《智能信息处理》课程作业

形式概念分析及改进算法分析研究

李征蔚

作业	分数[20]
得分	

2021 年 11 月 28 日

形式概念分析及改进算法分析研究

李征蔚

（大连海事大学　信息科学与技术学院　大连　116026）

摘　要　人类在认知过程中，把所感觉到的具有共同特点的事物抽取出了，加以概括，称为概念。又概念被理解为由外延和内涵两个部分所组成的思想单元。基于概念的这一哲学思想，德国的 R. Wille 教授首先提出了形式概念分析理论。形式概念分析理论是对领域本体进行抽象描述，并且以概念的形式呈现。经过多年的研究探索，该理论已经在概念格构造、属性约简、概念约简、规则获取、粒计算等方向获得了大量的研究成果，广泛应用于信息检索、数据挖掘、知识发现、人工智能等诸多领域。

关键词 形式概念分析, 领域本体, 概念格

Formal Concept Analysis and Improved Algorithm Analysis Research

Li Zhengwi

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

Abstract　In the process of cognition, human beings extract and generalize the things they feel with common characteristics, which are called concepts. The concept is also understood as a unit of thought composed of two parts: denotation and connotation. Based on this philosophical idea of concept, Professor R. Wille of Germany first proposed the theory of formal concept analysis. Formal concept analysis theory is an abstract description of the domain ontology and presents it in the form of concepts. After years of research and exploration, the theory has obtained a large number of research results in concept lattice construction, attribute reduction, concept reduction, rule acquisition, granular computing, etc., and is widely used in information retrieval, data mining, knowledge discovery, and artificial intelligence. And many other fields.

Keywords Formal concept analysis, Domain ontology, Concept lattice

1 引 言

在形式概念分析中，概念的外延是指涵盖所有对象的集合，内涵是指由所有对象共有属性组成的集合。形式概念分析是数据的内在结构、关联和相互依存关系的数据分析方法，通过内涵与外延两方面对概念进行数学化表达，可发现和推理数据中的概念以及发现数据中的依赖关系，已经被广泛运用于多个领域。 其并不像其他数据分析方法那样大大减少了给定的信息，只能获得极少的“重大参数”[1]；相反，通过数据集中对象和共同属性之间的二元关系，运用简单明了的图形反映复杂的网络关系，能最大程度保留数据之间的各类特殊关系。因此，利用形式概念分析法，从概念的整体角度描述云化形式背景下的数据多层次结构，通过概念格全面地反映概念的聚类和关联特性，使其直观化、可视化，从而能有效挖掘云化形式背景下的数据所蕴含的潜在信息

在当今大数据时代，知识获取变得愈加复杂。众所周知，概念格的构造本质上是一个 NP-hard 问题。近年来，格理论在形式概念分析框架下得到了广泛的应用。形式概念分析的主要特点是形式概念的发现，在发现的概念中推导出层次的偏序关系，并将其可视

化为概念格。概念格作为一种重要的知识发现的方法，可以用于构造满足约束的角色层次结构。该理论从形式背景的概念开始，指定哪些对象具有哪些属性。形式概念分析被视为一种有效的数据分析方法，被广泛的使用在决策、信息检索、数据挖掘和知识发现等领域。

2 形式背景

形式概念分析首先要建立形式背景。形式背景被定义为一个三元组，形式背景是一个集合，结构为 K：=(O,A,R),在这里，O 和 A 是集合，而 R 是在 O 和 A 之间的一个二元关系（即：R⊆O×A)。元素 o 和 a 分别叫做（形式化）对象利（形式化）属性。

实际上，形式背景一般都不是直接存在的，需要从数据源中提取，从而就需要对数据源进行分析，采取不同的策略和算法来提取形式背景。

表 1 形式背景的示例

	a	b	c	d	e
1	1	1	1	1	0
2	1	1	1	0	0
3	0	1	1	0	0
4	0	1	0	0	1

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。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 2 单值属性形式背景表的示例

	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

3 概念格

建格的过程实际上是概念类聚的过程[1]。因此，在概念格中，建格算法具有很重要的地位对于同一批数据，所生成的格是唯一的，即不受数据或属性排列次序的影响，这也是概念格的优点之一。概念格的建格算法可以分为两类：批处理算法和增量算法。概念格可以添加背景知识，这些知识以 if…then 的规则形式出现。概念格甚至可以只用背景知识建造。

批处理算法根据其构造格的不同方式，可分为 3 类，即从顶向下算法、自底而上算法、枚举算法。从顶向下算法首先构造格的最上层节点，再逐渐往下。自底而上算法则相反，首先构造底部的节点，再向上扩展。枚举算法则是按照一定顺序枚举格的所有节点，然后再生成 Hasse 图，即各节点之间的关系。增量算法和批处理算法不同，增量算法的思想都是大同小异的——基本思想都是将当前要插入的对象和格中所有的概念交，根据交的结果采取不同的行动[2]。主要区别在于连接边的方法。

概念格是对概念进行形式化表达的方式，通过将数据集转化为形式背景，挖掘数据的规则关系，并通过Hasse图实现可视化。其相关定义及性质如下[12,13]：定义 1 三元组(O, A,R) ，其中 O 表示对象集，A 表示属性集， R 表示两者之间的关系，即 R ⊆ O ×A，则三元组 (O, A, R) 称为形式背景。定义 2 形式背景 (O, A,R) 所有概念的集合所构成的完备格，称为概念格 B(O, A,R) 。定义 3 若形式背景 (O, A,R) 中的概念可用序偶(M, N) 表示， M ⊆O ， N ⊆A ，则称 M 为概念(M, N) 的内涵， N 为概念 (M, N) 的外延。满足如下关系：

M =α (N)={m ∈O| ∀n ∈A, m Rn}(3) N =β (M) ={n ∈A| ∀n ∈O, m Rn}[2] 。

在实际操作中，大多属性具有多种值，例如，距离可分为远、近，则这种形式背景为多值属性形式背景。对于多值形式背景，概念格有两种处理方式：一种是将多值形式背景转化为单值形式背景，这种方式增加了概念的节点数，但易于理解和提取规则；另一种方式是直接在多值形式背景下，依据底

层概念关系提取高层的概念。概念格可以从数据集中提取关联规则、分类规则和特征规则： ① 关联规则：与 Apriori 算法原理类似，通过挖掘数据之间的关系，消除冗余，表示为 A ⇒B 。 ② 分类规则：规则一般为左右件分离，左件为属性的特征，右件为类别标号。③特征规则：将属性的特征作为规则的右件，左件作为特征的对象。概念格扫描数据集获取概念，实质上是概念聚类的过程。当数据集对象或属性数量很大时，概念格不具备实际应用性，因此，需要对概念格进行化简[6]。在实际应用中可通过属性约简，设定兴趣属性等方式降低算法复杂度，也可根据具体情况和所转化的形式背景，选择代表性数据提取无冗余的规则集，建立部分格可代表全局数据的规则[3]。

以下，本文根据亲属关系的简单形式背景产生对应概念格。首先，根据所给形式背景约减生成单值形式背景，再确定单值形式背景中的父子关系，根据父子继承关系绘制 Hasse 图，最后补充各形式概念的上确界和下确界，形成概念格。

3.1 约简形式背景

形式背景的约减包括聚类（行约减）和关联（列约减）。通过表 2 可看出，1、4 与是一组有相同属性的行，故将其合并。而表中并没有相同的列，故不进行列约减。最后得到约减后的形式背景如表 3。

表 3 亲属关系-约减形式背景

	a	b	c	d	e
1,4	1	1	1	1	0
2	1	1	1	0	0
3	0	1	1	0	0
5	0	1	0	1	0
6	0	0	1	1	0
7	0	0	1	0	1
8	0	1	0	0	1

3.2 生成单值形式背景

单值的形式背景即根据前一步约减后的形式背景，把值为“1”的位置改为“×”，去掉其他位置的“0”以表示该形式对象有此属性[7]。最后得出结果见表 4。

表 4 亲属关系-单值形式背景

	a	b	c	d	e
1,4	x	x	x	x	
2	x	x	x		
3		x	x		
5		x		x	
6			x	x	
7			x		x
8		x			x

3.3 确定父子关系

父子关系也称基于属性个数的排序。在获取到的单值形式背景的基础上做顺序的调整，找到属性继承的父子关系，例如 2 可由对 3 的全部属性继承的基础上添加自身属性 a 得到。通常情况下，为方便查找，从上倒下按属性的多少进行排列。表 5 所示为基于属性个数的排序。

表 5 亲属关系-基于属性个数的排序

	a	b	c	d	e
8		x			x
5		x		x	
6			x	x	
7			x		x
3		x	x		
2	x	x	x		
1,4	x	x	x	x	

3.4 绘制 Hasse 图

Hasse 图也称哈斯图[8]，在数学分支序理论中，是用来表示有限偏序集的一种数学图表，它是一种图形形式的对偏序集的传递简约[4]。

具体的说，对于偏序集合（S,≤），把 S 的每个元素表示为平面上的顶点，并绘制从 x 到 y 向上的线段或弧线，只要 y 覆盖 x（就是说，只要 x < y 并且没有 z 使得 x < z < y）。这些弧线可以相互交叉但不能触及任

何非其端点的顶点。带有标注的顶点的这种图唯一确定这个集合的偏序。

Hasse 图的作图法为：以“圆”表示元素；若 x<y，则 y 在 x 的上层；若 y 覆盖 x，则连线；不可比的元素在同层。应用 Hasse 图表示各结点所组成的偏序集及节点间的关系，由上到下表示的即为两节点间的父子关系，根据表 5 所绘 Hasse 图如图 1 所示。

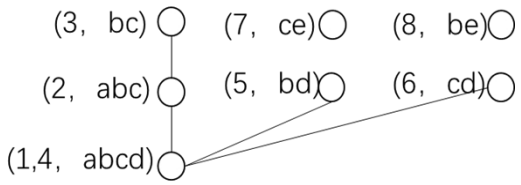

图 1　Hasse 图

3.5 生成概念格

针对表 5 的简单形式背景，采用手工方式生成概念格。图 1 已经给出 Hasse 图，即已得出概念间的偏序关系，只需补出上下确界即可得到概念格[5]。图 2 是产生的概念格。

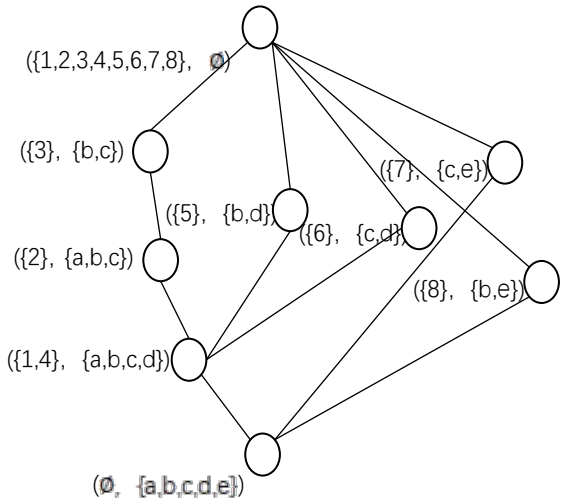

图 2　概念格

3.6 基于概念格改进云推理分类预测算法

通过云模型对数据的双向转化，建立具有不确定性的形式背景，通过构造概念格实现对分类规则的提取，创建规则集，取代了传统云推理模型中的“软与”算法，减少规则数量，更适用于多维度数据。 设数据集中有 n 个条件属性，则传统概念格产生的概念节点数最大值为 2n.设数据集中的数据量为 m，则传统云推理模型需要计算至少 m 个激活强度[8]，大部分数据不只激活一条规则。本文的改进主要是通过概念格对云化形式背景下的数据进行规则提取，无需计算软与程度调节参数，且云模型对概念的模糊化，使概念格算法复杂度可降低为 O(n) 。

4 结 论

在使用形式概念分析过程中，我们可以利用已知的知识或者常识，经过形式化以后得到新的知识和常识，从而帮助我们更好的理解研究的系统发现潜在的知识。

本文通过使用概念格及相关知识，对云推理模型网络进行了改进，得到以下结论。结论一：通过云形式背景将条件属性的确定度转化为区间形式，并对应不同数据集等级，可以保证形式背景的不确定性与算法的普适性。 结论二：取消云推理过程中的“软与”过程，将规则中的云概念语言集所对应的确定度区间范围作为规则后件，并通过 Y 条件云发生器进行输出，从而降低了推理过程中的规则数量以及人为因素的影响。

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] 智慧来,李逸楠.形式概念分析中的面向对象概念约简[J].海南热带海洋学院学报,2021,28(05):66-71.DOI:10.13307/j.issn.2096-3122.2021.05.09.

[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] 徐立,白金牛,孟海东.基于概念格的云推理分类预测算法研究[J].控制工程,2020,27(11):1892-1900.DOI:10.14107/j.cnki.kzgc.20200475.