

梁吉业 著
李德玉



信息系统中的 不确定性 与知识获取

 科学出版社
www.sciencep.com

G202
39



信息系统中的 不确定性 与知识获取

梁吉业 著
李德玉

科学出版社
北京

5CH/09/01

内 容 简 介

本书是一部研究信息系统中的不确定性与知识获取的专著。书中在粗糙集理论框架下,以熵理论为工具,系统地阐述了信息系统中不确定性的表示、度量与计算,并将其应用于完备信息系统和不完备信息系统的知识约简与知识获取中,反映了当前该领域的最新研究成果。

本书可作为计算机科学、应用数学、自动控制、信息科学和管理工程等专业高年级本科生的参考书及硕士、博士研究生教材,同时对相关学科领域的研究人员和工程技术人员也有重要的使用和参考价值。

图书在版编目(CIP)数据

信息系统中的不确定性与知识获取/梁吉业,李德玉著. —北京:科学出版社,2005

ISBN 7-03-015192-5

I. 信… II. ①梁…②李… III. 信息系统-研究 IV. G202

中国版本图书馆 CIP 数据核字(2005)第 021828 号

责任编辑:李 敏 李晓华 / 责任校对:张 琪

责任印制:钱玉芬 / 封面设计:朱 平

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>

双 青 印 刷 厂 印 刷

科学出版社发行 各地新华书店经销

*

2005年3月第 一 版 开本:B5(720×1000)

2005年3月第一次印刷 印张:8

印数:1—3 000

字数:156 000

定价:20.00 元

(如有印装质量问题,我社负责调换(环伟))

序

知识获取是人工智能领域最为核心的内容之一,对知识(信息)的不确定性研究与探索贯穿于人工智能近半个世纪发展历程的始终.20世纪80年代初,波兰学者 Z. Pawlak 创立了粗糙集理论,发现并描述了信息系统中知识的粗糙性.由于信息系统可以作为人工智能众多分支领域中问题的描述框架,所以粗糙集理论创立不久就得到了国内外学者的广泛关注,并在机器学习、模式识别、决策分析、医疗诊断、金融数据分析、过程控制、数据库知识发现等领域得到成功应用.1995年 ACM Communication 将粗糙集理论列为计算机学科的新研究课题,并在 Internet 上定期发布电子公告,许多国际会议将其列为专题.在我国,从2001年开始,每年举办一届粗糙集与软计算学术会议.目前,粗糙集理论已成为最为重要的知识发现方法之一.

该书汇集了梁吉业教授、李德玉教授在博士、博士后期间的主要研究成果.这些成果主要发表在“*International Journal of General Systems*”、“*Information Sciences*”、“*International Journal of Computational Cognition*”和“*International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*”等国际学术刊物上,引起了同行学者的广泛关注.书中详细地介绍了作者在粗糙集理论框架下,对信息熵、粗糙熵、知识粒度、包含度和知识获取等信息系统中一系列重要问题的研究成果,内容丰富、文献翔实,既注重理论推导的严谨性,同时又兼顾应用.相信该书的出版将对我国在粗糙集理论与软计算方面的研究与应用起到积极的推动作用.

张钊 院士

2004年12月

前 言

信息系统这个概念作为智能信息处理领域的一个具有形式化定义和明确语义的专业术语,出现于波兰学者 Z. Pawlak 提出的粗糙集理论中,数据库模型、数据的决策与分析、模式识别、机器学习等领域的许多问题的描述框架可以归结为信息系统.信息系统中的不确定性问题、知识获取问题一直是信息科学领域关注的热点.

粗糙集理论是波兰学者 Z. Pawlak 于 1982 年提出的一种数据分析理论.经过二十余年的发展,由于其在数据的决策与分析、模式识别、机器学习与知识发现等方面的成功应用,引起了世界范围内各国学者的广泛关注.该理论与统计方法、模糊方法、证据理论等其他处理不确定性问题的方法相比,最显著的特点是无需提供所需处理的数据集合之外的任何先验信息.目前,粗糙集理论、神经网络、演化计算、模糊系统及混沌系统已被公认为人工智能的五大新兴技术,其影响已经渗透到信息科学的几乎所有分支.

众所周知,对知识(信息)之不确定性研究与探索贯穿于人工智能这一综合性边缘学科近半个世纪发展历程的始终.熵的概念最早被用于度量热力学系统的无序性和紊乱性,1948 年 C. E. Shannon 借用熵的概念度量系统的不确定性.现在,熵这一概念几乎成了不确定性的代名词,在过去的五十余年中涌现了许多成功的范例.

本书在粗糙集理论框架下,以熵理论为工具,系统地阐述了信息系统中不确定性的表示、度量与计算,并将其应用于完备信息系统和不完备信息系统的知识约简与知识获取中,反映了当前该领域的最新研究成果.全书共分 8 章.第 1 章在信息系统框架下介绍信息系统、粗糙集与粒度计算的基本概念;第 2 章建立了包含度与粗糙集数据分析中的度量之间的关系;第 3 章阐述了完备信息系统下的信息熵、粗糙熵、粒度度量与知识粒度的概念,并建立了它们之间的关系;第 4 章提出了不完备信息系统下的信息熵、粗糙熵、粒度度量与知识粒度的概念,并建立了它们之间的关系;第 5 章从区分性的角度提出了粗糙性的拓扑熵,并讨论了其

在知识约简中的应用;第6章建立了完备、非协调信息系统中各类约简之间的关系,并提出了决策规则优化方法;第7章建立了不完备信息系统知识表示与知识获取的极大相容块技术;第8章给出了信息系统的同态不变特征.

本书是著者近年来在不确定性信息处理、基于粗糙集理论的知识获取方法等方面主要研究成果的总结.本书的出版得到了国家自然科学基金(No. 60275019, No. 70471003)、国家863计划(No. 2004AA115460)、国家重点基础研究发展计划(973)(No. 2004CB318108)、国家自然科学基金重点项目(No. 60135010)、国家自然科学基金创新研究群体项目(No. 60321002)、山西省高等学校拔尖创新人才基金、山西省自然科学基金(No. 20031036, No. 20041040)、中国博士后基金(No. 2003-20452)的部分资助,在此一并表示感谢.

梁吉业 李德玉

2004年12月

目 录

1 信息系统、粗糙集与粒度计算的基本概念	(1)
1.1 信息系统	(1)
1.2 近似与粗糙集	(2)
1.3 属性约简	(5)
1.4 属性依赖性	(7)
1.5 区分矩阵与区分函数	(8)
1.6 粒度计算	(11)
2 包含度与粗糙集数据分析中的度量	(13)
2.1 粗糙包含与包含度	(13)
2.2 粗糙集中基本概念的包含度表示	(15)
2.3 包含度与粗糙集数据分析中度量之间的关系	(15)
2.3.1 粗糙集的近似精度与粗糙隶属度可以归结为包含度	(15)
2.3.2 近似分类精度与近似分类质量可以归结为包含度	(16)
2.3.3 属性依赖性度量可以归结为包含度	(17)
2.3.4 属性重要性度量可以归结为包含度	(18)
2.3.5 决策规则的可信度和覆盖度可以归结为包含度	(19)
2.3.6 可变精度粗糙集模型的包含度表示	(19)
3 信息系统中的不确定性与知识粒度	(21)
3.1 Shannon 熵	(21)
3.2 信息熵与知识粒度	(23)
3.3 粒度度量与粗糙熵	(32)
4 不完备信息系统中的不确定性与知识粒度	(37)
4.1 不完备信息系统	(37)
4.2 信息熵与知识粒度	(39)
4.3 粒度度量与粗糙熵	(42)
5 粗糙性的拓扑熵及其应用	(47)
5.1 现有不确定性度量方法评述	(47)
5.2 有限拓扑空间及其熵	(49)
5.2.1 有限拓扑空间及其原子基	(49)
5.2.2 一个在给定集合上产生拓扑的方法	(50)

5.2.3	有限拓扑空间的熵	(51)
5.2.4	拓扑熵的图论表示	(53)
5.3	信息系统的数学本质——有限拓扑空间	(54)
5.3.1	信息系统及其信息粒度	(54)
5.3.2	概念的近似	(55)
5.3.3	决策表中利用条件知识对决策近似	(56)
5.3.4	由近似空间导出的拓扑空间	(56)
5.4	应用	(57)
5.4.1	信息系统中知识的粗糙性	(58)
5.4.2	信息系统中属性的重要性	(58)
5.4.3	基于粗糙熵的信息系统知识约简算法	(60)
5.4.4	知识依赖	(62)
5.4.5	信息系统中的概念粗糙性	(64)
5.4.6	决策表的熵	(66)
5.4.7	决策表中的属性重要性	(67)
5.4.8	基于粗糙熵的决策表知识约简算法	(67)
6	完备、非协调信息系统的约简与决策规则优化	(70)
6.1	非协调决策系统中关于对象各类知识约简之间的关系	(70)
6.2	非协调决策信息系统中关于系统的各类约简之间的关系	(76)
6.3	约简与最优决策规则提取	(78)
6.3.1	优化一个可能规则	(80)
6.3.2	优化一个近似规则	(81)
6.3.3	优化一个广义决策规则	(82)
6.3.4	优化一个 μ -决策规则	(84)
6.3.5	优化一个 μ -规则	(86)
7	不完备信息系统与决策表的知识获取	(87)
7.1	不完备信息系统	(87)
7.2	极大相容块及其性质	(89)
7.3	集合的近似	(92)
7.4	不完备信息系统与决策表中的知识约简	(95)
7.4.1	决策表和决策规则	(95)
7.4.2	不完备信息系统的约简	(97)
7.4.3	不完备决策表的约简	(101)
8	信息系统的同态不变特征	(105)
8.1	信息系统同态	(105)

8.2 几个引理	(106)
8.3 主要结论	(108)
8.4 一个例子	(111)
参考文献	(114)

1 信息系统、粗糙集与粒度计算的基本概念

信息系统这个概念作为智能信息处理领域的一个具有形式化定义和明确语义的专业术语,出现于波兰学者 Z. Pawlak^[1~3]提出的粗糙集理论中. 数据库模型、数据的决策与分析、模式识别、机器学习等领域的许多问题的描述框架可以归结为信息系统. 目前,粗糙集不但在数学理论上得到不断完善,而且在许多领域得到成功的应用^[4~8],如在机器学习、决策分析、近似推理、图像处理、医疗诊断、金融数据分析、专家系统和知识发现等领域. 本章主要介绍信息系统、粗糙集理论与粒度计算的基本概念,并作为后面各章节的基础.

1.1 信息系统

形式上,四元组 $S = (U, A, V, f)$ 是一个信息系统 (information system), 其中 U : 对象的非空有限集合, 称为论域;

A : 属性的非空有限集合;

$V = \bigcup_{a \in A} V_a$, V_a 表示属性 a 的值域;

$f: U \times A \rightarrow V$ 是一个信息函数, 即对 $\forall x \in U, a \in A$, 有 $f(x, a) \in V_a$.

通常 $S = (U, A, V, f)$ 也简记为 $S = (U, A)$.

信息系统的数据以关系表的形式表示. 关系表的行对应要研究的对象, 列对应对象的属性, 对象的信息是通过指定对象的各属性值来表达的.

容易看出, 一个属性对应一个等价关系, 一个表可以看作是一族等价关系.

令 $P \subseteq A$, 定义属性集 P 的不可区分关系 (indiscernibility) $IND(P)$ 为

$$IND(P) = \{(x, y) \in U \times U \mid \forall a \in P, f(x, a) = f(y, a)\} \quad (1.1)$$

如果 $(x, y) \in IND(P)$, 则称 x 和 y 是 P 不可区分的. 容易证明 $\forall P \subseteq A$, 不可区分关系 $IND(P)$ 是 U 上的等价关系, 符号 $U/IND(P)$ 表示不可区分关系 $IND(P)$ 在 U 上导出的划分, 也称为信息粒度, $IND(P)$ 中的等价类称为 P 基本集. 符号 $[x]_P$ 表示包含 $x \in U$ 的 P 等价类.

在不产生混淆的情况下, 也用 P 来代替 $IND(P)$, $U/IND(P)$ 可简记为 U/P .

令 $P, Q \subseteq A$, $P \leq Q$ (或 $U/IND(P) \leq U/IND(Q)$) 表示对任意 $X \in U/IND(P)$, 存在 $Y \in U/IND(Q)$, 使得 $X \subseteq Y$. 这意味着 P 的划分比 Q 更精细, 或者 Q 的划分比 P 更粗糙. $P < Q$ (或 $U/IND(P) < U/IND(Q)$) 表示对于任意 $X \in U/IND(P)$, 存在 $Y \in U/IND(Q)$, 使得 $X \subseteq Y$, 且存在 $X_0 \in IND(P)$, $Y_0 \in U/IND(Q)$, 使得

$$X_0 \subset Y_0.$$

下面讨论两个信息系统之间的关系.

令 $S_1 = (U, A_1)$ 和 $S_2 = (U, A_2)$ 为两个信息系统. 若 $\text{IND}(A_1) = \text{IND}(A_2)$, 即 $U/A_1 = U/A_2$, 则称 S_1 和 S_2 (或 A_1 和 A_2) 是等价的, 记作 $S_1 \simeq S_2$ (或 $A_1 \simeq A_2$). 这个概念意味着可以用不同的属性集对对象进行描述, 以表达关于论域的完全相同的事实.

对于 $S_1 = (U, A_1)$ 和 $S_2 = (U, A_2)$ 两个信息系统, 当 $\text{IND}(A_1) \subseteq \text{IND}(A_2)$ 时, 称信息系统 S_1 (或属性集 A_1) 比信息系统 S_2 (或属性集 A_2) 更精细, 或者称 S_2 比 S_1 更粗糙. 当 S_1 比 S_2 更精细时, 也称 S_1 为 S_2 的特化, S_2 为 S_1 的推广.

信息系统 $S = (U, A)$ 也称为知识表达系统或知识 A .

令

$$U/A = \{R_1, R_2, \dots, R_m\}$$

当然, 有趣的是最小划分

$$\hat{A}(U) = \{\{x\} \mid x \in U\}$$

和最大划分

$$\bar{A}(U) = \{U\}$$

如果论域不会引起混淆, 也可以将 $\hat{A}(U)$ 和 $\bar{A}(U)$ 记为 $\hat{A}$ 和 $\bar{A}$.

1.2 近似与粗糙集

令 $X \subseteq U, P \subseteq A$. 当 X 能表达成某些 P 基本集的并时, 称 X 是 P 可定义的; 否则称 X 为 P 不可定义的.

P 可定义集是论域的子集, 它可在信息系统 S 中精确地定义, 而 P 不可定义集不能在这个信息系统中精确定义. P 可定义集也称作 P 精确集, 而 P 不可定义集也称为 P 非精确集或 P 粗糙集 (rough set).

当存在等价关系 $P \subseteq A$ 且 $X \subseteq U$ 为 P 精确集时, 集合 X 称为 S 中的精确集; 当对于任何 $P \subseteq A, X \subseteq U$ 都为 P 粗糙集, 则 X 称为 S 中的粗糙集.

粗糙集可以被两个精确集近似地刻画, 即通过粗糙集的上近似 (upper approximation) 和下近似 (lower approximation) 来描述.

给定信息系统 $S = (U, A)$, 对于每个子集 $X \subseteq U$ 和 $P \subseteq A$, 定义两个子集

$$\underline{P}X = \bigcup \{Y \in U/P \mid Y \subseteq X\} \quad (1.2)$$

和

$$\bar{P}X = \bigcup \{Y \in U/P \mid Y \cap X \neq \emptyset\} \quad (1.3)$$

分别称它们为 X 的 P 下近似集和 P 上近似集.

下近似集、上近似集也可用下面的等式来表达:

$$\underline{PX} = \{x \in U \mid [x]_P \subseteq X\} \quad (1.4)$$

$$\bar{PX} = \{x \in U \mid [x]_P \cap X \neq \emptyset\} \quad (1.5)$$

集合 $\text{bn}_P(X) = \bar{PX} - \underline{PX}$ 称为 X 的 P 边界域; $\text{pos}_P(X) = \underline{PX}$ 称为 X 的 P 正域; $\text{neg}_P(X) = U - \bar{PX}$ 称为 X 的 P 负域. 显然, $\bar{PX} = \text{pos}_P(X) \cup \text{bn}_P(X)$.

$\underline{PX}$ 或 $\text{pos}_P(X)$ 是由那些根据属性集 P 判断肯定属于 X 的 U 中元素组成的集合; $\bar{PX}$ 是那些根据属性集 P 判断可能属于 X 的 U 中元素组成的集合; $\text{bn}_P(X)$ 是那些根据属性集 P 既不能判断肯定属于 X 又不能判断肯定属于 $\sim X$ (即 $U - X$) 的 U 中元素组成的集合; $\text{neg}_P(X)$ 是那些根据属性集 P 判断肯定不属于 X 的 U 中元素组成的集合.

下列性质是显而易见的.

定理 1.1 (1) X 为 P 可定义集当且仅当 $\underline{PX} = \bar{PX}$.

(2) X 为 P 粗糙集当且仅当 $\underline{PX} \neq \bar{PX}$.

我们也可将 $\underline{PX}$ 描述为包含在 X 中的最大可定义集, 将 $\bar{PX}$ 描述为包含 X 的最小可定义集.

从近似的定义, 可以直接得到 P 下近似集和上近似集的下列性质.

定理 1.2 (1) $\underline{PX} \subseteq X \subseteq \bar{PX}$.

(2) $\underline{P}\emptyset = \bar{P}\emptyset = \emptyset, \underline{P}U = \bar{P}U = U$.

(3) $\bar{P}(X \cup Y) = \bar{PX} \cup \bar{PY}$.

(4) $\underline{P}(X \cap Y) = \underline{PX} \cap \underline{PY}$.

(5) $X \subseteq Y \Rightarrow \underline{PX} \subseteq \underline{PY}$.

(6) $X \subseteq Y \Rightarrow \bar{PX} \subseteq \bar{PY}$.

(7) $\underline{P}(X \cup Y) \supseteq \underline{PX} \cup \underline{PY}$.

(8) $\bar{P}(X \cap Y) \subseteq \bar{PX} \cap \bar{PY}$.

(9) $\underline{P}(\sim X) = \sim \bar{PX}$.

(10) $\bar{P}(\sim X) = \sim \underline{PX}$.

(11) $\underline{P}(\underline{PX}) = \bar{P}(\bar{PX}) = \underline{PX}$.

(12) $\bar{P}(\bar{PX}) = \underline{P}(\underline{PX}) = \bar{PX}$.

集合的不精确性是由于边界域的存在而引起的. 集合的边界域越大, 其精确度则越低. 为了更准确地表达这一点, 我们引入精度 (accuracy measure) 的概念. 定义等价关系 P 刻画集合 X 的近似精度为

$$\alpha_P(X) = \frac{|\underline{PX}|}{|\bar{PX}|} \quad (1.6)$$

其中 $X \neq \emptyset, |X|$ 表示集合 X 的基数.

显然, 对每一个 P 和 $X \subseteq U$, 有 $0 \leq \alpha_P(X) \leq 1$. 当 $\alpha_P(X) = 1$ 时, X 的 P 边界域为空集, 集合 X 为 P 可定义的; 当 $\alpha_P(X) < 1$ 时, 集合 X 有非空 P 边界域, 集合

X 为 P 不可定义的.

当然,其他一些量度同样可以用来定义集合 X 的不精确程度.

例如,可用 $\alpha_P(X)$ 的一种变形,即 X 的 P 粗糙度 $\rho_P(X)$ 来定义

$$\rho_P(X) = 1 - \alpha_P(X) \quad (1.7)$$

下面定义四种不同类型的粗糙集:

(1) 如果 $\underline{P}X \neq \emptyset$ 且 $\overline{P}X \neq U$, 则称 X 为 P 粗糙可定义.

(2) 如果 $\underline{P}X = \emptyset$ 且 $\overline{P}X \neq U$, 则称 X 为 P 内不可定义.

(3) 如果 $\underline{P}X \neq \emptyset$ 且 $\overline{P}X = U$, 则称 X 为 P 外不可定义.

(4) 如果 $\underline{P}X = \emptyset$ 且 $\overline{P}X = U$, 则称 X 为 P 全不可定义.

这种对粗糙集划分的直观意义如下:

如果集合 X 为 P 粗糙可定义, 则意味着可以确定 U 中某些元素属于 X 或 $\sim X$.

如果集合 X 为 P 内不可定义, 则意味着可以确定 U 中某些元素是否属于 $\sim X$, 但不能确定 U 中的任一元素是否属于 X .

如果集合 X 为 P 外不可定义, 则意味着可以确定 U 中某些元素是否属于 X , 但不能确定 U 中的任一元素是否属于 $\sim X$.

如果集合 X 为 P 全不可定义, 则意味着不能确定 U 中任一元素是否属于 X 或 $\sim X$.

下面给出集合拓扑划分的一个有用性质.

定理 1.3 (1) 集合 X 为 P 粗糙可定义(或 P 全不可定义), 当且仅当 $\sim X$ 为 P 粗糙可定义(或 P 全不可定义).

(2) 集合 X 为 P 外(内)不可定义, 当且仅当 $\sim X$ 为 P 内(外)不可定义.

粗糙集的数字特征和粗糙集的拓扑特征之间存在一种关系. 首先, 如果集合为内不可定义或全不可定义, 则其精度为 0; 其次, 当集合为外不可定义或全不可定义, 则它的补集的精度为 0. 这样即使知道了集合的精度, 也不能确定它的拓扑结构; 反过来, 集合的拓扑结构也不具有精度的信息. 因此在粗糙集的实际应用中, 需要将边界域的两种信息结合起来, 既要考虑精度因素, 又要考虑到集合的拓扑结构.

下面讨论近似分类问题.

令 $\mathcal{F} = \{X_1, X_2, \dots, X_n\}$ 是 U 的一个分类或划分, 这个分类独立于属性集(或等价关系) P . $\mathcal{F}$ 的 P 下近似和上近似分别定义为

$$\underline{P}\mathcal{F} = \{\underline{P}X_1, \underline{P}X_2, \dots, \underline{P}X_n\}$$

$$\overline{P}\mathcal{F} = \{\overline{P}X_1, \overline{P}X_2, \dots, \overline{P}X_n\}$$

定义两个量度来描述近似分类的不精确性.

第一个量度为根据 $P, \mathcal{F}$ 的近似分类精度

$$\alpha_P(\mathcal{F}) = \frac{\sum_{i=1}^n |\underline{P}X_i|}{\sum_{i=1}^n |\overline{P}X_i|} \quad (1.8)$$

另一个量度为根据 $P, \mathcal{F}$ 的近似分类质量

$$\gamma_P(\mathcal{F}) = \frac{\sum_{i=1}^n |\underline{P}X_i|}{|U|} \quad (1.9)$$

近似分类精度描述的是当使用属性集 P 对对象分类时,可能的决策中正确决策的百分比;分类的质量表示的是应用属性集 P 能确切地划入 $\mathcal{F}$ 类对象的百分比.

1.3 属性约简

属性约简是粗糙集理论的核心内容之一. 众所周知,信息系统中的属性并不是同等重要的,甚至其中某些属性是冗余的. 所谓属性约简,就是在保持信息系统分类能力不变的条件下,删除其中不相关或不重要的属性.

属性约简中有两个基本概念:约简(reduct)和核(core). 在对约简和核进行讨论之前,先给出如下定义.

令 A 为一属性集, $a \in A$, 如果

$$\text{IND}(A) = \text{IND}(A - \{a\}) \quad (1.10)$$

则称 a 为 A 中不必要的; 否则称 a 为 A 中必要的.

如果每一个 $a \in A$ 都为 A 中必要的, 则称 A 为独立的; 否则称 A 为依赖的.

定理 1.4 如果 A 是独立的, $P \subseteq A$, 则 P 也是独立的.

证明 用反证法. 假设 $P \subseteq A$ 且 P 是依赖的, 则存在 $Q \subset P$, 使得 $\text{IND}(Q) = \text{IND}(P)$. 这意味着 $\text{IND}(Q \cup (A - P)) = \text{IND}(A)$, $Q \cup (A - P) \subset A$. 因此 A 为依赖的, 与假设矛盾, 故定理得证. 证毕.

设 $Q \subseteq P$, 如果 Q 是独立的, 且 $\text{IND}(Q) = \text{IND}(P)$, 则称 Q 为 P 的一个约简. 显然, P 可以有多个约简. P 中所有必要的属性组成的集合称为 P 的核, 记作 $\text{core}(P)$.

核与约简有如下关系.

定理 1.5 $\text{core}(P) = \bigcap_{Q \in \text{red}(P)} Q$, 其中 $\text{red}(P)$ 表示 P 的所有约简的集合, 且 $\bigcap_{Q \in \text{red}(P)} Q = \text{core}(P)$.

证明 (1) $\text{core}(P) \subseteq \bigcap_{Q \in \text{red}(P)} Q$. 设 $a \in \text{core}(P)$, 我们需要证明 $a \in \bigcap_{Q \in \text{red}(P)} Q$. 若 $a \notin \bigcap_{Q \in \text{red}(P)} Q$, 则存在某个 $Q \in \text{red}(P)$ 使得 $a \notin Q$. 从而有 $Q \subseteq P - \{a\}$, 且

$$\text{IND}(Q) \supseteq \text{IND}(P - \{a\}) \supseteq \text{IND}(P).$$

另一方面, $\text{IND}(P) = \text{IND}(Q)$, 因为 Q 是 P 的一个约简, 因而有 $\text{IND}(P - \{a\}) = \text{IND}(P)$, 这说明 a 在 P 中是不必要的, 即 $a \notin \text{core}(P)$, 矛盾.

(2) $\text{core}(P) \supseteq \bigcap \text{red}(P)$. 设 $a \in \bigcap \text{red}(P)$, 我们需要证明 $a \in \text{core}(P)$. 若 $a \notin \text{core}(P)$, 则 a 在 P 中是不必要的, 于是有 $\text{IND}(P) = \text{IND}(P - \{a\})$. 另一方面, 因为约简总是存在的, 我们知道 $P - \{a\}$ 有一个约简 P_0 . 从 $P_0 \subseteq P - \{a\}$ 知 $a \notin P_0$. $P - \{a\}$ 的约简 P_0 满足:

$$(i) \text{IND}(P_0) = \text{IND}(P - \{a\}).$$

$$(ii) \text{如果 } P' \subset P_0, \text{ 则 } \text{IND}(P - \{a\}) \subset \text{IND}(P').$$

这两个条件可以重写为

$$(i') \text{IND}(P_0) = \text{IND}(P).$$

$$(ii') \text{如果 } P' \subset P_0, \text{ 则 } \text{IND}(P) \subset \text{IND}(P').$$

注意到 $P_0 \subseteq P - \{a\} \subset P$. 我们知道 P_0 是 P 的一个约简, 因此存在某个 $Q \in \text{red}(P)$, 使得 $P_0 = Q$. 因此 $a \notin Q$, 从而 $a \notin \bigcap \text{red}(P)$, 矛盾. 证毕.

可以看出, 核这个概念的用处有两个方面: 可以作为所有约简的计算基础, 因为核包含在所有的约简之中, 并且计算可以直接进行; 可解释为在属性约简时, 它是不能消去的属性集合.

设 $S = (U, A)$ 为一个信息系统, 其中 $A = C \cup D$, 且 $C \cap D = \emptyset$, C 称为条件属性集, D 称为决策属性集. 具有条件属性和决策属性的信息系统称为决策表.

在应用中, 一个分类相对于另一个分类的关系十分重要, 因此我们将介绍属性的相对约简(relative reduct)和相对核(relative core)的概念. 首先定义一个分类相对于另一个分类的正域.

令 $P, Q \subseteq A$, Q 的 A 正域记为 $\text{pos}_P(Q)$, 即

$$\text{pos}_P(Q) = \bigcup_{x \in U/Q} \underline{P}x \quad (1.11)$$

Q 的 P 正域是 U 中所有根据分类 U/P 的信息, 可以准确地划分到 Q 的等价类中去的对象集合.

令 $P, Q \subseteq A, a \in P$, 如果

$$\text{pos}_P(Q) = \text{pos}_{P - \{a\}}(Q)$$

则称 a 为 P 中 Q 不必要的; 否则 a 为 P 中 Q 必要的.

如果 P 中的每个 a 都为 Q 必要的, 则称 P 为 Q 独立的 (或 P 相对于 Q 独立).

设 $R \subseteq P$, R 为 P 的 Q 约简当且仅当 R 是 P 的 Q 独立子集, 且 $\text{pos}_R(Q) = \text{pos}_P(Q)$. P 的 Q 约简常简称为相对约简.

P 中所有 Q 必要的属性构成的集合称为 P 的 Q 核, 简称为相对核, 记为

$\text{core}_Q(P)$.

相对核与相对约简的关系由下述定理来刻画.

定理 1.6 $\text{core}_Q(P) = \bigcap \text{red}_Q(P)$, 其中 $\text{red}_Q(P)$ 表示所有 P 的 Q 约简构成的集合.

定理 1.6 的证明与定理 1.5 的证明类似.

注: 对于决策表 $S = (U, A)$, 其中 $A = C \cup D$, 且 $C \cap D = \emptyset$, C 称为条件属性集, D 称为决策属性集, C 的 D 约简称为决策表的相对约简, C 的 D 核称为决策表的相对核.

1.4 属性依赖性

属性的依赖性可形式化地定义如下: 令 $S = (U, A)$ 是一个信息系统, $P, Q \subseteq A$.

- (1) 属性集 Q 依赖于属性集 P (记作 $P \Rightarrow Q$), 当且仅当 $\text{IND}(P) \subseteq \text{IND}(Q)$.
- (2) 属性集 P 与属性集 Q 等价 (记作 $P \equiv Q$), 当且仅当 $P \Rightarrow Q$ 且 $Q \Rightarrow P$.
- (3) 属性集 P 与属性集 Q 独立 (记作 $P \not\Rightarrow Q$), 当且仅当 $P \Rightarrow Q$ 与 $Q \Rightarrow P$ 均不成立.

显然, $P \equiv Q$ 当且仅当 $\text{IND}(P) = \text{IND}(Q)$.

当属性集 Q 依赖于属性集 P 时, 也说属性集 Q 是由属性集 P 导出的.

通过简单推导可得下列性质.

定理 1.7 下列条件是等价的:

- (1) $P \Rightarrow Q$.
- (2) $\text{IND}(P \cup Q) = \text{IND}(P)$.
- (3) $\text{POS}_P(Q) = U$.
- (4) 对于所有 $X \in U/Q$, 有 $\underline{P}X = X$.

下面给出属性依赖性的一些重要性质.

定理 1.8 (1) 如果 $P \Rightarrow Q$ 且 $Q \Rightarrow R$, 则 $P \Rightarrow R$.

(2) 如果 $P \Rightarrow Q$ 且 $Q \Rightarrow R$, 则 $P \cup Q \Rightarrow R$.

(3) 如果 $P \Rightarrow R \cup Q$, 则 $P \Rightarrow R$ 且 $P \Rightarrow Q$.

(4) 如果 $P \Rightarrow Q$ 且 $Q \cup R \Rightarrow T$, 则 $P \cup R \Rightarrow T$.

(5) 如果 $P \Rightarrow Q$ 且 $R \Rightarrow T$, 则 $P \cup R \Rightarrow Q \cup T$.

(6) 如果 $P \Rightarrow Q$ 且 $P' \supset P$, 则 $P' \Rightarrow Q$.

(7) 如果 $P \Rightarrow Q$ 且 $Q' \subset Q$, 则 $P \Rightarrow Q'$.

有时候属性的依赖性可能是部分的, 这意味着属性集 Q 仅有部分是由属性集 P 导出的, 部分导出可由属性集的正域来定义.

令 $S = (U, A)$ 是一个信息系统, $P, Q \subseteq A$. 当

$$k = \gamma_P(Q) = |\text{POS}_P(Q)| / |U| \quad (1.12)$$

时,称属性集 Q 是 $k(0 \leq k \leq 1)$ 度依赖于属性集 P 的,记作 $P \Rightarrow_k Q$.

当 $k=1$ 时,称 Q 完全依赖于 P ; 当 $0 < k < 1$ 时,称 Q 粗糙(部分)依赖于 P ; 当 $k=0$ 时,称 Q 完全独立于 P . 如果 $P \Rightarrow_1 Q$, 也记为 $P \Rightarrow Q$.

由依赖性的定义可知,当 $P \Rightarrow_k Q$ 时,由 Q 导出的分类 U/Q 的正域覆盖了信息系统的 $k \times 100\%$ 个元素;另一方面,只有属于分类正域的元素能被唯一地分类,即对象的 $k \times 100\%$ 可以通过属性集 P 划入分类 U/Q 的等价类中.

系数 $\gamma_P(Q)$ 可以看作 Q 关于 P 的依赖度.

部分依赖性 $P \Rightarrow_k Q$ 的量度 k 不能够完全反映 U/Q 中类之间的分布情况. 例如,一些决策类可能完全由 P 描述,但另一些可能仅仅由 P 部分描述. 因此,需要使用一个系数 $\gamma_P(X) = |\underline{P}X| / |X| (X \in U/Q)$ 来表明通过属性集 P 能将 U/Q 中每个类的多少个元素正确分类.

这样,两个值 $\gamma_P(Q)$ 和 $\gamma_P(X) (X \in U/Q)$, 给出了属性集 P 的“分类能力”关于分类 U/Q 的全部信息.

1.5 区分矩阵与区分函数

利用区分矩阵(discernibility matrix)来表达属性集有许多优点,特别是它能够容易地计算约简和核^[9].

令 $S = (U, A)$ 是一个信息系统, $|U| = n$. S 的区分矩阵是一个 $n \times n$ 矩阵, 其任一元素为

$$\alpha(x, y) = \{a \in A \mid f(x, a) \neq f(y, a)\} \quad (1.13)$$

因此, $\alpha(x, y)$ 是区别对象 x 和 y 的所有属性的集合.

下面引入一个布尔函数,称其为区分函数(discernibility function),用 Δ 表示. 对每个属性 $a \in A$, 指定一个布尔变量“ a ”. 若 $\alpha(x, y) = \{a_1, a_2, \dots, a_k\} \neq \emptyset$, 则指定一个布尔函数 $a_1 \vee a_2 \vee \dots \vee a_k$, 用 $\Sigma \alpha(x, y)$ 来表示; 若 $\alpha(x, y) = \emptyset$, 则指定布尔常量.

区分(布尔)函数 Δ 可定义如下:

$$\Delta = \prod_{(x, y) \in U \times U} \Sigma \alpha(x, y) \quad (1.14)$$

现在介绍几个与布尔函数有关的概念.

(1) 布尔表达式 E 是一个范式, 如果 E 仅仅由布尔变量和常量通过析取与合取运算来表达.

(2) 布尔表达式 E 是一个合取范式, 如果 E 是由一些析取式组成的合取所构成的范式.