《智能信息处理》课程作业

概念格理论与方法及其研究展望

王晓丽

作业	分数[20]
得分	

2020 年 11 月 12 日

概念格理论与方法及其研究展望

王晓丽

(大连海事大学 计算机科学与技术 辽宁省大连市 中国 116026)

摘　要　概念格是基于属于某一概念的所有对象与其共有属性之间的二元关系而建立的一种概念层次结构，是一种知识分析与处理的有效工具。概念格理论与方法是形式概念分析研究中的基本内容，该研究已取得一系列的重要成果,主要集中在概念格模型推广、概念格构造、概念格约简、基于概念格的规则提取、概念知识空间、概念格的粒计算方法及概念格应用等研究方向。为了进一步促进形式概念分析的研究与发展，本文从概念格理论的起源、定义、研究内容、发展方向等方面总结了概念格的研究进展，最后探讨了概念格理论未来的研究趋势。

关键词　概念格；形式概念分析；规则提取；粗糙集；

Concept Lattice Theory and Method and Their Research Prospect

Xiaoli　Wang

(Computer science and technology，Dalian maritime university, LiaoningDalian,116026,China)

Abstract　Concept lattice is a Conceptual hierarchical structure, which is based on the binary relationship between all objects and their attributes. It is an effective tool for knowledge analysis and processing. Concept lattice theory and method are the basic topics in the study of formal concept analysis, and important achievements are obtained. The previous study mainly focus on the generalization of concept lattice models, concept lattice construction, concept lattice based rule acquisition, conceptual knowledge space, granular computing method for concept lattice and concept lattice applications. To further promote the study and the development of formal concept analysis theory, This papertry to summarize the research progress of concept lattice theory from the aspects of its origin, definition, research content and development direction. And finally it discusses the research trend of concept lattice theory in the future.

Key words　Concept lattice；Formal concept analysis；Rule extraction；Rough Set;

1 引言

概念格，是根据二元关系提出的概念层次结构，用于数据的分析和规则的提取[1]。从数据集中生成概念格的过程实质上是一种概念聚类的过程，通过Hasse图可以清楚的反应出概念间的层次结构，以及相互之间泛化与特化的关系，从而实现数据的可视化。它提出的初衷是希望通过形式化的方式刻画现实中的实体对象或抽象概念，并建立相应的层次知识结构,描述概念之间的泛化与特化关系。近年来，随着互联网、大数据和计算机技术的飞速发展，概念格作为数据挖掘和分析的重要方法和工具，在概念分析、概念展示、概念构造、概念关联等方面的优势愈发凸显，概念格与其他学科理论（认知计算、粒计算等）的融合发展也越来越深入，被广泛运用于数据挖掘、数据分析、软件工程、信息检索、人工智能、知识发现、本体研究、Web语义检索等领域。概念格的研究涵盖多方面的内容,主要工作包括概念格模型推广、概念格构造、概念格约简、基于概念格的规则提取、概念知识空间、概念格的粒计算方法、概念格应用等[2]，已逐步发展成一门成熟的知识发现体系，并演变出很多不同的研究范式，备受国内外计算机科学、数学、医学、情报学等领域专家学者关注。

2 概念格基础

2.1 形式背景和概念格

定义 1 设 U 是对象的集合，M 是属性的集合，I 是 U 与 M 间的关系，则称三元组 K=(U,M,I)为一个形式背景。(u.m)∈I 表示对象 u 具有属性 m。背景可以用一个对象集合、属性集合以及他们之间的二元关系建立起来的表格来表示，它的每行表示某一对象，每列则表示某一属性。

定义 2 设 K=(U,M,I)为形式背景，对象集合 U={u1,u2,..,um}，属性集合 M={m1,m2,..,mn}。

设 m∈M，具有属性 m 的对象集合。定义

$$g(m)=\{u\in U/(u,m)\in I\}$$

设 u∈U，具有属性 m 的对象集合。定义

$$f(u)=\{m\in M/(u,m)\in I\}$$

定义 3 设 K=(U,M,I)为形式背景，若 A⊆U，B⊆M，那么存在 $f(A)=\{m\in M/\forall u\in A,(u,m)\in I\}$ 和 $g(B)=\{u\in U/\forall m\in B,(u,m)\in I\}$

如果集合 A,B 满足 f(A)=B,g(B)=A。则称由(A,B)二元组被称为概念。A 是概念(A,B)的外延，B 是概念(A,B)的内涵。

定义 4 设 K=(U,M,I)为形式背景，存在两个概念(A1,B1)和(A2,B2)满足

$$(A_1,B_1)\le(A_2,B_2)\Leftrightarrow A_1\subseteq A_2(\Leftrightarrow B_1\supseteq B_2)$$

则称(A1,B1)为(A2,B2)的子概念，(A2,B2)为(A1,B1)的超概念。由形式背景(U,M,I)中所有概念根据它们之间的层次关系有序组成的集合，常称为(U,M,I)的概念格。

2.2 概念格基本原理

概念格将每一个节点表示为一个形式概念, 每个形式概念包含概念的外延 (extent) 和内涵 (intent) 两部分内容。外延表示此概念所包含的所有对象的集合, 即此概念所涵盖的实例, 内涵则表示概念中所有对象的共有特征。对于给定的形式背景 K= (G, M, I) (其中 G 为对象集合, M 为属性集合, I 是 G 与 M 之间的一个二元关系) , 存在惟一一个偏序集合与之相对应。由偏序集构成一种格结构, 并且此偏序集满足自反性、反对称性和传递性。若 g∈G, m∈M, g Im 表示对象 g 具有 m 属性。格中的每个节点称之为概念, 记作 C (X, Y) , X∈G 是概念 C (X, Y) 的外延, Y∈M 是概念中对象的共有属性(内涵) 。节点概念与节点概念之间存在着偏序关系, 若有概念 C1= (X1, Y1) , C2= (X2, Y2) , 并且 X1<X2<=>Y1<Y2, 称 C1 为 C2 的父节点。概念格的形式背景通常是由如表 1 所示的二维数表来表示, 横向维表示属性, 纵向维表示对象, 第 i 行 j 列的数为 1 表示存在该属性, 为 0 表示不存在该属性。由此可知该形式背景中包含五个对象, 四个属性, 可形成八个概念, 与其相对应的外延与内涵如表 2 所示。由概念间的继承关系可得出如图 1 所示的 Hasse 图。

G	A	B	C	D
1	0	1	1	0
2	1	0	0	1
3	1	0	1	1
4	0	1	1	0
5	1	0	0	0

表 2.2.1 形式背景示例

编号	外延	内涵
0	{1,2,3,4,5}	{}
1	{3,5}	A
2	{3,4}	B
3	{1,2,4}	C
4	{3}	A,B
5	{4}	B,C
6	{1,2}	C,D
7	{}	A,B,C,D

表 2.2.2：所生成的概念

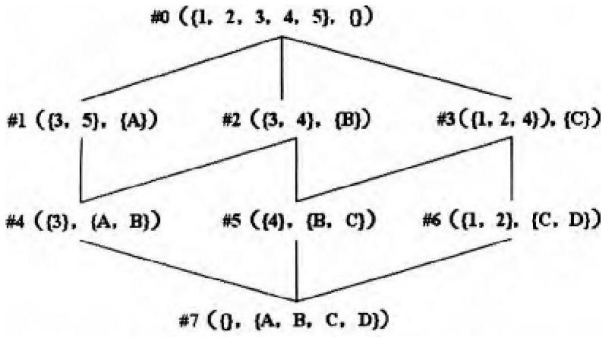

图 2.2.1 与表 1 相对的 Hasse 图

3 概念格的主要研究内容

3.1 概念格构造

概念格的结构通常具备完备性，构造效率呈现指数级发展，国内外很多学者专家 Bordat、

Nourine、刘宗田、胡可云等从提高概念格构造效率的方向入手，对概念格构造方法和规则算法进行研究，优化和改进了一些算法[3]。目前概念格构造的主流方法已基本形成，分别是批处理算法和增量算法。这些方法也是构建扩展概念格的主要思想，后续其它概念格构造算法均能找到上述主流方法的影子，如批处理 Choi，渐增式 AddIntent，In-Close 等。

3.1.1 批处理算法

批处理算法根据构造的不同方式，批处理算法可以大致分为三类:自顶向下算法、自底向上算法和枚举算法。自顶向下算法是指首先构造格的最顶层结点，再依次往下构造，这种算法直观、明了、简洁、便于实现并行化，但也存在一些不足，格的子结点在生成过程中会产生很多重复结点。如 Bordat 算法、Choi 算法和 OSHAM 算法等;自底向上算法是首先构造格的最底层结点，再逐渐向上扩张，如 Chein 算法、Norris 算法、MAC 算法和 Cbo 算法等;枚举算法是指按照一定顺序先枚举出格的所有结点，然后在形成格结点的关系，即 Hasse 图。如 Ganter 算法[4]等。

3.1.2 增量算法

增量算法，又称为渐进式算法。增量算法的主要思想是将待插入的对象与格内已存在的概念节点进行交运算，根据结果的不同使用相应的处理办法。目前较为流行的增量算法有 Capineto 算法、T.B.HO 算法、Godin 算法等[5]。Godin 对增量算法进行了改进，在新对象插入时，用遍历所有的节点，仅仅检查是否至少和新对象有一个共同属性的节点。该操作通过维护一个可包含每个属性首次在格内出现的指针来实现，该指针能自顶而下进行深度优先搜索。渐进式生成概念格的求解过程中，要着重解决三类问题:如何生成新节点、如何避免重复节点的产生和如何更新连接节点的边。

3.2 概念格约简

概念格的约简能够有效地提高概念格的维护效率，使形式背景中所蕴含的知识易于发现，简化知识的表示方式。约简概念格实际上是在保持对象集不变的条件下，如何求得最小的属性集的过程。概念格约简有三种方式:对象约简、属性约简和概念约简. 相对而言对象约简受关注较少，下面主要介绍属性约简和概念约简的研究进展. 实际上,属性约简是一个保持形式背景或概念格某种特性不发生改变以计算属性全集的极小子集的过程，而概念约简是保持某种需求得到满足的情况下求解所有概念的最小子集的过程。

3.2.1 属性约简

属性约简是概念格研究领域中的一个热点问题,张文修等提出的概念格属性约简理论与方法是该研究方向开创性工作。概念格属性约简是在形式背景的对象集不发生改变时计算极小属性子集,使约简后的形式背景生成的概念格与原始数据生成的概念格同构。属性约简的一般性描述可概括如下:对于形式背景(U,A,I),设ϕ(U,A,I)为客户要求的信息刻画,若存在 E⊆A,使ϕ(U,A,I)=ϕ(U,E,IE),则称 E 为(U,A,I)保持ϕ信息的协调集. 进一步,如果∀a∈E, ϕ(U,A,I)≠ϕ(U,E-{a},IE-{a}),称 E 为(U,A,I)保持ϕ信息的约简。

3.2.2 概念约简

实际上,相对对象约简和属性约简,概念约简则是综合考虑如何避免对象与属性的冗余问题。曹丽等在保持二元关系不变的前提下建立概念约简理论. 类似于经典概念格属性约简,讨论如下问题:1) 提出概念协调集和约简,给出相应的判断性质;2)依据功能和作用的不同,将全体概念区分为 3 类(核心概念、相对必要概念、绝对不必要概念),并从二元关系的角度对它们进行了刻画。具体地,记 ∂(U,A,I)为形式背景(U,A,I)的概念格,Ω⊆∂(U,A,I),若 I=∪(X,B)∈ΩX×B, 则称Ω为保持二元关系不变的概念协调集. 进一步,如果对∀(X,B)∈Ω, Ω′=Ω−{(X,B)},I≠∪(X,B)∈Ω′X×B, 称Ω为保持二元关系不变的概念约简. 进一步,魏玲等[2]基于概念辨识矩阵,给出计算所有概念约简的方法. 谢小贤等通过关系矩阵运算生成所有概念约简.需要指出的是,除了保持二元关系不变的概念约简之外,还有其它方式研究概念约简。

3.3 基于概念格的规则提取

概念格上的规则提取具有广泛的应用前景。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。由于概念格结点反映了概念内涵和外延的统一，结点间关系体现了概念之间的泛化和例化关系，因此非常适合作为规则发现的基础性数据结构，基于概念格的分类规则的提取在知识发现等方面有着广泛的应用。规则挖掘是近年来数据挖掘的研究

课题，每个概念格节点本质上就是一个最大项目集，为关联规则挖掘提供了平台，体现了概念之间的包含与分类关系，更加易于理解和表示。根据数据挖掘任务的不同（如蕴涵规则、关联规则、分类规则、聚类分析、序列模式、时序摸索、决策规则等），国内研究人员做了大量研究，并且对概念格结构做了不同程度的扩展以适应规则挖掘的要求。提取分类规则的模型又很多种（判断树、贝叶斯网、神经网络、概念格和粗糙集等），概念格模型只是其中的一种。目前，对于概念格上分类规则的研究主要集中在优化概念格的构建和求解算法上。

3.4 概念格与其他学科结合产生的方法

为了扩大和充实概念格理论,也为了该理论的广泛应用,必须将该理论与其他学科相结合.由目前研究成果得知,概念格与其他学科的结合较为广泛,其中与粗糙集和模糊集的结合成果相对丰富.

3.4.1 与模糊集的结合

由于各个应用领域中存在的信息具有复杂性和不确定性，在处理以上问题时，传统的形式概念分析很难表述不确定的信息。为了找到表述模糊信息的表达方式，人们将模糊理论与形式概念分析结合起来，由此产生了模糊形式概念分析。模糊形式概念分析是建立在模糊形式背景上的，模糊形式背景可用一个三元组 K=（G，M，I=Φ（G×M）来表示，其中 G 为所有对象集合，M 为所有属性集合，I 是在域 G×M 上定义的模糊关系。每个概念（g m）均包含一个隶属度 μ，取代普通概念格中的二元赋值。隶属度的引入是为了能够进行概念间语义相似度的计算，也为了能够更加简单的表述模糊概念格。一般形式概念使用概念的属性来描述，所以任一对象与概念之间的关系即是该对象与概念内各个属性之间的关系，并且取属性间关系的交集。由于在 Zadeh[6]的模糊集合论中，隶属度的数值取该对象与该对象概念中各属性间隶属度最小的值，隶属度 μ 的取值在闭区间[0，1]上。若对象 g 与属性 m 之间的 μ=1，则表示 m 完全属于 g，即该对对象与属性具有完全确定的关系，称 g 为由（g，m）所形成的概念的外延，m 为该概念的内涵，这与一般概念格保持一致;若 μ=0，则 g 与 m 完全没有关系。利用隶属度函数可以方便地计算任意概念与其超概或是子概念之间的语义相似度。

3.4.2 与粗糙集的结合

粗糙集是发现知识、挖掘知识的数学工具,能有效地分析和处理不精确、不完备的信息.粗糙集利用等价关系对数据表进行分类[3],而概念格是基于相同数据表,并结合序理论对概念分层加以讨论. Xu 等将变精度粗糙集 β-上、下分布约简算法的优势与概念格形式背景相结合,提出了基于变精度粗糙集的概念格约减算法,同时分析了变精度粗糙集模型中的 β 值的选取算法、可辨识矩阵属性约简,以及传统算法中存在的问题,并且对传统算法进行了改进.Mao 等通过利用变精度粗糙集中的 β-上、下近似与概念格中概念相结合，提出了概念格上的变精度粗糙集近似算子,并根据这一近似运算对形式背景中任意不可定义的对象集进行近似,求出与其最接近的概念的外延,并得到了上、下近似概念。

4 概念格存在的问题与展望

4.4.1 概念格约简存在的问题

概念约简能保持对概念信息的需求不变,涉及更少的概念节点,理论意义非常明确.作为一个有前景的研究方向,仍需进一步探讨制约其发展的关键问题,例如保持二元关系不变的概念约简:一个形式背景对应的概念约简通常是不唯一的,那么如何快速获取概念个数最少的概念约简值得考虑.其次,概念约简一般不再是一个格(而是偏序集),那么是否存在某个偏序集能形成格结构? 再者,既然概念约简并不改变原始的形式背景,而概念约简又只是全体概念的一部分,那么是否存在完备的概念约简(按照某种方式能生成剩余的概念)也是一个重要课题. 决策形式背景上的概念约简如何实施及其语义如何解释,这些都是当前面临的困难问题.

4.4.2 概念格应用存在的问题

第一，基于概念格的本体研究：该方向已取得一些初步的研究成果,可适应时间序列模式和多维空间结构融合的本体研究,但这还远远不够,如在多粒度多维度动态数据等复杂环境下如何实现本体快速构建与有效融合仍需进一步研究。第二，基于概念格的认知计算.概念格自身的形成过程在某种程度上体现知识认知的规律,也较适合模拟人脑的认知过程[7].但是,人脑的认知过程不是简单机械地重复一个过程,它还包含一定的不确定性.为了更好地模拟人脑的认知过程,需要将认知的修正功能及认知遗忘和认知重现等复杂因素考虑在

内,这是一个有挑战性的研究课题。

总之，传统的研究方向已较成熟,像概念格模型推广、概念格构造、概念格约简和基于概念格的规则提取,这些研究方向已逐渐开始转入深层次问题的探讨。新兴的研究方向才刚刚起步,像概念知识空间、概念格的粒计算方法和概念格应用中的认知计算等研究工作,目前尚处于理论和方法的完善阶段,有些问题甚至都还没有明确的定论。现阶段需要突破本文指出的各类关键问题，这些方面一旦取得实质性的突破,将会极大促进人工智能、认知科学等领域的进一步发展。

结束语

概念格理论经过几十年的发展,取得一系列成果，本文从概念格理论的定义、研究内容、发展方向等方面总结了概念格的研究进展，希望能使更多读者关注和了解概念格理论，共同努力推进概念格理论与技术科学发展。另外，概念格理论与方法的研究不再只局限于形式概念分析的理论框架下进行探讨,它发展到今天已经与其它理论深度交叉融合,如粗糙集、模糊集、粒计算、三支决策及认知计算等,从而可更好地解决机器学习、人工智能、认知科学等领域中存在的关键科学问题。通过近几年来概念格在国内外发展的现状统计和分析表明,概念格以其独有的特性不断地赢得众多学者的关注,其应用范围也不断地拓展.从概念格理论研究历史中得出,将其他学科与概念格理论相结合会创造出新思想、新方法。

参 考 文 献

[1]WILLE R.Restructuring lattice theory:Anapproach based on hier-archies of concepts[C].Ordered Sets.Dordrecht:Reidel,1982

[2]李金海,魏玲,张卓,翟岩慧,张涛,智慧来,米允龙.概念格理论与方法及其研究展望[J].模式识别与人工智能,2020,33(07):619-642.

[3]降惠.概念格理论研究进展与发展综述[J].办公自动化,2019,24(09):18-21.

[4] Ganter B, Stumme G, Wille R. Formal Concept Analysis: Theory and applications - J. UCS Special Issue. Journal of Universal Computer Science, 2004, 10(8): 926-926.

[5]邓硕. 多粒度标记粗糙集与概念格的规则比较[D].昆明理工大学,2019.

[6]Zadeh L A.Fuzzy logic and approximate reasoning[J].Syntheses, 1975, 4 (4) :203～216

[7]李金海,闫梦宇,徐伟华,折延宏,张文修.概念认知学习的若干问题与思考[J].西北大学学报(自然科学版),2020,50(04):501-515.