Journal of Computer Applications
计算机应用，2015，3(4)：193-196
ISSN 1001. 9081
CODEN JYIIDU
2015. 11. 20
http://www.joca.ca

文章编号：1001-9081(2015)11-193-04
doi：10.11772 / j.issn.1001-9081.2015.11.193

基于形式概念分析的知识发现方法研究

朱雪

（大连海事大学 信息科学技术学院，大连 116026）
（ 通信作者电子邮箱 1874110235@qq. com）

摘 要：形式概念分析是由德国的 Wille 教授于 20 世纪 80 年代初提出的，它反映了概念的哲学理解，其核心数据结构概念格，也称 Galois 格，准确而简洁地描述了概念之间的层次关系，已成为一种重要的知识表示方法。随着研究的深入，形式概念分析越来越多地被应用到数据挖掘、信息检索、软件工程等领域，成为组织和处理大规模数据的有效工具。基于以上背景，本文主要从以下五个方面展开研究：基于搜索空间划分的概念生成、基于概念格的用户关联挖掘 、基于概念格的分类器研究、GDT 与扩展概念格模型、 基于概念格的数据挖掘系统原型。

关键词：形式概念分析；概念格；分类器；数据挖掘系统

中图分类号：P391.4　　**文献标志码：**A

Knowledge Discovery Methods Research Based on Formal Concept Analysis

ZHU Xue
(Institute of Computer 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, reflects the comprehension for concept in philosophy. 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 knowledge. With the development of FCA, it has been used widely in the machine learning, data mining and knowledge discovery, information retrieval, software engineering and become one of the most efficient tools to deal with large scale of data. Under this context this paper makes the researches on the bellowin g aspect: concept generation based on search space partition、user association mining based on concept lattice、research on classifier based on concept lattice、GDT and extended concept lattice model、prototype of data mining system based on concept lattice.

Key words： Formal Concept Analysis;Concept Lattice; Classifier; Data Mining System

1 绪论

在哲学中，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理解，德国的 Wille 教授提出了形式概念分析，用于概念的发现、排序和显示[1]。在形式概念分析中，概念的外延被理解为属于这个概念的所有对象的集合，而内涵则被认为是所有这些对象所共有的特征或属性集，这实现了对概念的哲学理解的形式化。所有的概念连同它们之间的泛化例化关系构成一个概念格。概念格结构模型是形式概念分析理论中的核心数据结构，它本质上描述了对象和特征之间的联系，表明了概念之间的泛化与例化关系，其相应的 Hasse 图[2]则实现了对数据的可视化。下面介绍一下形式概念分析的几个基本概念：

[定义 1.1] 形式背景定义为一个三元组 K=(G,M,I)，其中，G 是对象集合，M 是属性集合，I⊆G × M 是 G 与 M 之间的一个二元关系。若(g, m)∈I，读作“对象 g 具有属性 m”。形式背景通常用交叉表（cross table）来表示。表 1.1 是一个例子形式背景的交叉表。

收稿日期：2016-10—30；修回日期：2016—11—12。

基金项目：国家自然科学基金资助项目(61373099)；国家青年科学基金资助项目(61402316)。

作者简介：朱雪(1994 一)，女，山东滨州人，硕士，主要研究方向：概念格、数据挖掘。

[定义 1.2] 形式背景 K= (G,M, I)上的一个形式概念（简称概念）定义为一个二元组(A,B)，满足：A⊆G, B⊆M, A’ = B, B’ = A 其中，A 称为概念(A,B)的外延，B 称为概念(A,B)的内涵。

[定义 1.3] (A,B)和(C, D)是形式背景 K = (G, M, I)上的任何两个概念，称(A, B)是(C, D)的超概念（等价地，(C, D)为(A, B)的子概念），当且仅当 B ⊆ D（等价地，C ⊆ A），记为 (C, D) ≤ (A, B)。即：(C, D) ≤(A, B) ⇔ B ⊆ D (⇔ C ⊆ A) 通过这种序关系，得到一个有序集 B(K) = (B(K), ≤)，称为形式背景 K 的概念格[3]。

表 1.1 一个例子形式背景

		M								
	I	a	b	c	d	e	f	g	h	i
	1	1	0	1	0	0	1	0	1	0
	2	1	0	1	0	0	0	1	0	1
G	3	1	0	0	1	0	0	1	0	1
	4	0	1	1	0	0	1	0	1	0
	5	0	1	0	0	1	0	1	0	0

2 基于概念格的用户关联挖掘

从评价表挖掘用户关联是基于规则的推荐系统的核心任务。传统的基于 Apriori 算法[4]的用户关联挖掘算法通常产生大量的与当前推荐无关的规则。本章将基于概念格的关联规则挖掘思想用于用户关联挖掘，定义了两个用户关联基：确定的用户关联基（规则的可信度=100％）和近似用户关联基（规则的可信度<100％）。从这两个用户关联基可以导出所有的有效用户关联规则。

2．1 问题描述

推荐系统收集所有用户对资源的评价信息形成评价表，评价表一般用二维表来表示，其中第 i 行第 j 列表示第 j 个用户对第 i 个资源的评价，取值为“1”表示喜欢，“0”为未作评价。表 2.1 是 6 个用户对 10 种资源的评价。

显然可以将评价表看作一个形式背景，得到其对应的概念格，这样，用户关联挖掘问题演化为在概念格上提取以当前用户为后件的关联规则。

表 2.1 例子评价表

	U_1	U_2	U_3	U_4	U_5	U_6
I_1	1	1	0	0	1	0
I_2	1	1	0	0	1	0
I_3	1	0	0	1	1	0
I_4	1	1	1	1	1	0
I_5	0	1	0	0	0	1
I_6	1	0	1	1	1	0
I_7	0	0	1	1	0	1
I_8	1	0	1	1	0	1
I_9	0	0	1	0	0	1
I_{10}	1	0	1	1	0	0

2．2 挖掘用户联基

概念格中每个概念的内涵是一个封闭项集，概念的外延的基数是该封闭项集的支持度。如果封闭项集的支持度大于支持度门限，则为频繁封闭项集。由于频繁项集的支持度等于它的闭包的支持度，并且最大频繁项集是最大频繁封闭项集，使用频繁封闭项集挖掘关联规则不但可以大量减少冗余规则的数量，而且不会丢失任何信息。正式由于上述原因，概念格成为关联规则挖掘的一个很自然的平台。

[定义 2.1]确定的用户关联基为：EB={r：g⇒Uc| Uc∈f∧g∈Gf∧Uc∉g}。其中，f 为包含当前用户的频繁封闭项集，Gf 表示 f 的缩减的集合。

[定义 2.2] 近似的用户关联基为： AB = {r：g→Uc|Uc∈f∧g∈G∧γ (g)⊂f}。其中，f 为包含当前用户的频繁封闭项集， G 表示所有频繁封闭项集的缩减的集合。

2．3 算法 User_Association

给出确定和近似的用户联基 EB、AB。首先根据当前用户 Uc 及支持度门限 min_sup，构造包含当前

用户节点集合 G+和边界节点集合 Gb，然后调用过程 Gen_EB 和 Gen_AB 生成确定的用户关联基 EB 和近似的用户关联基 AB。过程 Gen_EB 考察 G+中的每一个节点 H，首先计算内涵 Hf 的缩减集 Gf，然后根据定义 2.1 对每个缩减 g 生成相应的确定规则(，规则的可信度为 1，支持度为节点 H 的外延的基数 Hc 与对象集的基数|G|的比。过程 Gen_AB 考察 Gb 中的每一个节点 H，利用一个堆栈搜索它的所以前驱节点 P，如果 P 不属于 G+、未访问过、不是根节点且满足可信度门限 min_conf 则计算内涵 Pf 的缩减集 Gf，然后根据定义 2.2 个缩减 g 生成相应的近似规则，规则的可信度为节点 H 的外延的基数 Hc 与节点 P 的外延的基数 Pc 的比，支持度为节点 H 的外延的基数 Hc 与对象集的基数|G|的比[5]。

表 2.2 表 2.1 例子评价表的确定的用户关联基

G_+中的频繁节点 H	H 中的封闭项集 f	f 的缩减 g	确定规则 $r: g \Rightarrow U_c$	$Sup(r)$
$(1,\{U_1\})$	$\{U_1\}$	$\{U_1\}$	G 中包含 U_c	
$(5,\{U_1,U_4\})$	$\{U_1,U_4\}$	$\{U_1,U_4\}$	G 中包含 U_c	
$(5,\{U_1,U_5\})$	$\{U_1,U_5\}$	$\{U_5\}$	$U_5 \Rightarrow U_1$	5/10
$(3,\{U_1,U_4,U_5\})$	$\{U_1,U_4,U_5\}$	$\{U_4,U_5\}$	$U_4 \wedge U_5 \Rightarrow U_1$	3/10
$(4,\{U_1,U_3,U_4\})$	$\{U_1,U_3,U_4\}$	$\{U_1,U_3\}$	G 中包含 U_c	
$(3,\{U_1,U_2,U_5\})$	$\{U_1,U_2,U_5\}$	$\{U_1,U_2\}$	G 中包含 U_c	
		$\{U_2,U_5\}$	$U_2 \wedge U_5 \Rightarrow U_1$	3/10
$(2,\{U_1,U_3,U_4,U_5\})$	$\{U_1,U_3,U_4,U_5\}$	$\{U_3,U_5\}$	$U_3 \wedge U_5 \Rightarrow U_1$	2/10

表 2.3 表 2.1 例子评价表的近似的用户的关联基

G_b中的频繁节点 H	H 的前驱节点 P	P 中的封闭项集 f	F 的缩减 g	近似规则 $r: g \rightarrow U_c$
$(7,\{U_1\})$	$(10,\varnothing)$	f 为$\varnothing$		
$(5,\{U_1,U_4\})$	$(6,\{U_4\})$	$\{U_4\}$	$\{U_4\}$	$U_4 \rightarrow U_1$
	$(10,\varnothing)$	f 为$\varnothing$		
$(4,\{U_1,U_3,U_4\})$	$(5,\{U_3,U_4\})$	$\{U_3,U_4\}$	$\{U_3,U_4\}$	$U_3 \wedge U_4 \rightarrow U_1$
	$(6,\{U_4\})$	已处理过		
	$(6,\{U_3\})$	$\{U_3\}$	$\{U_3\}$	$U_3 \rightarrow U_1$
	$(10,\varnothing)$	f 为$\varnothing$		
$(3,\{U_1,U_2,U_5\})$	$(4,\{U_2\})$	$\{U_2\}$	$\{U_2\}$	$U_2 \rightarrow U_1$
	$(10,\varnothing)$	f 为$\varnothing$		

3 GDT 与扩展概念格模型

GDT 与扩展概念格模型泛化分配表（GDT）来是一种用来表示概念和实例间概率关系的表。本章将 GDT 作为假说搜索空间，提出了一种从不一致和不完全的数据中提取规则的算法；并给出了一种基于 GDT 的扩展概念格模型和相应的规则提取算法。

3．1 基于 GDT 的规则发现算法

我们将 GDT 作为搜索空间，搜索所有泛化，试图从中发现能覆盖所有实例且所含泛化数目最小的规则集。搜索策略用来决定考察泛化的顺序。我们将所有的泛化根据泛化中“*”的数目分为若干级别，“*”的数目越多，级别越高。我们的搜索策略优先考虑那些级别高的，即更一般的泛化。很明显不同级别的泛化之间存在蕴含关系。例如实例 a0b1c1 所对应的泛化为 a0、b1、c1、a0b1、a0c1 和 b1c1，它们之间的关系如图 3.1 所示。

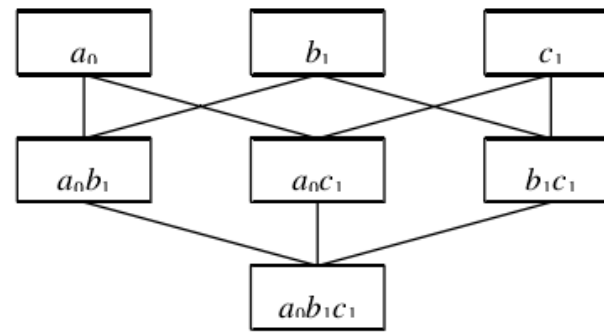

图 3.1 泛化之间的关系

3．2 基于 GDT 的规则发现算法

可见上层泛化蕴含下层与之有关的所有泛化。即 a0 蕴含 a0b1 和 a0c1，b1 蕴含 a0b1 和 b1c1，c1 蕴含 a0c1 和 b1c1。

我们采用自上而下，宽度优先的搜索策略。在搜索过程中考虑下述三种情况，可有效减少所生成规则的数目，同时缩短搜索时间。

●若上层某泛化 X 能生成有效规则 X→Y，即 C（X→Y）≥θ，且 S（X→Y）≥φ，则他的所有子泛化不

必考虑。因为 X 的子泛化生成的规则相对于 X→Y 是冗余的。例如 a0→y 是有效规则，那么 a0b1 和 a0c1 不必考虑，因为既便它们也能生成有效规则，这些规则相对 a0→y 也是冗余的。这样，当某一泛化可生成有效规则时，可遍历并标识它的所有子泛化，搜索下层泛化时略去这些已标识的泛化。

●若上层某泛化 X 不能生成有效规则 X→Y，且 S(X→Y)<φ，则它的所有子泛

●若上层某泛化 X 不能生成有效规则 X→Y，且 C(X→Y) <θ，S (X→Y)≥φ，这时它的所有子泛化可以进一步考虑。

3．3 结合 GDT 的扩展概念格模型

将 GDT 作为假说搜索空间，可以从不一致和不完全的数据中挖掘规则，还可以对未见实例进行预测。于是我们考虑结合 GDT 与概念格，提出一个用于处理不完备数据的扩展概念格模型。由于我们只关心概念的内涵（对应泛化）及其所覆盖的实例所属的类别，所以将概念的外延中的对象用其所属类别来代替，同时引入一个概率值，来表征内涵（泛化）的强度[6]。扩展后的概念具有如下形式：

(s (X); n1, n2, … , nm; X)

其中， X 为概念的内涵； s (X)为概念内涵的强度，定义同泛化的强度；ni（i = 1, 2, …, m）为概念外延中属于类别 Yi 的对象的数目。

4 结语

形式概念分析作为一种用于数据组织和数据分析的形式化的工具，无论是对于理论研究还是实际应用都具有重要意义，并且被广泛而成功地应用于众多的领域。在此背景下，本文在以下方面展开了研究：

(1)将基于概念格的关联规则挖掘思想用于用户关联挖掘，定义了确定的用户关联基（规则的可信度=100％）和近似用户关联基（规则的可信度<100％）。这两个用户关联基是使用形式概念分析中的频繁封闭项集及其缩减来描述的，并且从它们可以导出所有的有效用户关联规则。进而提出了基于概念格的用户关联基提取算法。实验结果表明，使用这两个基于概念格的用户关联基可以在很大程度上减少用户关联规则的数目，并且不会丢失任何信息。

(2)将 GDT 作为假说搜索空间，提出了一种规则发现算法。使用该算法可以从不一致和不完全的数据中挖掘规则，同时该算法可过滤数据中的噪声，并且可以对未见实例进行预测。在分析 GDT 与概念格关系的基础上，结合 GDT 与概念格，提出一个用于处理不确定数据的扩展概念格模型，并给出了相应的规则挖掘算法。

参考文献：

[1] Agrawl R, Imielinski T, and Swami A (1993). Mining association rules between sets of items in large databases. In Proceedings of 1993 ACM_SIGMOD International Conference Management of Data, Washington, D.C., 207-216.

[2] Agrawl R, and Srikant R (1994). Fast algorithm for mining association rules. In Proceedings of 1994 International Conference on Very Large Data Bases, Santiago, Chile, 487-499.

[3] Amazon: Online shopping for electronics, apparel, music, books, DVDs & more. [http://www.amazon.com/]

[4]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):219- 226·

[5]孙士保,秦克云·基于剩余蕴涵的模糊概念格构造方法[J]·西南交通大学学报, 2006,41(2):259- 263·

[6] Burusco A, Gonzalez R F. The study of the L-fuzzy concept lattice[J]. Mathware and SoftComputing,1994,1(3): 209- 218.