为概念的单元来进行形式化的表达。

5.1 单值属性的形式背景

形式背景是一个集合，结构为 (O, A, R)，在这里，O 和 A 是集合，而 R 是在。和 A 之间的一个二元关系。元素 0 和 a 分别叫做(形式化)对象和(形式化)属性。如果对于每个属性项，我们只关心它是否有值，如果该项有值，我们就用 1 来表示，否则就用 0 表示，这样得到的形式化背景就是单值属性背景。例如，对于例 1 的数据，我们可以定义单值属性的形式背景如表 1 所示。利用这种单值属性的形式背景，可以十分方便地对事务型数据进行处理。

表 1　单值属性形式背景表

	a	*b*	*c*	*d*	*e*	*f*
T1	1	1	0	1	0	0
T2	1	1	1	1	0	0
T3	1	1	0	1	1	0
T4	0	1	0	0	1	1
T5	1	1	0	1	0	1
T6	1	0	1	1	1	0

表 2　形式背景所生成的形式概念

(0, {a, b, c, d, e, f})	(1, {a, c, d, e})	(1, {a, b, d, e})
(1, {a, b, c, d})	(1, {a, b, d, f})	(2, {a, c, d})
(1, {b, e, f})	(4, {a, b, d})	(2, {a, d, e})
(2, {b, f})	(2, {b, e})	(5, {a, d})
(5, {b})	(3, {e})	(6, {∅})

5.2 多值属性的形式背景

在进行数据处理的过程中，更多的情况是，每个属性项具有多种值，例如，对于属性项“颜色”、“重量”、“性别”、“级别”等，都具有多种属性值，对于这种具有多种值的属性就称为多值属性。

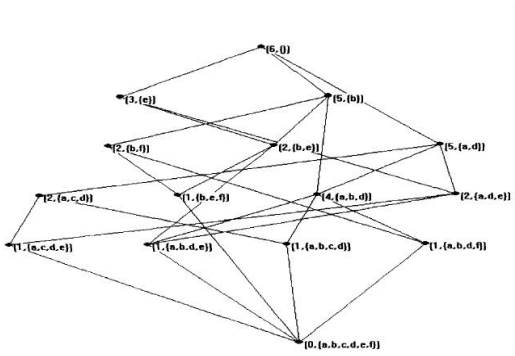

图 1　根据表 2 生成的 Hasse 图

6　概念格的知识表达和处理

概念格实际上是一种概念格节点之间的关系的表达。因为概念格具有清晰的数学背景，概念格的属性可以以一种整洁的、可计算的方法进行证明。利用概念格可以用来进行知识的表达与处理。如果给定一个对象集合 G 的一个子集，求取其相应的属性集合 M 的子集时，可以通过在概念格中寻找包含该对象子集的“最小”概念来实现。

规则是知识表达的一种最重要的方式，具体的包括关联规则、分类规则、特征规则等。基于概念格的概念之间的相互关系可以进行知识规则的表达和处理。

(1) 关联规则:关联规则反映概念之间的蕴含关系和依赖关系。

(2) 分类规则:分类规则用于对一种类型的概念与其他类型的概念进行区别。

(3)特征规则:特征规则用于反映一个类型独立于其他类型的特征，反映的是某个概念。

7　结 论

概念格理论是一个利用数学的形式化的方法描述概念形成过程的形式化的方法，可以作为规则表达的自然基础。本章详细分析讨论了概念格的基本理论;分析了基于概念格的知识的表达与处理方法，将关联规则、分类规则、特征规则用统一的规则形式“AFB”来表达，从而建立了集关联规则挖掘、分类规则挖掘和聚类规则挖掘为一体的

统一的数据挖掘的框架。本章对基于概念格的数据挖掘的原理进行了分析，研究讨论了相关的数据挖掘算法，研究出两种集概念格的构建和 Hasse。

参考文献

[1]. Wille R.Restructuring lattice theory： an approach based on hierarchies of concepts[M]//Rival I，ed. Ordered Sets. Dordrecht：Reidel，1982：445—470

[2] 王娜. 基于概念格的知识获取 [J]. 科技创业, 2010, 6(4): 118-120.

[3] 张文修,魏玲,祁建军. 概念格的属性约简理论与方法[J]. 中国科学(E 辑), 2005, 25(5): 490-495.

[4] 宋炜,张铭. 语义网简明教程[M]. 北京:高等教育出版社, 2004.

[5] 谢志鹏,刘宗田. 概念格的快速渐进式构造算法[J].计算机学报, 2002, 35(6): 628-639.

[6] 杨帆,翟岩慧,曲开社,李德玉. 基于形式概念分析的词义理解研究[J]. 2011, 38(10): 189-191.

[7] 曲开社,翟岩慧.偏序集、包含度与形式概念分析[J].计算机学报,2006,29(2):32- 33.

[8] 马叔良等.离散数学[M]. 北京:电子工业出版社,199