科学版

研究生教学丛书

信息系统与知识发现

张文修　梁怡　吴伟志 编著

科学出版社
www.sciencep.com

信息系统与知识发现

科学版研究生教学丛书

信息系统与知识发现

张文修　梁怡　吴伟志　编著

本书得到国家重点基础研究发展计划(973 计划)项目“复杂生产制造过程实时智能控制与优化理论和方法研究”资助

科学出版社
北京

内 容 简 介

本书以粗糙集、模糊集、随机集理论为工具，论述信息系统上的知识发现与知识约简理论.内容包含经典信息系统与知识发现，模糊信息系统与知识发现，随机信息系统与知识发现，格值信息系统与知识发现，以及知识发现形成的知识系统与知识逻辑.本书坚持严格的数学模式与方法的实际可实现性紧密结合,既可以作为数学与信息科学研究生教材，又可以作为从事信息科学研究人员与工程人员的参考书.

图书在版编目(CIP)数据

信息系统与知识发现/张文修,梁怡,吴伟志编著.—北京:科学出版社,2003

科学版研究生教学丛书

ISBN 7-03-011521-X

Ⅰ.信… Ⅱ.①张… ②梁… ③吴… Ⅲ.①信息系统-研究生-教学参考资料②知识学-研究生-教学参考资料 Ⅳ.①G202②G302

中国版本图书馆 CIP 数据核字(2003)第 042963 号

责任编辑:杨 波/责任校对:宋玲玲

责任印制:安春生/封面设计:黄华斌

科学出版社 出版

北京东黄城根北街16号

邮政编码:100717

http://www.sciencep.com

源海印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2003 年 9 月第 一 版 开本:B5(720×1000)

2003 年 9 月第一次印刷 印张:15 3/4

印数:1—4 000 字数:299 000

定价:24.00 元

(如有印装质量问题,我社负责调换〈新欣〉)

前　言

知识是人类认识客观世界的结果，同时也是指导人们行为的准则，在知识经济的时代里，知识是社会发展的重要动力，是决定生产力发展的主要因素.特别是随着时代的变化，环境的变化，认识的深入，人们必须不断地获取与发现新的知识.人们有各种获得知识与发现知识的手段，而其中最重要的一种手段是从数据中发现知识，称为数据库中的知识发现(knowledge discovery in database，简称 KDD).

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(信息爆炸)，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(复杂系统).如何从大量的、杂乱无章的、强干扰的数据(海量数据)中挖掘潜在的、新颖的、正确的、有利用价值的知识(有用知识)，这给智能信息处理提出了严峻的挑战.

同样是基于数据库的知识发现，有着完全不同的方法，使用着完全不同的数学工具.比如基于网络结构的神经网络算法，基于训练选优的遗传算法，基于统计理论的数据挖掘与支持向量机方法，基于归纳学习的机器学习方法，基于范例的推理方法，基于生物信息的知识发现方法等.本书重点是在信息系统上，利用粗糙集(rough set)、模糊集(fuzzy set)和随机集(random set)的三集理论的知识发现方法.这种方法也称为概率方法.

信息系统是一个具有对象和属性(条件属性与目标属性)关系的数据库，这种数据库通过数据隐含着知识的对象与属性之间的关系，最终表达的知识模式是用属性来表达的，它有明确的直观意义，因此是可以理解的.由于数据表的规模性和多样性(定性值、定量值、离散值、连续值、缺省值、集合值等)，知识表达的对象与属性的关系不是能直接观察到的，必须依赖于一定的数学方法与计算工具，因此模式的获得是非平凡的.

作为知识发现的概率分析方法的工具，粗糙集、模糊集、随机集等几乎同时产生于20世纪的下半世纪.特别是1982年，波兰数学家 Z. Pawlak 提出粗糙集以后，粗糙集方法在知识发现中的作用日益显著，同时也使随机集与模糊集在知识发现中的作用日益增强.使三集理论成为处理不精确、不确定、不协调、不完全信息系统的数学基础.

本书以信息系统为基本研究对象，以三集理论(粗糙集、随机集、模糊集)为工具，以知识发现为目的，系统地介绍了各种信息系统上的知识发现与知识约简，既有明确的应用目标，又有严格的数学模式与结构，力求达到理论与实际、方法与应用的统一，使对于知识发现的研究人员与工程人员都能受益.本书中包含了国内外许多专家的重

要成果，同时也包含了作者的某些重要成果，也包含了作者指导下的博士生与同事的研究成果.特别要指出的是，书中包含有徐宗本教授、梁吉业教授、米据生博士、陈德刚博士后、李德玉博士、张梅博士等人的重要成果，在此表示衷心感谢.

本书虽经再三修改，仍会有不少漏洞，热忱欢迎广大读者批评指正.

目　录

第一章　信息系统与知识发现

§1.1　信息系统与知识发现

知识是人类认识客观世界的结果，同时也是人们指导自己行为的准则．人们可以从不同的途径获取知识，比如实践中获得的经验，各种渠道 (网络、书刊、交流) 获得的信息，领导和教授提供的指导与结论，自己头脑的思考等，都是获取知识与发现知识的重要手段．但是不同的知识发现手段有着不同的方法，我们这里讲的知识发现是一种特定的知识发现，它是从数据集中识别正确、新颖、有潜在应用价值以及最终可为人们理解的模式的方法．这种方法的特点是：

(1) 基础信息是数据库．数据库中的数据不是孤立存在的，它必须与一定的研究对象以及对象所反映的一定属性相联系．也就是说某个数据 v 同时与对象 x 以及属性 a 相关联．

(2) 模式是可以理解的．最终表达的知识模式是用属性表达的，它有明确的直观意义，符合人们的直观理解，且方便人们的应用．

(3) 模式的获取是非平凡的．从数据中获取的模式不是直观的，不可能是直接观察的结果，它依赖于一定的数学方法和计算机工具．

知识发现一直是人工智能的核心问题，但是这样一种特定的知识发现被正式提出来，当属于 1989 年 8 月在美国底特律召开的第 11 届国际人工智能联合会议的专题讨论会上．从那以后，知识发现，或者说数据库中的知识发现 (knowledge discovery in database, 简称 KDD) 备受重视．

同样是基于数据库的知识发现，有着完全不同的方法，使用着完全不同的数学工具．比如基于网络结构的神经网络算法，基于训练选优的遗传算法，基于统计理论的数据挖掘方法，基于归纳学习的机器学习方法等．我们这里的重点是基于粗糙集 (rough set) 、模糊集 (fuzzy set) 和随机集 (random set) 的三集理论的知识发现方法．这种方法也可以称为概率分析方法，因为其他的方法基本上属于统计训练方法．

首先，我们叙述知识发现的数据库的表示方法．

定义 1.1　称 (U, A, F) 为一个信息系统，或者数据库系统，其中 U 为对象集，即

$$U = \{x_1, x_2, \cdots, x_n\}.$$

U 中的每个 $x_i(i \leq n)$, 称为一个对象. 而 A 为属性集, 即

$$A = \{a_1, a_2, \cdots, a_m\}.$$

A 中的每个 $a_j(j \leq m)$, 称为一个属性. F 为 U 和 A 的关系集, 即

$$F = \{f_j : j \leq m\}.$$

其中 $f_j : U \to V_j$ $(j \leq m)$, V_j 为属性 a_j 的值域.

在信息系统中, 关系集是非常重要的. 如果 F 不存在, 对象集 U 与属性集 A 之间是孤立的. 关系集 F 表达了对象集 U 与属性集 A 之间的联系, 这正是知识发现所需要的信息基础. 比如 $f_j(x_i) = v$ 表示了对象 x_i 的属性 a_j 具有属性值 v.

属性值域 $V_j(j \leq m)$ 可以是定量值, 也可以是定性值. 由于在知识发现的问题中主要是分类问题, 我们用不同的数值表示不同的定性属性并不会影响知识发现的过程与结果. 这里 V_j 一般取有限个数值, 并且取

$$V = \bigcup_{j=1}^{m} V_j.$$

这样使 $f_j : U \to V$ $(j \leq m)$.

例 1.1 表 1.1 我们给出有三种症状 a_1, a_2, a_3 和 10 个病例 $x_1 \sim x_{10}$ 所表示的信息系统:

表 1.1 病例信息系统

U	a_1	a_2	a_3
x_1	2	1	3
x_2	3	2	1
x_3	2	1	3
x_4	2	2	3
x_5	1	1	4
x_6	1	1	2
x_7	3	2	1
x_8	1	1	4
x_9	2	1	3
x_{10}	3	2	1

例 1.1 给出了一个信息系统, 其中对象集为

$$U = \{x_1, x_2, \cdots, x_{10}\}.$$

属性集为

$$A = \{a_1, a_2, a_3\}.$$

属性的值域分别为

$$V_1 = \{1, 2, 3\},$$
$$V_2 = \{1, 2\},$$
$$V_3 = \{1, 2, 3, 4\}.$$

从而

$$V = \{1, 2, 3, 4\}.$$

$F = \{f_1, f_2, f_3\}$, f_1 表示症状 a_1 对于不同病例对象的取值, f_2 表示症状 a_2 对于不同病例对象的取值, f_3 表示症状 a_3 对于不同病例对象的取值. 例如: $f_1(x_2) = 3$, $f_2(x_5) = 1$, $f_3(x_8) = 4$ 等.

由例 1.1 不难看出, 一个信息系统对应着一个关系数据表, 一个关系数据表也对应着一个信息系统. 信息系统是数据表的抽象描述.

定义 1.2 设 U 是对象集, 记

$$U^2 = U \times U = \{(x_i, x_j) : x_i, x_j \in U\}. \tag{1.1}$$

$R \subseteq U^2$ 称为 U 上的一个等价关系, 若 R 满足以下条件:

(1) 自反性: $(x_i, x_i) \in R$, $(\forall i \leq n)$.

(2) 对称性: $\forall i, j \leq n$, $(x_i, x_j) \in R \Longrightarrow (x_j, x_i) \in R$.

(3) 传递性: $\forall i, j, k \leq n$,$(x_i, x_j) \in R, (x_j, x_k) \in R \Longrightarrow (x_i, x_k) \in R$.

定义 1.3 设 U 是对象集, 若存在 $C_i \subseteq U$ $(i \leq k)$, 满足以下条件:

(1) $C_i \neq \emptyset$ $(i \leq k)$,

(2) $C_i \cap C_j = \emptyset$ $(i \neq j)$,

(3) $\bigcup_{i=1}^{k} C_i = U$,

称 $\mathcal{A} = \{C_1, C_2, \cdots, C_k\}$ 为 U 的一个分划. 我们用 $\mathcal{G}$ 表示 U 上的分划全体.

定理 1.1 U 上的等价关系必产生 U 上的一个分划, U 上的一个分划必由 U 上的一个等价关系产生. U 上的等价关系与分划一一对应.

证明 设 R 是 U 上的一个等价关系, 记

$$[x_i]_R = \{x_j : (x_i, x_j) \in R\}. \tag{1.2}$$

由于 $x_i \in [x_i]_R, [x_i] \neq \emptyset$. 一方面, 当 $x_j \in [x_i]_R$ 时, 有 $x_i \in [x_j]_R$, 由 R 的对称性和传递性有 $[x_i]_R = [x_j]_R$; 而另一方面, 若 $x_j \notin [x_i]_R$, 则 $[x_i]_R \cap [x_j]_R = \emptyset$. 最后显然

有 $\bigcup_{i=1}^{n}[x_i]_R = U$. 于是 R 产生了 U 上的一个分划：

$$\mathcal{A} = U/R = \{[x_i]_R : x_i \in U\}.$$

$[x_i]_R$ 称为含 x_i 的等价类. 反之，若

$$\mathcal{A} = \{C_i : i \leq k\}$$

为 U 上的一个分划，记

$$R = \{(x_i, x_j) : \text{存在 } l \leq k \text{ 使 } x_i, x_j \in C_l\},$$

则 R 是 U 上的等价关系，且当 $x_i \in C_l$ 时必有 $[x_i]_R = C_l$. □

定理 1.2 设 (U, A, F) 是一个信息系统，对于 $B \subseteq A$,

$$R_B = \{(x_i, x_j) : f_l(x_i) = f_l(x_j) \quad (\forall a_l \in B)\} \tag{1.3}$$

是 U 上的一个等价关系，从而产生 U 上的一个分划：

$$\mathcal{A} = U/R_B = \{[x_i]_B : x_i \in U\},$$

其中

$$[x_i]_B = \{x_j : (x_i, x_j) \in R_B\} = \{x_j : f_l(x_j) = f_l(x_i) \ \ (a_l \in B)\}.$$

证明 可直接验证 R_B 满足自反性、对称性和传递性. □

在例 1.1 中，有

$$R_A = \{(x_i, x_j) : f_l(x_i) = f_l(x_j) \ \ (l = 1, 2, 3)\}.$$

从而可产生分划：

$$U/R_A = \{C_1, C_2, C_3, C_4, C_5\},$$

其中

$$\begin{aligned} C_1 &= \{x_1, x_3, x_9\}, \\ C_2 &= \{x_2, x_7, x_{10}\}, \\ C_3 &= \{x_4\}, \\ C_4 &= \{x_5, x_8\}, \\ C_5 &= \{x_6\}. \end{aligned}$$

每一个 C_l 表达了一个可以理解的知识. 如果用 (a_j, l) 表示知识 "属性 a_j 具有属性值 l", "$\wedge$" 表示逻辑 "与" 运算, 则

$$
\begin{aligned}
C_1 &\sim (a_1, 2) \wedge (a_2, 1) \wedge (a_3, 3),\\
C_2 &\sim (a_1, 3) \wedge (a_2, 2) \wedge (a_3, 1),\\
C_3 &\sim (a_1, 2) \wedge (a_2, 2) \wedge (a_3, 3),\\
C_4 &\sim (a_1, 1) \wedge (a_2, 1) \wedge (a_3, 4),\\
C_5 &\sim (a_1, 1) \wedge (a_2, 1) \wedge (a_3, 2).
\end{aligned}
$$

如果把每个 $C_i (i \leq 5)$, 冠以不同的名称, 则得到不同的概念, 而且这个概念可以用满足某些条件的属性来表达, 因而是可以理解的.

信息系统的知识发现问题本质上是按照属性特征将对象进行分类的问题. 面对巨型数据库系统, 以及数据库系统表现的多样性, 能够识别出正确、新颖和有潜在应用价值的模式, 仍然需要非平凡的手段, 这样, 就自然产生了一系列需要研究的问题:

(1) 知识约简问题. 一般说来, 描述不同对象特征的属性集是较大的, 但是对于信息系统分类的知识发现来说有些属性并不总是必要的. 不同属性对于分类知识发现来说, 有些属性是绝对不必要的, 去掉这种属性并不影响分类的知识发现; 而有些属性是绝对必要的, 去掉这种属性必然会影响分类的知识发现. 还有一些属性是相对必要的属性, 它可能与其他一些属性联合起来确定分类的知识发现, 但是也存在另外一种不需要这种属性的属性集也可以确定知识发现. 知识约简问题就是要在属性集中寻找一个最小的属性集, 它能完全确定知识发现, 也即由这个最小属性集确定的分类知识与用全体属性集确定的分类知识是相同的. 通过例 1.1 可以看出, 去掉属性 a_1 后仍然分为 $C_1, C_2, \cdots, C_5$ 这 5 类, 即分类不变, 因此属性 a_1 对于分类知识发现来说是绝对不必要的. 但是去掉属性 a_3 以后, 使 C_4 与 C_5 合并为一类, 使 5 类变为 4 类, 这时就影响到分类知识; 去掉属性 a_2 后, 使 C_1 与 C_3 合并为一类而成为 4 类, 也影响到分类知识. 属性 a_2 与 a_3 都是不能去掉的. 这样, 属性集 $\{a_2, a_3\}$ 就构成了知识分类的一个最小属性集. 特别是在属性较多的情况下, 如何寻找这些不影响分类知识发现的最小属性集成为分类知识发现的一个重要课题. 最小属性集可以使分类知识表示简化, 而又不丢失任何信息, 这正是人们所期望的. 同时, 通过知识约简, 去掉了不必要的属性, 深化了人们对于知识的认识, 同样也是人们所期望的.

(2) 知识融合问题. 我们知道, 知识的概念是内涵与外延的统一体, 内涵由属性表示, 外延用对象表示. 分类知识即是满足一定属性性质的对象集与相应的对象集满足的属性性质是完全一致的. 对于信息系统的分类知识, 就存在分类知识以外的对象集. 在例 1.1 中有 10 个对象, 它有 $2^{10}-1$ 个非空子集, 而只有 $C_1, C_2, \cdots, C_5$ 这 5 个子

集是确定的元知识,这就存在着其他的对象子集如何用确定的元知识 $C_1, C_2, \cdots, C_5$ 表达的问题,显然其中有些可以表示成为 $C_1, C_2, \cdots, C_5$ 这 5 个集合中的一些并集,这时称为确定的知识,它也是可以理解的,可以通过元知识的逻辑“或”表示,但是这种并集也只有 2^5-1 个,绝大部分对象子集还不能用 $C_1, C_2, \cdots, C_5$ 及其并集来表示. 如果视 $C_1, C_2, \cdots, C_5$ 为已知的知识,那么就存在着其他的对象子集如何用已知的知识表示的问题. 另外,从属性的角度来看,分类知识是可以理解的,在于它可以用属性值来描述. 但是属性值的组合是很多的,而分类知识远远小于属性值的组合,那么就存在如何用已知的分类的属性去理解刻画其他的属性值组合. 在例 1.1 中, $|V_1|=3$, $|V_2|=2$, $|V_3|=4$, (a_1, a_2, a_3) 可以有 24 种属性值表达,而分类知识只有 5 种,这就存在着其他的属性值组合怎样用已知的 5 种属性值组合表示的问题.

(3) 信息系统多样性. 我们在例 1.1 中给出的信息系统是最基本的数据库系统. 它直接通过属性给出对象间的联系,从而产生对象集的一个明确的分类,即分划,任何两类之间是互不相交的. 如果作为分划的不相交性不成立,它构成了一个对象集的覆盖,即 $\bigcup_{l=1}^{k} C_i = U$. 由于存在 C_i, C_j, 使 $C_i \cap C_j \neq \emptyset$ $(i \neq j)$, 则有 $x \in U$ 使 $x \in C_i$, 且 $x \in C_j$, 那么对于 $x \in U$ 就存在进一步识别的问题. 对于其他的对象集的描述也带来了困难.

(4) 信息缺省问题. 在一个信息系统中,对于每一个属性 a_l $(l \leq m)$, 及对象 x_i, 必有 $f_l(x_i) \in V_l$ 存在. 如果存在 a_l $(l \leq m)$, 及对象 x_i, 使得 $f_l(x_i)$ 是缺省值,这时称信息系统是信息缺省的. 对于信息缺省的对象一般无法归类. 在通常的知识发现中,一般采用数据的预处理. 可以在有关领域专家的指导下,通过对数据集的分析,填补缺省数据,也可以从样本数据出发,通过某种训练算法,推算缺省的数据. 在信息缺省的情况下,不经过信息的预处理,直接建立知识发现的方法,是一个值得探索的问题.

(5) 信息不确定问题. 信息不确定性问题与信息缺省问题不同,对于某些属性 a_l, 以及某些对象 x_i, $f_l(x_i)$ 不是取 V_l 中的一个值,也不是缺省值,而是取 V_l 中的几个值,即 $f_l(x_i)$ 取 V_l 中的一个子集,这时也称为信息系统是有噪声的. 在通常的知识发现中,要进行数据清理去掉噪声. 这可以在领域专家指导下,通过对数据集的分析,清除噪声;也可以从样本数据出发,通过某种训练算法去掉噪声数据. 但是在信息不确定的情况下,直接建立知识发现方法,将是知识发现的一种新思路.

(6) 数据的连续性问题. 由于要通过信息系统得到分类知识,一般要求属性值域是有限的. 对于属性值域是连续的情况,一般是将其离散化. 但是离散化的结果,可能会使信息大量丢失. 如果能够直接建立连续值情况下的知识分类与知识发现方法,将会使发现的知识更符合实际.

通过信息系统去发现知识，本质上是一个分类问题，即通过相同的属性汇聚不同的对象，或者对于不同的对象寻找相同的属性. 对象集与属性集正好相反，对象集越大，相同的属性就越少；反之，对象集越少反映的属性就越多. 人们可以理解的知识，一般是用属性表达的，属性越多，知识所包含的信息越多. 在知识的发现过程中，人们可能对于某一个对象集只发现了一部分属性，或者只发现了满足某些属性的一部分对象，这时知识还是不完全的. 只有在属性集与对象集完全对应的情况下才是完全的知识. 因此，我们所能做的知识发现，是在一定条件下的知识发现，比如在已有的信息系统下的知识发现. 这种知识一般来说，仍然是不完全的知识，随着信息系统的变化，或者是属性的增加与减少，或者是观察对象的增加与减少，发现的知识也会在不断地变化. 因此，知识发现的过程本质上是一个动态过程. 我们不仅要研究通过信息系统的知识发现的方法，还需要研究信息系统变化情况下知识发现的修正方法与递推方法，不断地完善我们发现的知识.

§1.2 目标信息系统与知识发现

信息系统的知识发现是根据不同的属性对于对象的分类问题，因此信息系统的知识发现是概念的发现，对于不同的分类产生不同的概念. 目标信息系统是研究条件属性与目标属性之间的关系问题，因此目标信息系统的知识发现是命题的发现，从条件属性与目标属性之间的不同关系，可得到不同的命题.

下面我们给出目标信息系统的概念.

定义 1.4 称 (U, A, F, D, G) 为目标信息系统或决策表，其中 (U, A, F) 是信息系统，A 称为条件属性集，D 称为目标属性集或决策属性集，即

$$D = \{d_1, d_2, \cdots, d_p\}.$$

G 为 U 和 D 的关系集，即

$$G = \{g_j : j \leq p\},$$

其中 $g_j : U \to V_j'$ $(j \leq p)$, V_j' 为目标属性 d_j 的值域.

在目标信息系统中，关系集 G 同关系集 F 一样，同样是重要的. 如果 G 不存在，对象集 U 与目标属性集 D 之间是孤立的，关系集 G 表达了对象集与目标属性集之间的联系，从而通过对象集使条件属性与目标属性之间建立了联系，这正是目标信息系统命题知识发现所需要的信息基础. 比如 $f_j(x_i) = v$ 表示对象 x_i 的条件属性 a_j 具有属性值 v, $g_l(x_i) = u$ 表示对象 x_i 的目标属性 d_l 具有属性值 u, 这样对于 x_i 来讲，$(a_j, v) \to (d_l, u)$.

目标属性值域 V'_j $(j \le p)$, 可以是定量值, 也可以是定性值, 同样可以用不同的数据值表示不同的定性属性, 并且记

$$V' = \bigcup_{j=1}^{p} V'_j$$

使 $g_j : U \to V'$ $(j \le p)$.

例 1.2 在例 1.1 的基础上, 我们进一步给出病例诊断信息系统 (见表 1.2).

表 1.2 病例诊断信息系统

U	a_1	a_2	a_3	d
x_1	2	1	3	1
x_2	3	2	1	2
x_3	2	1	3	1
x_4	2	2	3	2
x_5	1	1	4	3
x_6	1	1	2	3
x_7	3	2	1	2
x_8	1	1	4	3
x_9	2	1	3	1
x_{10}	3	2	1	2

例 1.2 给出了一个目标信息系统, 其中 U, A, V_1, V_2, V_3, V 与例 1.1 相同, F 与例 1.1 相同, 而

$$D = \{d\}, \quad V' = \{1, 2, 3\}.$$

$G = \{g\}$, g 表示目标属性 d 对于不同的病例诊断结果. 例如 $g(x_2) = 2$, $g(x_5) = 3$, $g(x_9) = 1$ 等.

由例 1.2 不难看出, 一个目标信息系统对应着一个含有条件属性与目标属性的数据表, 一个含有条件属性与目标属性的数据表对应着一个目标信息系统. 目标信息系统是含有条件属性与目标属性的数据表的抽象描述.

通过表 1.2 可以将对象集分为三个集合:

$$\begin{aligned} D_1 &= \{x_i : g(x_i) = 1\} = \{x_1, x_3, x_9\}, \\ D_2 &= \{x_i : g(x_i) = 2\} = \{x_2, x_4, x_7, x_{10}\}, \\ D_3 &= \{x_i : g(x_i) = 3\} = \{x_5, x_6, x_8\}. \end{aligned}$$

同样 $\mathcal{D} = \{D_1, D_2, D_3\}$ 也构成了 U 的一个分划. 分划中的每个元素 $D_k (k \le 3)$ 表

达了一个知识：

$$D_1 \sim (d,1),$$
$$D_2 \sim (d,2),$$
$$D_3 \sim (d,3),$$

其中 (d,l) 表示目标属性 d 具有属性值 l.

通过条件属性得到了 U 的一个分划：

$$\mathcal{A} = \{C_1, C_2, C_3, C_4, C_5\}.$$

通过目标属性也得到了 U 的一个分划：

$$\mathcal{D} = \{D_1, D_2, D_3\}.$$

且两个分划之间有关系：

$$C_1 = D_1, \quad C_2 \cup C_3 = D_2, \quad C_4 \cup C_5 = D_3.$$

如果 $A_j \subseteq D_l$, 我们就得到了一条命题知识或决策规则，即通过 A_j 的条件属性值可以得到 D_l 的目标属性值，于是有

$$(a_1,2) \wedge (a_2,1) \wedge (a_3,3) \to (d,1),$$
$$(a_1,3) \wedge (a_2,2) \wedge (a_3,1) \to (d,2),$$
$$(a_1,2) \wedge (a_2,2) \wedge (a_3,3) \to (d,2),$$
$$(a_1,1) \wedge (a_2,1) \wedge (a_3,4) \to (d,3),$$
$$(a_1,1) \wedge (a_2,1) \wedge (a_3,2) \to (d,3).$$

这样就得到了 5 条命题知识. 对于由目标信息系统得到的命题同样有以下一些需要研究的问题：

(1) 知识约简问题. 这里的知识约简不同于信息系统的知识约简，它是与目标属性 d 密切相关的. 不同的条件属性对于目标分类来讲作用是不同的. 有些条件属性对于目标分类来讲是绝对不必要的，去掉这种条件属性并不影响目标分类；而有些条件属性对目标分类来讲是绝对必要的，去掉这种属性必然会影响目标分类；还有一些条件属性对于目标分类来讲是相对必要的，它可能与绝对必要属性联合起来确定目标分类，但是也存在另外一些属性与绝对必要属性联合起来确定目标分类. 目标信息系统的知识约简就是在条件属性中寻找一个最小的属性集，可以完全确定目标分类，它与信息系统的知识约简不完全相同. 比如在信息系统中，a_1 是不必要属性，$\{a_2, a_3\}$ 是信息系统中的知识约简属性集. 但是在目标信息系统中，