形式概念中的知识获取与知识表示

任梦圆

（大连海事大学 信息科学技术学院 大连 116000）

摘 要

形式概念分析是德国学者 Wille 于 1982 年作为一种数学理论首先提出的，概念格结构模型是它的核心数据结构。概念格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和特化关系，其相应的 Hasse 图则实现了对数据的可视化。目前形式概念分析已被广泛地研究，并应用到机器学习、软件工程和信息获取等领域。

本文跟踪国际学术前沿，主要对形式概念分析中的知识表示和知识获取进行系列研究，所获研究成果不仅从理论上丰富和发展了形式概念分析，而且由于它们的广泛应用背景，这些结果同样具有重大的应用价值。

将形式概念分析理论引入到粗糙集中，提出基于形式概念分析的粗糙集模型。该模型首先解决信息系统上的代数结构问题，即在信息系统上诱导出一个格结构，格中的结点称为粗糙概念;然后探讨基于粗糙概念的信息系统中的一些常见问题的求解，如核和约简;最后给出粗糙概念在决策表中的应用。另外，决策依赖以其描述性强、便于理解等优点已之成为决策表中一种常用的知识表示形式，从而得到了广泛的应用因此，我们在决策表中应用推理规则得到一个完备且无冗余的决策依赖集。该模型为粗糙集提供一种新的合理化的解释，有利于人们从形式概念分析的角度加深对粗糙集的理解。

关键词：形式概念分析 粗糙集 决策规则

中图法分类号：TP301

Knowledge Representation and Knowledge Acquisiton on Formal Concept Analysis

RenMengyuan

Abstract Formal Concept Analysis (FCA) is an order-theoretic method for the mathematical analysis of scientific data, pioneered by R. Wille in mid 80's.Over the past twenty years, FCA has been widely studied and become a powerful tool for machine learning, software engineering and information retrieval.

Tracking the international research status, the thesis mainly researches knowledge acquisition models based on formal concept analysis theory. Achieved results not only enrich and improve formal concept analysis theory,but also are expected enormous applied value due to the widespread applied background of these theories.

By introducing formal concept analysis into rough set theory, we propose the rough set model based on formal concept analysis. In this model,a solution to an algebraic structure problem is first provided in an information system: a lattice structure is inferred from the information system and corresponding nodes are called rough concepts. How to deal with common problems in rough set theory based on rough concepts is then explored, such as upper and lower approximation operators,

作者简介：任梦圆（1994-）女，辽宁人，硕士研究生，研究方向：人工智能

reducts and cores. Decision dependency has become a common form of knowledge representation owing to its properties of expressiveness and ease of understanding, so it has been widely used in practice. Finally, application of rough concepts to the extraction of decision dependencies from a decision table is studied; a complete and non-redundant set of decision dependencies can be obtained from a decision table. This model not only provides a better understanding of rough set theory from the perspective of formal concept analysis, but also demonstrates a new way of combining rough set theory and formal concept analysis.Decision dependency has become a common form of knowledge representation owing to its properties of expressiveness and ease of understanding, so it has been widely used in practice. Finally, application of rough concepts to the extraction of decision dependencies from a decision table is studied; a complete and non-redundant set of decision dependencies can be obtained from a decision table.

Keywords： formal concept analysis; rough sets; decision rule;

1 背景

1.1 论文的研究背景

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急剧增加，并且由于人类的参与使数据与信息中的不确定性更加显著，信息与数据中的关系更加复杂。如何从大量的、杂乱无章的、强干扰的数据中挖掘潜在的、新颖的、正确的、有利于价值知识，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一一数据挖掘(DM)和数据库知识发现(KDD)。目前已有许多的数学挖掘工具，比如神经网络、遗传算法、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 Concept Analysis, FCA)对于处理复杂的信息不失为一种有效的方法。

1.2 形式概念分析的基本思想

由于许多应用数学问题需要借助于格论来实现其最终的应用，因此德国 Darmstadt 的研究小组便开始系统地研究和发展一种基于格论的应用软件。Darmstadt 研究小组从事多达上百个基于格论的应用软件研究，此时，Darmstadt 研究小组开始建立了一个小小的原型系统，并进一步使其逐步实用化。其中，从一个二维表中来构造完全的格结构在 Birkhoff 的格论[I]中已经加以解释和说明，的，这两个子背景分别称为条件子背景

但是由于新的应用目的和发展的需要，格论需要进一步的扩展和更深入的研究，而形式概念分析正是在这种情况下诞生的，它的首次描述是在 1981 年关于有序集合的 Banff 会议的专题演讲上，而人们真正了解形式概念分析是在 1982 年，是由 Wille 教授把它作为一种数学理论提出来的。形式概念分析作为一种强有力的数据挖掘工具，它的诞生引起了
人工智能工作者的很广泛关注，数以百计的相关论文开始发表和出版，甚至包括一些关于形式概念分析理论的数学基础的书籍。由此，形式概念分析从应用的角度开始走向理论的形式化研究，而理论研究将进一步促进形式概念分析在各个领域中的应用。

2 形式概念分析的基本概念

2.1 决策背景和决策规则

定义 1 形式背景 K=(G, M, I)被称为决策背景，如果 M=$C \cup D$，I=$I_C \cup I_D$ 其中 C 为条件属性，D 为决策属性，$I_C \subseteq G \times C$ 为条件关系集合，$I_D \subseteq G \times D$ 为决策属性关系集合。

事实上，决策背景是由两个子背景组成 $K_C \subseteq (G, C, I_C)$ 和决策子背景

$K_D \subseteq (G, D, I_D)$。其中 C 为条件属性，D 为决策属性。

定义 2 设 K=(G, M, I)为一决策背景，该决策背景被称为协调的决策背景，如果对于任意 g，h∈G, $g^{CC} = h^{CC}$ 意味着 $g^{DD} = h^{DD}$.

定义 2 表明，在一个协调的决策背景中，如果两个对象的条件属性相同，那么它们的决策属性也一定相同，这就意味着当我们进行决策时，对相同的条件不会得出不同的结果，即不会出现矛盾决策。下文中所指的决策背景都是协调的决策背景。

定义 3 设 K = (G, M, I)为一决策背景，$B_{c1} \subseteq C$ $B_{D1} \subseteq D$ 则 K 上的规则 $B_{C1} \to B_{D1}$ 为决策规则当且仅当 $B_{C1}^{C} \subseteq B_{D1}^{D}$

2.2 粗糙集理论

定义 3 一个信息系统 S 是一个四元组 S = (U, AT, V, f)，这里 U 称为论域，它的元素称为对象，AT 称为属性集，它的元素称为属性，U 与 AT 是有限的非空集合，$V = U_{a \in AT} V_a$ 称为属性值域，V_a 为属性 a 的值域;f 为 $U \times AT$ 到 V 的一个映射，对任意的 $x \in U$，$a \in AT$， $f(x,a) \in V_a$，通常称 f 为信息函数或描述函数。

设 S = (U, AT, V, f)为一信息系统，$B \in AT$ 为一属性子集，$x, y \in U$，等价关系

$$IND(B) = \{(x,y) | \forall a \in B, f(y,a)\}$$

称为 B 不可分辨关系。令

$$[x]_B = \{y \in U | (x,y) \in IND(B)\}$$

容易验证 $U/B = \{[X]_B | x \in U\}$ 为论域 U 的一个划分。

定义 4 设 S = (U, AT, V, ,f)为个信息系统，$x \in U$ 为一对象子集，$B \in AT$ 为一属性子集，X 关于 B 的下近似定义为:

$$\underline{B}(x) = \{x \in U | [x]_B \subseteq X\}$$

X 关于 B 的上近似定义为:

$$\overline{B}(x) = \{x \in U | [x]_B \cap X \neq \varnothing\}$$

3 形式概念分析对粗糙集理论的表示

定理 1(等价类表示)设((G, M, W , I)为一信息系统，(G, N, J)为相应的平面梯级衍生背景，$A \in G, B \in M$ 为信息系统中属性集 B 在衍生背景中相应属性集，即，

$$B_N = \{(m,n) | m \in B, n \in M_m\}$$

对于 $g \in G, [g]_B$ 表示在信息系统中由 g 决定的 B 等价类，在衍生的形式背景中令 $[g]_B = \{h \in G | g^J \cap B_N = h^J \cap B_N\}$，其中 g^J 和 h^J 表示衍生背景中对象 g 和对象 h 相对应的内涵，则有 $[g]_B = [g]_B$

定理 2(上下近似表示)设(G, M, W, 1)为一信息系统，(G, N, J)为相应的平面梯级衍生背景，$A \in G, B \in M$，$\overline{B}(A)$ 和 $\underline{B}(A)$ 分别为信息系统中 A 关于 B 的上近似和下近似，则

$\overline{B}(A)=\{g\in G\left|[g]_B\cap A\neq\varnothing\right.\}$。

$\underline{B}(A)=\{g\in G\left|[g]_B\subseteq A\right.\}$

4 形式概念分析的应用

4.1 形式概念分析在软件工程中的应用

在软件工程领域，形式概念分析为软件再生工程、面向对象程序设计和软件重用等应用提供了有效的理论支撑，并已经取得了大量的研究成果。在面向对象的程序设计方面，Arfi 等开发了一个用于生成和浏览 Smalltalk-80 类层次的工具;Godin 等开发了一个在类的规范说明中用于计算类层次的原型工具;针对类层次的一些设计上的缺陷，Snelting 等提出了一个用于检测和补救存在设计上缺陷的类层次的框架结构。在软件再生工程方面，Funk 等提出了基于概念格分解的软件配置管理;Lindig 等通过分析全局变量和过程变量间的关系构造出一个用于评估模块候选项之间祸合/内聚度的概念格结构;Snelting 利用概念格来组织和显示配置文件的结构;
在软件重用方面，Godin 等基于概念层次来组织和检索库中的产品;Lindig 允许用户渐进式地应用一系列的关键字来检索目标可重用软件构件。

4.2 形式概念分析在数据挖掘中的应用

关联规则获取是形式概念分析在数据挖掘中应用的最广也是研究成果最多的领域。例如 Godin 等提出了从概念格中提取蕴涵规则(精确规则)的算法; Missaoul 等提出了近似规则的提取算法;王志海等提出了概念格中提取规则的渐进式算法;Hu 提出了一种在概念格中提取关联规则和分类规则的集成算法;谢志鹏等提出了基于内涵缩减的关联规则提取算法。形式概念分析还被应用于数据挖掘中的分类知识获取，一些经典的分类系统包括:GRAND 系统，RULEARNER 系统 GALOIS 系统，CIBLe 系统，LEGAL 系统和 CLNN&CLNB 系统;胡学钢等提出了基于分布式扩展概念格的分类规则提取方法;Gupta 等提出了 CBALattice 分类方法;胡可云等讨论了分类/关联规则的集成挖掘方法，并提出 CLACF 分类方法。

4.3 形式概念分析在其他领域的应用

概念格还被应用于信息检索等多个方面。Cole 等[46]设计的 CEM 电子邮件管理系统将 Email 存储在概念格而非树状结构中;Neuss 等应用概念格结构完成了 Internet 文档信息的自动分类和分析;Eklund 等展示了概念层次在 Web 文档索引和导航中的能力;金阳等提出基于有序概念格的个性化搜索引擎导航;Godin 等对基于概念格结构的信息检索进行了实验;Fernandez-manj on 等把概念格当作教育软件设计过程中的支撑工具;Priss 提出一个可处理具有层次特征且附加有属性(术语)的文档检索系统;Bhatia 等提出一个基于概念聚类的信息检索;Carpinet。等基于概念格对文本数据库的混合导航和自动组织设计了 ULYSSES 检索系统。

5 总结与展望

形式概念分析理论是用来进行数据分析和处理的一种数学工具，由于其良好的数学性质，目前已引起人工智能工作者的广泛关注。本文跟踪国际学术前沿，把经典形式概念分析与多种软技术数据分析工具进行结合研究，提出了多个基于形式概念分析的知识获取模型。这些模型的提出不仅有助于丰富形式概念分析的理论基础和拓展形式概念分析的应用范围，而且有助于人们从形式概念分析的角度加深对其它学科的认识和理解。将形式概念分析引入到离散信息系统中，提出了基于形式概念分析的粗糙集模型。该模型解决了离散信息系统上的代数结构问题，即构建了粗糙概念格，同时探讨了如何基于粗糙概念来解决粗糙集中的一些重要问题，如上、下近似算子、属性约简等。

6 参考文献

[1] R.Wille. Restructuring lattice theory: an approach based on hierarchies of concepts.

In: Rival I ed. Ordered Sets. Dordrecht, Reidel, 1982, 445-470.

[2] B. Ganter, R. Wille. Formal Concept Analysis: Mathematical Foundations. Berlin, Springer-Verlag, 1999.

[3] L. Nourine,O. Raynaud. A fast algorithm for building lattices. Information Processing Letters, 1999, 71(5-6):199-204.

[4] M. Chein. Algorithme de recherche des sous-matrices premieres d'une matrice. Bull.Math. Soc. Sci. Math. R.S. Roumanie, 1969,13: 21-25.

[5] G. Stumme, R.Taouil, Y Bastide, N. Pasquier, L. Lakhal. Fast computation of

concept lattices using data mining techniques. In: Proceedings, 7th Int. Workshop on Knowledge Representation Meets Databases (KRDB 2000), Berlin, Germany, 2000,129-139.

[6] B. Ganter. Two basic algorithms in concept analysis. (Technical Report FB4- Preprint No. 831). TH Darmstadt, 1984.

[7]H. Alaoui. Algorithmes de manipulation du treillis de Galois d'une relation binaire et applications. Master Thesis, Universite du Quebec a Montreal, 1992.

[8]C.Lindig.Algorithmen zur begriffsanalyse and ihre anwendung bei softwarebiblio-theken.(Dr.-Ing.)Dissertation, Techn. Univ. Braunschweig, 1999.

[9] J.P. Bordat. Calcul pratique du treillis de Galois d'une correspondance. Mat. Sci.Hum., 1986, 96: 31-47.

[10] R. Godin, R. Missaoni, H. Alaoui. Incremental concept formation algorithms based on Galois lattices. Computation Intelligence, 1995, 11(2):246-267

[11]谢志鹏，刘宗田. 概念格与关联规则发现. 计算机研究与发展，2000

[12]基于形式概念分析理论的知识获取模型研究，康向平。2012