基于形式概念分析的数据挖掘方法的研究

胡森博

（大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘要 形式概念分析是由德国的Wille教授在20世纪80年代初所提出的，其核心数据结构概念格，也被称为Galois格，准确而简洁地描述了概念之间的层次关系，已成为一种极其重要的 知识表示方法。作为一种优良的数学工具，概念格已经被广泛的应用于知识表示、数据挖掘、信息检索等众多领域。讨论了概念格的基本原理，介绍了概念格的相关构造算法，并对各种建格算法加以论述，接着分析了在数据挖掘中的应用，最后提出了未来概念格的相关研究方向。

关键词 形式概念分析，概念格，数据挖掘

中图法分类号 TP311 **文献标识码** A

Research on Data Mining Method Based on Formal Concept Analysis

HuSenbo

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

Abstract Formal Concept Analysis (FCA), which was elaborated by Professor Wille of German in the eighties of the twentieth century. The Concept Lattice, which is the core data structure of FCA and is also called Galois Lattice, can describe the hierarchy relationship between concepts and has become an important method for the representation of the knowledge. As a kind of excellent mathematic tool, Concept Lattice has been widely applied in knowledge representation, data mining and many other fields. This article discusses the basic principles of the concept lattice and introduced the construction algorithm which related to the concept lattice, and a variety of algorithms to build the grid to be addressed, and then analyze its practical application. Finally, introduced the future research directions related to the concept lattice.

Keywords Formal Concept Analysis,Concept Lattice ,Data Mining

1 引言

在计算机与网络信息技术飞速发展的今天，各个领域的信息与数据都急剧增加，同时由于人类的参与使得数据与信息中的不确定性更加显著，信息与数据间的关系更加复杂。如何从大量杂乱无章和强干扰的数据中挖掘出来潜在的、新颖的、正确的、有利用价值的知识，这给智能信息处理提出了严峻的挑战，由此产生了人工智能领域研究的一个崭新领域——数据挖掘（DM）和数据库知识发现（KDD）。目前已有许多的数据挖掘工具，比如神经网络、遗传算法、支撑向量机、决策树、粗糙集、形式概念分析等等。在DM和KDD诸多方法中形式概念分析（Formal Concep Analysis，FCA）对于处理复杂的信息不失为一种有效的方法。

而概念格则是FCA的核心数据结构。概念格理论最早由Wille R等提出，是应用数学的分支，它来源于哲学相关领域内对概念的理解。随着研究的深入，很多学者逐渐认识到概念格自身结构的巨大优势，研究从开始的单纯理论扩展发展到

收稿日期：2016-10-9
作者简介：胡森博（1994-）男，硕士生在读。

理论与实际应用相结合，并且融合交叉多个相关理论，成为许多专家学者关注的热点。作为数据分析和知识处理的形式化研究方法，概念格在知识发现、信息检索等方面均得到了广泛的应用。概念格理论的研究不仅能用于解决知识发现领域中所涉及的关联规则、蕴含规则、分类规则的提取，还能够实现对信息的有机组织，减少冗余度，简化信息表，所以对于概念格理论及其构造方法的研究具有十分重要的意义。

本文首先介绍了形式概念分析的基本概念和其形式背景，又介绍了概念格的基本概念以及构建形式背景和概念格，进行简单的概念分析，并从分析中获得关联规则，再通过对构造概念格的算法进行论述，使人们更加了解构造概念格的算法，最后介绍了概念格在数据发掘方面的应用，方便以后的运用。

2 基本概念

2.1 形式背景

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。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 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。

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。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。

定义 2.1 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)∈R 或 gRm 表示对象 g 具有属性 m。

定义 2.2 设 A 是对象集合 G 的一个子集，我们定义 f(A)={m∈M|∨g∈A, gRm}(A 中对象共同属性的集合)。

相应的地设 B 是属性集合 M 的一个子集，我们定义 g(B)={g∈G|∨m∈B,gRm}（具有 B 中所有属性的对象的集合)。

定义 2.3 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，则有下面的一些性质:

(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) $A\subseteq g(B)\Leftrightarrow B\subseteq f(A)\Leftrightarrow A*B\subseteq R$

定义 2.4 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 2.5 $C_1=(A_1, B_1)$，$C_2=(A_2, B_2)$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。

一般来说，由一个规范的形式背景及其概念格可挖掘到大量的关联规则。用图形方式表示概念格是传播知识和建立透明的高层次知识表示的有效方法。知识的各种连接和解释可以通过各种概念格的 Hasse 图来实现可视化。

定义 2.6 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，如果 $C_1 < C_2$，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 概念格的构造

自 1982 年提出形式概念分析以来，概念格作为形式概念分析理论中的一种核心数据结构，已经在知识发现领域、软件工程领域、知识工程领域、经济分析及 Web 挖掘等众多领域取得了广泛应用。然而在应用概念格时，概念格的构造效率始终是一大难题。它的构造过程实际上是概念聚类的过程，是应用形式概念分析的前提。通常，概念格的大小是在指数量级上的，而且要处理的数据又多数是海量的，因此概念格构造算法的研究是形式概念分析中的一个主要问题。

目前，学者们提出的构造算法主要可分为两大类:批处理算法和渐进式算法。

3.1 批处理算法

使用批处理算法构造概念格要完成两项任务：① 生成所有的格节点，即所有概念的集合；② 建立这些格节点间直接前趋直接后继关系。按这两项任务完成的次序不同，我们可以将批处理算法分为任务分割生成模型和任务交叉生成模型。任务分割生成模型是首先生成全部的概念集合，然后再找出这些概念之间的直接前驱/直接后继关系；任务交叉生成模型是在生成概念的过程中同时确定概念之间的关系。

已有的任务分割生成模型算法包括 Nourine 算法和 Alaoui 算法等。其中，NextClourse 算法、Chein 算法、NextClourse 算法和 Chein 算法只生成所有概念集合，并未确定概念之间的父子关系。而 Alaoui 算法的功能是生成概念间的父子关系。已有的任务交叉生成模型算法包括:Bordat 算法和 LATTICE 算法等。下面简要介绍一下几个典型的算法。

NextClourse 算法是著名的概念格构造算法，它按字典序从形式背景中计算出所有的概念。具体做法是:以位向量来表示特征子集，某位为 1 表示含有该特征，为 0 表示不包含该位所对应的特征。位向量从小到大的次序反映了特征集合上一种字典排序。通过对位向量的枚举来生成所有的内涵集，从而得到所有的概念。然而 NextClourse 算法本身并不能直接得到格结构。

Chein 算法采用自底向上的次序生成所有的节点，即从对象概念开始逐层地求取两个概念的内涵交集来生成新的概念，并不断地去除已生成的概念。这个算法在对象数据较多的情况时，在开始阶段需要产生大量的重复节点，因此在数据分析中使用的较少。

Alaoui 算法则是用计算概念节点之间的父一一子关系，从而将概念节点集合链接成概念格结构的一个简单算法，它的时间复杂度为 $O_{(|C|^2)}$，其中|C|表示概念格中节点的数目。

Nourine 算法首先生成所有的概念节点,这些概念节点通过一个词典排序树进行索引，接着通过一个算法计算出所有节点的父子关系。其中生成所有节点的时间复杂度为 $O_{((|O|+|A|)*|O|*|C|)}$,完成这些节点之间链接的时间复杂度为 $O_{((|O|+|A|)*|O|*|C|)}$。

Bordat 算法从格的顶端节点开始构造，为每个节点生成它的所有子节点并完成子节点到父节点之间的链接。其算法的思想在于:如果当前节点为(X_1, Y_1)，找出属性子集 $Y_2 \subseteq A - Y_1$，使得 Y_2 在 X_1 中能保持完全二元组的性质，则 $Y_1 \cup Y_2$ 构成了当前节点的一个子节点的内涵。这个算法存在的一个问题是，每个二元组被生成多次。为了避免重复，必须检查每个节点是否已经生成。

LATTICE 算法是一个较为高效的算法。对于第一个过程，它利用了类似于 NextClourse 算法的方法来为概念格中的每个节点生成它的所有子节点;对于第二个过程，它则将所有已经生成的概念节点通过一棵词典排序树组织起来，作为一个索引结构，从而可以快速地判断某个节点是否已经生成。实验结果表明，LATTICE 算法要明显优于 NextClourse 算法。

3.2 渐进式算法

渐进式算法的基本思想是将当前要插入的对象与格中所有的概念求交，根据交的结果进行不同的操作。典型的算法有:Godin 算法、Capineto 算法,Addintent 算法。Ho 等也提出了一个渐进算法，但和 Godin 算法的思想基本相同。下面简单介绍几个这类算法。

Godin 算法在插入一个新实例时,格中的节点被分为三类:一类是不变节点，这些节点的内涵和要插入实例的特征集没有交集，它们将新格保持不变;第二类是更新节点，这些节点的内涵包含于要插入实例的特征集，因此只需将其外延更新，包括要插入实例即可;第三类是新增节点，当所有要插入的实例的特征集与原来格中某个节点的内涵交集在格中没有出现过时，需要增加新的节点，该新节点的内涵即为该交集。可以证明，新增节点的父节点必然是某个新增节点或者更新节点，这使得连接过程很容易实现。Godin 还给出了一个改进算法，当一个新实例插入时，不必对格中所有节点进行检查，只需检查那些和新实例有共同属性的节点。可以用通过维护记录每个属性首次在格中出现的指针来实现。

Capinet。算法与Godin算法的基本思想类似，它将生成新概念的条件分为:“交为空”、“交已经存在”和“交包含在已有概念中”。主要的不同处出现在连接过程中。Capineto 算法是找到该新节点的最小上界和最大下界，删除它们之间的边，并将其连接到新概念。

Addintent 算法不但生成概念集,也生成概念的格结构。算法渐进地将下一个对象合并到前面对象已生成的图结构中。因此，该算法更适合于既需要得到概念集又需要得到格结构的相关应用。实验结果表明，在某些时候，Addintent 算法构造整个格结构的时间与其它算法从概念集中构造格结构的时间相比，所用时间要少很多。

4 形式概念分析在数据挖掘中的应用

随着计算机技术的不断发展,计算机应用领域不断扩大。人们收集和处理数据的能力和数量在不断变大,直接从大量的数据中找到用户感兴趣或对用户有指导意义的知识的难度也在不断变大,从而出现了“数据丰富、知识贫乏”的窘境。因此,数据挖掘技术得到了广泛的研究。

关联规则是从数据库中提取知识的主要表现形式,也是数据挖掘研究的核心内容之一。形式概念分析以概念格形式把数据有机地组织起来,数据之间的关系通过概念格节点的特化一例化关系体现出来,体现了概念的内涵和外延的统一,所以概念格非常适合作为规则发现的基础性数据结构用来发现规则型的知识。概念格应用于关联规则提取,是概念格在数据挖掘中应用的最广、取得成果最丰的一个领域。国内外学者在基于概念格提取关联规则方面都有深入的研究。Godin.R 等提出了基于概念格模型提取蕴涵规则的方法。他们首先由形式背景构造概念格,再从格中产生连接规则,最后再去掉冗余的蕴涵规则。他们把蕴涵规则分为连接蕴涵规则(Conjunction Implication Rules)和分离蕴涵规则(Disjunctive Implication Rules),并给出了蕴涵规则的确定算法。但蕴涵规则属于确定性的规则（精确规则），不具备描述概率规则（近似规则）的能力。Missaoui.R 等对 Godin 等提出的基于概念格模型的蕴涵规则进行了扩展,提出了从概念格中提取近似规则的算法。在国内,王志海等提出了概念格中提取规则的一般算法和渐进式算法。他们把概念格中的节点根据其双亲节点个数的不同,分为只有一个双亲节点、两个双亲节点和个双亲节点的情况,分别给出了其中规则的提取原则。胡可云等提出了概念格节点中对象用对象的势来代替,并用概念格来渐进产生最大项集集合,再提取规则的算法。赵奕等对胡可云等提出的方法中的一些不足,提出了一种改进的递增修正规则挖掘的算法,以适应实际数据不断递增或递减更新时的要求,并记录概念格节点在数据中出现的频率值,在无需构造全格的情况下提取规则。谢志鹏等提出了利用内涵缩减来提取关联规则,试图减少提取规则的数目。由于提取出的规则很多具有依赖关系,也就是说在提取的规则中许多是冗余的,虽然可以再进行冗余规则的去除,但毕竟耗费时间和空间。因此,人们又开始研究提取无冗余规则的问题。Y.Bastide 等提出了一种最小无冗余关联规则的定义,并提出了基于 Apriori 算法的改进算法,通过提取频繁封闭项集来获取最小无冗余关联规则。

N.Pasquier 等提出了由封闭项集构成格，再提取关联规则的思想，P. Valtchev 等提出了利用概念格获取频繁封闭项集的算法框架和相应的算法;针对在数据挖掘中概念格的理论研究，Y.Bastide 等定义的最小无冗余关联规则,我们通

过改变格的结构，形成所谓的量化封闭项集格，并提出采用多种方法直接从量化封闭项集格中求取最小封闭项集，以达到提取这种规则的目的。

概念格还可应用于数据挖掘中的分类知识的获取。Sahami.M 首先根据条件属性构造出概念格，然后从格中提取出分类规则用于支持对象的分类，并形成了被称为RULEARNER 分类系统；Njiwoda 等使用学习参数来生成部分格，采用投票的方式对新对象的分类进行群体决策，并通过特征选择方法设计了LEGAI-F 分类系统；胡可云等通过改进 Bordat 的建格算法，使之适应集成挖掘分类和关联规则的需要，成功实现了关联规则和分类规则在概念格框架下的统一。

5 结束语

随着大规模数据库的广泛应用和的迅猛扩展，全球范围内数据库中存储的数据量迅速增大。如何从海量的、多样的数据中挖掘潜在的、有利用价值的信息，即数据挖掘，已成为当前知识发现的主要研究课题之一。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。目前，概念格正在广泛应用于机器学习、模式识别、专家系统、计算机网络、数据分析、决策分析等领域。

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]. 黄天民.格、序引论及其应用[M].成都:西南交通大学出版社,1998

[2]. Boda J P.Calcul pratique dutreillis de galois dunecorrespon-dence[J].MathSci Humaines,1986,96(2):31-47.

[3]. Wang Yuanyaun ,Hu Xuegang，A Fast Algorithm for Mining AssociationRules Based on Concept Lattice[C].In:IEEE The Third International Conference on Machine Learning and Cybernetices，Shanghai，2002，8:1687 一 1691P.

[4]. CheinM.Algorithmde recherche des sous-matrices premieres dunematrice[Z].1969.

[5]. 杨 强，赵明清.概念格研究进展[J].计算机工程与设计,2008,29(20):5293-5296.

[6]. 何淑贤,刘桂枝.形式概念分析及其应用进展[J].应用技术，2007(5):77-79

[7]. 谢志鹏,刘宗田.概念格与关联规则发现[J].计算机研究与发展,2000,37(12):1415-1421

[8]. 王甦菁，陈震.基于概念格的数据挖掘方法研究[J].计算机应用,2005,25(4):157-161

[9]. 徐清泉，朱玉文，刘万春，基于概念格的关联规则算法阴.计算机应用 2005，25(8)：1856-1860

[10]. Hu Xuegang,Wang DeYing，Liu Xiaoping，Guo Jun，Wang Hao，The Analysis on Model of Association Rules Mining Based on Concept Lattice and Apriori Algorithm[C].In: IEEE The Third International Conference on Machine Learning and Cybernetices，Shanghai，2002，8:1620-1624.