加权概念格理论与应用

张素兰　著

科学出版社

加权概念格理论与应用

张素兰 著

科学出版社
北京

内 容 简 介

加权概念格是一种拓展概念格结构，具有数据约简和知识提取针对性强等特点。本书结合作者十几年来在该领域的研究成果，对加权概念格的理论和应用进行了较为系统和深入的阐述。

全书共分 6 章。第 1 章简要叙述了数据挖掘和概念格的相关理论；第 2～4 章重点介绍了加权概念格的理论，主要涉及加权概念格的定义、权值获取和构造方法，以及加权概念格的代数结构等；第 5～6 章主要探讨了加权概念格在图像语义自动标注和天体光谱数据挖掘中的应用。

本书可供从事软件工程、数据挖掘、机器学习、人工智能、图像理解和天文学等相关专业的科研人员阅读参考，也可作为高等院校计算机、自动化、电子工程等专业的高年级本科生与研究生的学习参考书。

图书在版编目 (CIP) 数据

加权概念格理论与应用/张素兰著. —北京：科学出版社，2014
ISBN 978-7-03-040520-3

Ⅰ. ①加… Ⅱ. ①张… Ⅲ. ①加权叠加语言－研究
Ⅳ. ①TP301.2

中国版本图书馆 CIP 数据核字(2014)第 087704 号

责任编辑：张 濮 刘志巧 / 责任校对：胡小洁
责任印制：阎 磊 / 封面设计：迷底书装

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

新科印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2013 年 12 月第 一 版 开本：720×1 000 1/16
2013 年 12 月第一次印刷 印张：8
字数：151 000

定价：38.00 元

（如有印装质量问题，我社负责调换）

前　言

概念格又称为 Galois 格，因其生动简洁地体现了概念之间的泛化和特化关系，具有典型的层次结构，同时又因其本质上体现了数据集上的实体和属性、概念上的内涵和外延的关系，从而成为一种有效的数据挖掘和知识表示工具，并已成功地应用于数字图书馆、文献检索、软件工程、本体、数据挖掘与知识发现等领域。因此，自 R. Wille 教授于 1982 年提出概念格以来，概念格的理论研究已经吸引了大量的国内外研究者。同时，它的研究对机器学习、数据挖掘、信息检索等领域实际问题的设计与实现具有指导意义。加权概念格是一种拓展的概念格结构，具有数据约简和知识提取针对性强等特点。本书主要对概念格模型的拓展结构——加权概念格的理论及其在图像语义自动标注和天体光谱数据挖掘中的应用进行较为系统的介绍。

本书首先对加权概念格的相关理论进行重点描述，然后探讨该理论在图像语义自动标注和天体光谱数据挖掘中的应用。全书共分为 6 章，第 1 章简要介绍数据挖掘和概念格的基本概念。第 2～4 章重点叙述了加权概念格的理论，主要包括加权概念格的定义、权值获取方法、加权概念格的代数结构及其知识提取的完备性，以及加权概念格的构造方法等。第 5～6 章讨论了加权概念格理论在图像语义自动标注和天体光谱数据挖掘中的应用。各章内容相对独立，章节之间又彼此紧密联系在一起，从基本理论介绍入手，再到具体的模型及方法，最后以实际的应用作为结束。

本书旨在进一步推动概念格的拓展模型理论研究，并为其在图像语义自动标注和天体光谱数据挖掘中的应用研究提供一种有效途径。本书的出版能够增进学者之间的交流，促进拓展概念格的理论及其应用研究的进一步发展。

本书的完成得到了太原科技大学计算机学院数据挖掘与智能信息系统实验室团队成员的大力支持，尤其是张继福教授提出了很好的建议，硕士研究生王欣欣和褚萌等做了大量的辅助性工作。另外，北京理工大学郭平教授对本书给予了重要的指导。在此，一并表示最诚挚的谢意！

本书所涉及的部分研究工作得到了国家自然科学基金项目(61373099)和太原科技大学博士启动基金项目(20132005)的资助，在此谨向国家自然科学基金委员会和太原科技大学表示衷心的感谢。

由于本人水平有限，时间紧迫，书中不妥之处在所难免，恳请读者批评指正。

作　者

2013 年 11 月

目　录

第1章 绪 论

科技的进步，特别是信息产业和互联网技术的发展，把我们带入了一个全新的大数据信息时代。随着计算机应用的普及和数据库技术的不断发展，数据库管理系统的应用越来越广泛。最近十几年中，数据库中存储的数据量急剧增大，大量信息给人们带来方便的同时，也带来了一系列问题。例如，信息量过大，超过了人们掌握、消化的能力；一些信息真伪难辨，给信息的正确运用带来困难；网络上的信息安全难以保障；信息组织形式的不一致性增加了对信息进行有效统一处理的难度等。人们意识到隐藏在这些数据之后的更深层次、更重要的信息能够描述数据的整体特征，可以预测发展趋势，这些信息在决策形成的过程中具有重要的参考价值[1]。

可见，激增的数据背后隐藏着许多重要的信息和知识，人们希望能够对其进行更高层次的分析，以便更好地利用这些数据，因此，数据和知识之间的鸿沟带来了对强有力的数据分析工具的需求，如何从这些浩如烟海的数据中挖掘出有用知识的需求激起了数据库中的知识发现(knowledge discovery in databases，KDD)，或者说数据挖掘(data mining，DM)的理论与技术研究的蓬勃发展。

1.1 数 据 挖 掘

1.1.1 数据挖掘的基本概念

数据挖掘又称为数据发掘或数据采掘，也称为知识提取(knowledge extraction)、数据考古学(data archaeology)、数据捕捞(data dredging)。它是一种从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、潜在的有用信息和知识的过程，是一种为决策支持服务的过程。更广义的定义是：数据挖掘是指从存放在数据库、数据仓库或其他信息库中的大量数据中提取人们感兴趣的、隐含的、尚未被发现的、有用信息和知识的过程[2]。

数据挖掘不同于传统数据分析及联机分析处理。传统数据分析是采用基于验证的方法，通过分析少量数据，了解已经发生了什么。联机分析处理是从不同角度、不同层次汇总、合并、聚集大量数据，以便多纬度多粒度观察、分析数据。而数据挖掘是采用基于发现的方法，通过分析大量数据，了解已经发生了什么，分析发生的原因并预测未来将会发生什么[3-4]。

许多人认为数据挖掘与知识发现同义，而另外一些人只是把数据挖掘视为知识发现过程的一个步骤。第一种观点在产业界、媒体和数据库研究界比较流行，而第二种观点却是从更广义的角度提出的。

1. 数据挖掘的过程

数据挖掘过程的一般步骤如图 1.1 所示[5-6]。

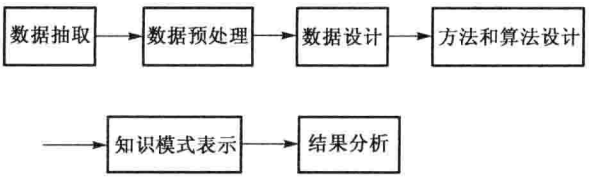

图 1.1　数据挖掘过程的一般步骤

1)数据抽取

大多数时候，与数据挖掘任务有关的数据是存储在应用数据库中的，而这些数据库往往是以应用为目的建立的，通常不能直接运行数据挖掘算法，需要进行必要的抽取和格式的整理工作。

2)数据预处理

数据预处理主要处理掉一些噪声数据(冗余的、不一致的)或添补一些丢失的数据，以便使被挖掘的数据保持完整和干净。

3)数据设计

数据设计的任务是对数据进行选择，该过程主要是去掉一些无关的属性，或者对数据量过大的数据库进行抽样等。

4)方法和算法设计

方法和算法设计主要是针对特定的数据挖掘任务，设计数据挖掘方法模型与高效的算法和相应的数据结构。

5)知识模式表示

知识模式表示主要从数据库或数据仓库中获取特定的知识类型，如分类、关联规则、聚类和序列模式等。

6)结果分析

结果分析由领域专家(domain expert)分析结果的可靠性、合理性和可用性，但是，有时需要对结果进行可视化处理。

从图 1.1 中可以看出，数据挖掘的核心步骤是数据挖掘方法和算法的设计，一个好的数据挖掘模型、一个好的算法(速度快、伸缩性好、结果容易使用且符合用户的特定需求)是影响数据挖掘效率的最重要的因素。

2. 数据挖掘系统的分类

由于数据挖掘是一个多学科交叉领域，所以根据数据挖掘的不同的研究期望会产生各种不同类型的数据挖掘系统。根据不同的标准，大致可以进行以下分类[2]。

(1)按挖掘对象的不同可分为：关系数据库、面向对象数据库、空间数据库、时态数据库、文本数据库、多媒体数据库、异构数据库、遗产数据库、Web 数据库等。

(2)按挖掘任务的不同可分为：分类、预测、时间序列分析、聚类分析、关联分析预测、偏差分析和数据挖掘可视化等。

(3)按挖掘方法的不同可分为：机器学习、统计方法、神经网络法、粗糙集方法和云模型等。

(4)按应用领域的不同可分为：零售业、银行、邮电、保险、医疗保健、运输业、行政司法、生物信息处理等。

(5)按挖掘知识的不同可分为：

广义型知识——事物共同性质的知识；

特征型知识——反映事物各方面特征；

差异性知识——事物之间属性的差别；

关联性知识——事物之间的依赖关系；

预测型知识——由过去和现在预测未来；

偏离型知识——揭示事物偏离常规的异常现象。

3. 数据挖掘的对象

数据挖掘的对象原则上可以是各种存储方式的信息。目前的信息存储方式主要包括关系数据库、数据仓库、事务数据库、高级数据库系统、文件数据库和Web数据库。其中，高级数据库系统包括面向对象数据库、关系对象数据库，以及面向应用的数据库[1]。

1)关系数据库

关系数据库由表组成，每个表有一个唯一的表名。属性(列或域)集合组成表结构，表中数据按行存放，每一行称为一个记录。记录间通过键值加以区别。关系表中的一些属性域描述了表间的联系。关系数据库是目前最流行、最常见的数据库之一，为数据挖掘研究工作提供了丰富的数据源。

2)数据仓库

数据仓库可以把来自不同数据源的信息以同一模式保存在同一个物理地点，其构成需要经历数据清洗、数据格式转换、数据集成、数据载入和阶段性更新等过程。尽管数据仓库中集成了很多数据分析工具，但仍需要数据挖掘等更深层次的、自动的数据分析工具。

3)事务数据库

一个事务数据库由文件构成，每条记录代表一个事务。典型的事务包含唯一的事务标识，多个项目组成一个事务。事务数据库可以用额外附加的关联表记录其他信息，而更深层次的数据分析只能利用数据挖掘思想来解决。

4)面向对象数据库

面向对象数据库是基于面向对象程序设计的范例，其每一个实体作为一个对象。与对象相关的程序和数据封装在一个单元中，通常用一组变量描述对象，等价于实体关系模型和关系模型中的属性。对象通过消息与其他对象或数据库系统进行通信。

5)关系对象数据库

关系对象数据库的构成基于关系对象模型。关系对象数据库在工业和其他应用领域使用得越来越普遍。与关系数据库上的数据挖掘相比，关系对象数据库上的数据挖掘更强调操作复杂的对象结构和复杂数据类型。

6)面向应用的数据库

面向应用的数据库包括空间数据库、时态数据库、时间序列数据库、文本数据库、多媒体数据库等。空间数据库包含空间关系信息，如地理数据库、医学图像数据库和卫星图像数据库等；时态数据库和时间序列数据库均存储与时间有关的信息，时间序列数据库存储随时间顺序变化的数据；文本数据库是包含用文字描述的对象的数据库，这里的文字不是通常所说的简单的关键字，可能是长句子或图形；多媒体数据库中存储图像、音频、视频等数据。

7)文件数据库

文件数据库又称为嵌入式数据库，是将整个数据库的数据信息保存在一个索引文件中，以便于数据库的发布。由于数据保存在单一文件中，所以数据库的部署和发布都比较简单，适用于内嵌在应用程序中，但是这类数据库的容量不能过大。

8)Web 数据库

Web 数据库也称为网络数据库。对于一个 Web 数据库，用户将浏览器作为输入接口，输入所需要的数据，然后浏览器将这些数据传递给网站，网站再将这些数据进行处理。因此，Web 数据库是 Web 技术与数据库技术相结合的产物，是存放和管理可以在互联网上访问的大量信息的数据库系统。

1.1.2 数据挖掘的任务

数据挖掘的任务是从海量数据中发现隐含的、有意义的知识。它的主要任务就是从实例集合中找出容易理解的规则和关系。这些规则可以用于预测未来趋势、评价顾客、评估风险或简单地描述和解释给定的数据[7]。通常数据挖掘的任务包括以下几个部分：分类、关联分析、聚类、预测、孤立点分析、时间序列分析和数据挖掘可视化等。

1. 分类

分类(classification)就是构造一个分类函数(分类模型)，把具有某些特征的数据项映射到某个给定的类别上。该过程由两步构成：模型创建和模型使用。模型创建通常是指通过对训练数据集的学习来建立分类模型；模型使用是指使用分类模型对测试数据和新的数据进行分类。其中，训练数据集是带有类标号的，也就是说，在分类之前，要划分的类别是已经确定的，分类模型通常是以分类规则、

决策树或数学表达式的形式给出的。最常用的数据分类方法有判定树(decision tree)、贝叶斯分类(Bayesian)、神经网络(neural network)、k 最近邻分类、基于案例的推理(case-based reasoning，CBR)、概念格方法、粗糙集方法和模糊集方法。对于分类来说，主要是提高分类的准确率与效率。

2. 关联分析

关联分析(association analysis)是寻找数据集中项与项之间的相互联系，发掘与描述蕴涵在关联规则中的有用知识的过程，即用量化的形式语言描述其内在规律的知识模式，具有很强的信息处理能力。它可以找出隐藏在数据背后的关联信息，这对现实生活中的商业决策具有重要意义。

关联规则模式由 Agrawal、Imielinski、Swami 于 1993 年提出[8]，它是描述在一个事务(项目)中物品(交易)同时出现的规律的知识模式，即通过量化的数字描述物品甲的出现对物品乙的出现的影响程度。关联规则模式的提取通常可以分为两步：找出满足用户的最小支持阈值的频繁模式集；提取出满足用户的最小置信阈值的关联规则。从大量商务事务记录中发现有趣的相关联系有利于许多商务决策的制定，如市场规划、广告规划、分类设计、交叉购物和贱卖分析等。又如，在现今中国贷款购买住房和汽车的顾客中，发现 70%的年龄为 35～45 岁，那么银行就可以通过分析这些客户的特点从而调整一些相应的政策，以便将贷款发放给这类客户群体。自从 Agrawal 等提出从大型数据库中挖掘关联规则以来，关联规则的挖掘已广泛地应用在电子通信行业、信用卡公司、股票交易所、银行和超市等。目前，国内外研究者从多种角度、多种渠道研究基于各种数据模型上的关联规则的提取。

3. 聚类

聚类(clustering)就是将数据项分组成多个类或簇，类之间的数据差别应尽可能大，类内的数据差别应尽可能小，即“最小化类间的相似性，最大化类内的相似性”原则，与分类模式不同的是，聚类中要划分的类别是未知的，它是一种不依赖于预先定义的类和带类标号的训练数据集的非监督学习，不需要背景知识，其中类的数量由系统按照某种性能指标自动确定。聚类分析是数据挖掘应用极其广泛的功能。

聚类分析是一种根据对象属性标识对象集的类(组、簇)的过程，对象按某种

聚类准则聚类后，对象组内的相异性最小，组间的相异性最大。例如，在保险业上，聚类能够帮助保险公司分析客户群体特征及其消费行为特征规律，进一步指导其对潜在客户的关注程度或者开展针对性的市场调研和新产品开发活动。

4. 预测

预测(prediction)就是通过分析历史数据，找出规律，建立模型，并使用该模型对未来数据的种类和特征进行分析，推导出其中可能存在的变化趋势，最终抽象出形式化的数据模型，进而由此模型对未知数据的种类及特征进行预测。预测是构造和使用模型评估未标号的样本类，或评估给定的样本可能具有的属性值或数据值区间。常见的预测方法主要包括线性回归、多元回归和非线性回归等。通常采用预测方差来度量预测结果的精度和不确定性。

5. 孤立点分析

孤立点(outlier)分析，又称为离群点分析，是一种挖掘出与数据的一般行为或模型不一致的数据对象的过程。孤立点是对差异和极端特例的描述，如聚类外的离群值，大部分数据挖掘方法都将这种差异信息视为噪声而丢弃，然而在一些应用中，就是这些极少数离群数据常常比其他常规数据的挖掘更有价值。因为这些数据可能就是一些非正常信息的真实反映，如信用卡的欺骗检测，通过检测一个给定账号与其历史上正常的付费相比较，可以根据某次付款数额特别大这一异常数据来发现信用卡可能被欺骗性使用。所以，孤立点分析往往可以发现一些真实的但又出乎意料的知识。实际生活中，孤立点分析已广泛地应用在网络入侵检测、贷款证明的审核和信用卡恶意透支等领域。

6. 时间序列分析

时间序列分析(time series analysis)是描述基于时间或其他序列的经常发生的规律或趋势，并对其建模，一个典型的例子就是：在购买计算机的顾客当中，70%的人会在半年内购买打印机。时间序列模式分析将关联模式和时间序列结合起来，重点考虑数据之间在时间维上的关联性，有3个参数的选择对序列模式挖掘的结果具有很大的影响：①时间序列的持续时间 t，也就是某个时间序列的有效时间或者是用户选择的一个时间段；②时间折叠窗口 $w(w \leqslant t)$，即在某段 w 时间内发生的事件可以被看作同时发生的；③所发现模式的时间间隔。

7. 数据挖掘可视化

数据挖掘可视化(data mining visualization)技术是建立在可视化和分析过程的基础上，它以刻画结构和显示数据的功能性，以及人类感知模式、倾向和关系的能力为基础，用可视化来加强数据挖掘处理。可视化可以使数据和挖掘结果更容易理解，允许对结果进行比较和检验，数据挖掘可视化功能使计算和数据内容对人是可理解的，它把信息转化为我们的感觉和大脑可以分析和遵循的经历，使数据和挖掘结果更容易理解和验证。数据可视化方法包括几何投影方法、分层表示方法、基于像素的方法等，如二元散点图、平行坐标、散布矩阵、多维测量图、放射性可视化图、数据立方体以及自组织映射图(self-organizing map，SOM)等。

1.1.3 数据挖掘的方法

为完成上节描述的数据挖掘的任务，采用的数据挖掘方法(技术)有：决策树、贝叶斯信念网络、模糊集、粗糙集、概念格和遗传算法等。

1. 决策树

决策树(decision tree)是一个类似于流程图的树结构，树的每个非叶结点均表示被考察数据项目的一个测试或决策，根据测试结果，选择某个分支。构造决策树采用自上而下的递归构造，直到叶结点，便形成了一个决策，即如果训练实例集合中的所有实例是同类的，就将其作为一个叶子结点，结点内容为该类别的标记。否则，根据某种策略确定一个测试属性，并按属性的各种取值把实例集合划分为若干个子集合，使每个子集上的所有实例在该属性上具有相同的属性值。然后，再依次递归处理各个子集，直到得到满意的分类属性为止。典型的算法有ID3[9]和C4.5。

2. 贝叶斯信念网络

贝叶斯信念网络(Bayesian belief networks)是概率分布的图形化表示。它是一种有向无环图，其结点表示属性变量，边表示属性变量间的概率依赖性，与各结点相关的是描述相应结点与其父结点之间关系的条件概率分布。

3. 模糊集

模糊集(fuzzy set)是一种表达和处理不确定性的重要方法。随着要处理的数

据日益增多，数据库模型中出现了多种形式的不确定性，如不精确、不完全、不一致、含糊等。模糊集合论是一种用隶属程度来表示处于中介过渡的事物对于差异一方所具有的倾向性程度的数学理论，是用精确的数学语言对模糊性进行描述的方法。因此，模糊集能够利用不确定性使系统的复杂性变得可以处理。当精确输入不可能或太昂贵时，模糊系统建模方法就是一种强有力的数据分析方法。

4. 粗糙集

粗糙集(rough set)理论是波兰数学家 Pawlak 于 1982 年提出的[10]，该理论是一种新的处理含糊性(vagueness)和不确定性(uncertainty)问题的数学工具。粗糙集理论是用一个集合的上下界来定义的。下界中的每个成员都是这个集合的成员，而上界中的每个非成员也一定是这个集合的非成员。粗糙集中的上界由下界和边界区域的并集构成。边界区域的成员可能但不一定是这个集合中的成员。因此，粗糙集可以被看成是一个有三级成员函数(是、非、可能)的模糊集，是能够处理数据不确定性的一种数学概念。目前粗糙集理论已广泛应用于数据挖掘、机器学习、决策支持、模式识别、专家系统、归纳推理等领域。

5. 概念格

概念格(concept lattice)也称为 Galois 格，又叫作形式概念分析，是 20 世纪 80 年代初由德国的 Wille 教授提出的[11]，它提供了一种支持数据分析的有效工具。概念格的每个结点是一个形式概念，由内涵(属性集)和外延(拥有该属性集的对象集)两部分组成，这种格的结构及其相应的 Hasse 图形式，反映了一种概念层次结构，本质上体现了实体(对象、记录、交易)和属性(特征、项目)之间的关系，概念内涵和外延的统一，生动而简洁地表明了概念之间的泛化和特化关系，成为一种很有用的数据分析和知识提取工具，这种形式概念分析工具已经被成功地应用于数字图书馆、文献检索、软件工程、基于案例数据分析和知识发现等领域。

6. 遗传算法

遗传算法(generic algorithm)的产生受自然界生物进化现象的启发，问题的解用一定长度的编码表示(最常用的编码方式是二进制编码)，个体的编码称为“基因型”或“染色体”，其编码对应的实际意义称为“表现型”。在种群中，每个个