基于粗糙集的知识获取理论与方法

王长忠 陈德刚 著

哈爾濱工業大學出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

基于粗糙集的知识获取理论与方法

王长忠 陈德刚 著

哈爾濱工業大學出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

内 容 简 介

粗糙集理论是一种不确定性数据分析工具，近年来在特征选择、分类学习和规则提取等应用方面取得了巨大成功，并逐渐成为信息科学最为活跃的研究领域之一。本书系统地论述了基于粗糙集的属性约简理论和方法，信息系统之间的信息通讯以及关于粗糙群的一些本质问题。主要内容有：基于覆盖粗糙集的属性约简模型建立及约简结构性质分析，基于广义粗糙集的属性约简模型建立及约简结构性质分析，基于广义粗糙集的信息系统之间等价属性约简理论与方法，基于模糊粗糙集的信息系统之间等价属性约简理论与方法，粗糙群的理论研究，以及这些理论在数据挖掘中的应用。

本书可以作为理工科大学计算机、应用数学、自动控制、信息科学以及管理工程等专业的高年级本科生、研究生以及博士生的教材，同时对有关领域的研究人员和工程技术人员也有重要的参考价值。

图书在版编目(CIP)数据

基于粗糙集的知识获取理论与方法/王长忠，陈德刚著.
—哈尔滨：哈尔滨工业大学出版社，2010.3
ISBN 978-7-5603-2877-5

Ⅰ.基… Ⅱ.①王…②陈 Ⅲ.①粗糙集-研究 Ⅳ.①O114

中国版本图书馆 CIP 数据核字(2010)第 043343 号

责任编辑 赵文斌
封面设计 卞秉利
出版发行 哈尔滨工业大学出版社
社　　址 哈尔滨市南岗区复华四道街 10 号 邮编 150006
传　　真 0451-86414749
网　　址 http://hitpress.hit.edu.cn
印　　刷 哈尔滨市石桥印务有限公司
开　　本 850mm×1168mm 1/32 印张 6 字数 130 千字
版　　次 2010 年 3 月第 1 版 2010 年 3 月第 1 次印刷
书　　号 ISBN 978-7-5603-2877-5
定　　价 24.80 元

前　言

粗糙集理论(Rough Set Theory)是一种刻画不完整性和不确定性的数学工具,能够有效地处理不确定和不精确的信息.该理论最初是由波兰学者 Pawlak Z 于 1982 年提出,此后,引起了数学领域、工程技术领域和计算机领域中研究人员的广泛关注.他们在粗糙集理论和应用方面作了大量的研究工作.从粗糙集的提出到与其他不确定性理论的结合,从理论研究到实际应用,粗糙集理论已遍地开花,并逐渐成为当今国际信息科学领域中的研究热点之一.

粗糙集理论的研究内容十分丰富,例如,粗糙计算方法、粗糙集的各种模型推广、粗糙集与其他不确定性理论的联系、粗糙集代数与粗糙逻辑、各种粗糙模型的实际应用等等,都是粗糙集理论中重要的研究领域.其中一个重要应用领域是信息系统中的属性约简.属性约简目的是从复杂数据中去掉冗余信息从而挖掘潜在有用的信息.近年来,基于粗糙集的属性约简理论和算法设计研究都取得了突破性进展,开始逐渐成为克服"冗余信息"而进行数据约简的有力手段.

迄今为止,国内外关于粗糙集理论及其应用的学术书刊很多.但专门从属性约简、信息系统通讯以及群的角度进行研究的不是很多.本书系统地论述了基于粗糙集的属性约简理论和方法,信息系统之间的信息通讯以及关于粗糙群的一些本质问题.本书的特色之处在于建立系统的属性约简理论体系和信息系统通讯模型,然后在此基础上研究和分析属性约简的结构性质和实现不同信息系统的等价快速属性约简的方法.这些理论和方法的形成和发展无疑极大丰富了粗糙集理论的内容.

本书的目的是介绍作者在粗糙集理论与方法的研究中的一点体会和心得，介绍我们在该领域研究中所取得的最新成果.虽然我们对粗糙集理论的研究还很肤浅，但是愿意抛砖引玉，期望为从事该理论研究的学者及研究生进入这一领域提供捷径.

借本书出版之际，我们要特别感谢哈尔滨工业大学吴从炘教授，没有他的指导，就不会有此书的出版，在此向吴从炘教授表示崇高的敬意.同时特别感谢王熙教授、吴伟志教授、米据生教授、梁吉业教授、李德玉教授、魏玲教授、邵明文教授、李同军教授、胡清华副教授、钱宇华博士、马建敏博士等，本书的许多思想，得益于与他们的讨论.

本书得到了辽宁省重点学科建设项目和国家自然科学基金项目(批准号:70871036 和 60703013)的支持，在此表示感谢.

鉴于我们的知识水平有限，书中难免有不妥或错误之处，真心地希望读者，特别是有关专家批评指正.

作　者

2010 年 3 月

目　录

第1章　绪　论

1.1　粗糙集理论的产生背景

在日常生活、生产实践中，人们常常需要面对和处理数量庞大而且结构复杂的数据.比如，金融分析、医疗诊断、管理决策以及大型设备故障诊断等领域的信息系统中保存了大量类型多样、结构复杂的数据.这些数据主要有以下特点.

(1)数据类型种类繁多.数据类型可简单地分为离散型、连续型、集值型等.离散型又分为名义型、有序型和整数型;集值型包括区间值型和模糊集型等.

(2)不同类型的数据蕴含了不同的信息结构.名义型数据之间蕴含着等价关系;有序型数据之间不仅存在等价关系，而且还存在序的结构信息;数值型数据则不仅存在等价关系与序关系，还有邻域关系;集值型和模糊型数据不仅包含邻域关系而且包含模糊信息等.

(3)在生产实践中往往是多种类型数据混杂在一起，而且还有一些数据是缺省的.

如何利用已掌握的数据，从数据中挖掘潜在有用的信息，迅速地做出正确的决策成为人们关心的核心问题.尤其是在讲究效率和节约成本的今天，从复杂数据中深入地研究数据挖掘理论并且设计出高效的算法成为当今计算机领域、数学领域和工程应用领域的研究人员面临的迫切需要解决的课题.

粗糙集(Rough Set)理论是由波兰数学家 Pawlak Z[1]于 1982 年提出的一种不确定性数据分析理论.该理论是经典集合论的一种推广,能够有效地对数据中潜在有用的信息进行挖掘[1-3].其主要思想就是在保持信息系统分类能力不变的前提下,通过知识约简剔除数据中冗余的信息,从而导出问题的正确决策或分类.

粗糙集理论的最初研究主要局限在东欧一些国家.到了 20 世纪 90 年代,由于该理论在人工智能领域的成功应用,才吸引了来自人工智能领域、数学领域以及工程应用领域的研究人员的广泛关注[3].经过二十多年的发展,粗糙集理论在特征选择[4-13]、分类学习[14-18]和规则提取[19-24]等方面取得了巨大进步,并成功地应用于机器学习、故障诊断、分析决策、过程控制以及关系数据库中的知识获取等人工智能领域,逐渐成为信息科学最为活跃的研究领域之一.

粗糙集理论作为一种不确定性数据分析理论,处理分析数据的一个显著的特点是不需要用户提供数据之外的任何先验知识,比如统计学中的概率分布,模糊集理论中的隶属函数.所以对问题的不确定性的描述和处理比较客观[25].然而由于该理论不包含处理不完备或实型数据的机制,不能够有效地描述和处理具有连续型和集值型数据,因此粗糙集理论与概率论、模糊集和证据理论等其他不确定性理论有很强的互补性[3,26].

粗糙集、模糊集和随机集都是处理不确定性和不精确性问题的集合理论.然而它们处理问题的侧重面不同,模糊集是基于元素对集合隶属程度的不同,强调的是集合边界的病态定义,即边界的不分明性.而粗糙集是以不可区分关系为基础,强调的是集合对象间的不可分辨性.随机集是以概率测度为工具,侧重的是事件的随机性[25,26].正因为这三个理论的特点不同,它们具有很强的互补性.许多学者将粗糙集与模糊集两者结合起来研究,于是产生了粗

糙模糊集与模糊粗糙集[27-31].而随机集与粗糙集相结合就产生了基于随机集的粗糙集模型[32-36].目前,对粗糙集理论的研究主要集中在包括粗糙代数与粗糙拓扑的数学性质,粗糙集模型的推广,粗糙集模型中的度量,与其他数学理论的关系与互补,基于粗糙集的逻辑,属性约简等方面.这些理论的形成和发展无疑极大地丰富了粗糙集理论的内容.

1.2 粗糙集和模糊粗糙集理论研究概况

粗糙集理论将研究对象的全体称为论域,利用等价关系将论域划分为若干互斥的等价类,作为描述论域中任意集合的基本信息粒子.对于空间中的任意概念,粗糙集理论是用两个可定义的,分别叫做下近似集合和上近似集合,来近似表达这一概念.这一方法自然地模拟了人类的学习和推理过程,学习得到的知识采用产生式规则表示,容易被用户理解、接受和使用,因此得到了广泛的重视[37-39].粗糙集理论的另一个重要应用是数据库中的属性约简.对于给定的由离散属性刻画的数据库,我们可以找出一个具有相同信息量的属性子集来刻画数据表.根据这种思想就产生了属性约简的概念.属性约简可以看做是粗糙集理论与其他不确定性理论区别的最具有特征的概念之一.

然而,Pawlak 粗糙集理论是建立在等价关系基础之上的,只适合于处理离散型变量,对于实值连续型数据却不能直接处理.在现实应用中,广泛存在着连续型数据,如股票分析中的波动价格[42-46],金融分析中的大量数据[47,48],故障诊断分析中的温度、电流、电压信号[49-58].经典粗糙集理论在处理这类数据时,往往先采用离散化算法把数值型属性转化为符号型属性作为预处理[59-63].这种数据类型转换的最大缺点是会带来大量的信息损

失,从而影响规则提取的准确性[64-66].为了解决这一问题,有必要对经典的粗糙集理论进行推广.正因为如此,Pawlak 粗糙集模型的推广成为粗糙集理论研究的主要方向之一.粗糙集理论推广的方法有两种:即构造性方法和公理化方法.

构造性方法是指从近似空间出发,以论域上的二元关系、邻域系统或布尔子代数作为基本要素,构造性地定义近似算子,建立粗糙集代数系统来研究粗糙集的数学结构和性质[25,26,67].以构造性方法推广的模型有基于一般关系的粗糙集模型[67-73]、基于相似关系的粗糙集模型[74-76]、基于邻域算子的粗糙集模型[77-80]、模糊粗糙集模型与粗糙模糊集模型[81-85]、α-粗糙集模型[86,87]、基于优势关系的粗糙集模型[88-90]、不完备信息系统下的粗糙集模型[91-93]、变精度粗糙集模型[94,95]以及概率粗糙集模型[3,34-36]等.构造性方法直观、容易理解,但不易深刻了解近似算子的数学性质.

公理化方法是指以满足某些公理的集合算子为基本要素,然后建立近似空间,使得由该近似空间导出的近似算子恰好就是由给定公理定义的近似算子[25,26,67].公理化方法的研究最初是由 Lin 和 Liu[96]提出的,他们首先研究了 Pawlak 粗糙集公理化的代数系统,后来逐渐发展到一般关系下的粗糙集代数系统[70,96,97]的研究.Yao[67]于 1998 年详细地研究了基于一般关系的粗糙集公理化的性质.在模糊粗糙集方面,Morsi 和 Yakout[98]首先研究了由 T-相似关系导出的模糊近似算子的公理化刻画.Wu[26,99]等利用截集去模糊化方法研究了一般模糊近似算子的公理化特征.Radzikowska 与 Kerre[100]讨论了由各种蕴含算子所定义的粗糙近似算子的性质与公理刻画,Mi[25,101,102]用三角模和蕴含算子建立了基于一般模糊关系的粗糙集模型,并对各类广义近似算子给出了公理化刻画.2005 年 Yeung 和 Chen[103]系统地研究了广义模糊粗糙集模型,

并用不同的公理集合刻画了广义模糊近似算子.公理化方法旨在研究粗糙集合与模糊粗糙集的数学特征而不在于应用方法.

基于不同的粗糙集模型,在实际生活中有不同的应用.例如,利用 Pawlak 粗糙集模型可以挖掘符号型变量的分类信息[104,105],邻域算子的粗糙集模型可以分析数据空间中的邻域拓扑信息[106,107],模糊粗糙集模型可以处理数据中的模糊信息[108-110]等.

为了深入理解从而更好利用各种粗糙集模型进行数据挖掘,研究者对粗糙集理论的数学性质进行了深入的研究,主要包括粗糙集的代数结构[111-137]、拓扑结构[106,107,138-142]、粗糙度量[143-148]以及粗糙集的分析性质[149,150]等,如粗糙半群[114,121,122,135]、粗糙群[113,117,120,123]、粗糙环[127-130,133]、粗糙拓扑[138-142]、粗糙积分和粗糙微分[149,150]等.随着研究的深入,必将不断涌现出新的富有生机的数学分支.

下面就本书相关的内容进行简要的综述.

1.2.1 Pawlak 粗糙集理论

Pawlak 粗糙集理论是建立在分类机制的基础之上的,将分类理解为在特定空间上的等价关系,而等价关系构成了对该空间的划分.粗糙集理论将知识理解为对数据的划分,每一个被划分的集合称为概念.粗糙集理论的主要思想是利用已知的知识库,将不精确或不确定的知识用已知的知识库中的知识来刻画.从认识的角度来看,人们需要通过分类去认识那些不能用等价关系分类信息精确表示的对象集[37,38,39].

设 $U=\{x_1,x_2,\cdots,x_n\}$ 是非空有限论域,$R\subseteq U\times U$ 是 U 上的二元等价关系,也称为不可区分关系.则称序对 (U,R) 为 Pawlak 近似空间.对于任意序对 $(x,y)\in U\times U$,若 $(x,y)\in R$,则称 x 和 y 关于 R 是不可区分的,否则,称 x 和 y 关于 R 是可区分的.

等价关系 R 确定论域的一个划分 U/R. 可以证明, U 上的划分可以与 U 上的等价关系之间建立一一对应. U/R 中的集合称为基本集或原子集. 若将 U 中的集合称为概念, 则 (U,R) 称为知识库, 称原子集为基本概念或知识模块. 空集或任意有限个的基本集的并集称为可定义集, 其他的集合称为不可定义的. 可以验证, 可定义集的全体构成 U 上的一个拓扑.

对于任意的 $X \subseteq U$, 不一定能用知识库中的知识模块来精确地描述 X, 即 X 可能为不可定义集. 这时就用 X 关于近似空间 (U,R) 的下近似集合 $\underline{R}(X)$ 和上近似集合 $\overline{R}(X)$ 来近似地描述 X, 其定义如下:

$$\underline{R}(X) = \{x:[x]_R \subseteq X\} = \bigcup \{[x]_R:[x]_R \subseteq X\}$$

$$\overline{R}(X) = \{x:[x]_R \cap X \neq \varnothing\} = \bigcup \{[x]_R:[x]_R \cap X \neq \varnothing\}$$

其中 $[x]_R$ 是 x 的 R 等价类. 我们称 $\underline{R}$ 和 $\overline{R}$ 为 Pawlak 近似算子, 称系统 $(2^U, \cap, \cup, \sim, \underline{R}, \overline{R})$ 为一个 Pawlak 粗糙集代数系统, 其中 $\sim$ 表示集合补.

可见, X 的下近似是包含在 X 中的最大可定义集合, 而 X 的上近似是包含 X 的最小可定义集合. 如果 X 的下近似与上近似相等, 则 X 是可定义的, 否则是粗糙的.

称集合 $\mathrm{Bn}_R(X) = \overline{R}(X) - \underline{R}(X)$ 为 X 的 R 边界域; 称 $\mathrm{Pos}_R(X) = \underline{R}(X)$ 为 X 的 R 正域; 称 $\mathrm{Neg}_R(X) = U - \overline{R}(X)$ 为 X 的 R 负域. 显然, $\overline{R}(X) = \mathrm{Pos}_R(X) \cup \mathrm{Bn}_R(X)$.

下近似 $\underline{R}(X)$ 或 $\mathrm{Pos}_R(X)$ 可以解释为论域中根据 R 判断肯定属于 X 的所有元素组成的集合; 上近似 $\overline{R}(X)$ 可以解释为论域中根据 R 判断可能属于 X 的所有元素组成的集合; 负域 $\mathrm{Neg}_R(X)$ 可以解释为论域中根据 R 判断肯定不属于 X 的所有元素组成的集合; 边界域 $\mathrm{Bn}_R(X)$ 可以解释为根据 R 既不能判断肯定属于 X 又

不能判断肯定不属于 X 的所有元素组成的集合.

根据下近似和上近似概念的定义,我们可以得到下面的定理.

定理 1.1　下近似 $\underline{R}(X)$ 和上近似 $\overline{R}(X)$ 有下列性质:

(1) $\underline{R}(\varnothing) = \overline{R}(\varnothing) = \varnothing$.

(2) $\underline{R}(U) = \overline{R}(U) = U$.

(3) $\underline{R}(X) \subseteq X \subseteq \overline{R}(X)$.

(4) $\underline{R}(X \cap Y) = \overline{R}(X) \cap \overline{R}(Y)$.

(5) $\overline{R}(X \cup Y) = \overline{R}(X) \cup \overline{R}(Y)$.

(6) $X \subseteq Y \Rightarrow \underline{R}(X) \subseteq \underline{R}(Y)$.

(7) $X \subseteq Y \Rightarrow \overline{R}(X) \subseteq \overline{R}(Y)$.

(8) $\underline{R}(X \cup Y) \supseteq \underline{R}(X) \cup \underline{R}(Y)$.

(9) $\overline{R}(X \cap Y) \subseteq \overline{R}(X) \cap \overline{R}(Y)$.

(10) $\underline{R}(\sim X) = \sim \overline{R}(X)$.

(11) $\overline{R}(\sim X) = \sim \underline{R}(X)$.

(12) $\underline{R}(\underline{R}(X)) = \overline{R}(\underline{R}(X)) = \underline{R}(X)$.

(13) $\overline{R}(\overline{R}(X)) = \underline{R}(\overline{R}(X)) = \overline{R}(X)$.

在 Pawlak 提出的粗糙集模型中,对于等价关系的要求只能处理完备的信息系统,因此限制了粗糙集理论的使用范围.为了使用粗糙集理论的思想处理更复杂的实际问题,Pawlak 的粗糙集模型被推广到更一般的情况.

1.2.2　覆盖粗糙集理论

将 Pawlak 粗糙集模型中的等价关系 R 生成的划分 U/R 推广到论域 U 上的覆盖就产生了基于覆盖的粗糙集模型. Zakowski[151] 是第一个将 Pawlak 粗糙集模型中的划分 U/R 推广到论域上的覆盖的学者,并建立了基于覆盖的粗糙集模型.后来的学者对如何有

效地描述论域的覆盖以及如何构造基于覆盖的粗糙近似算子这两个问题展开了研究[106,152-162].论域覆盖定义的有效性直接影响到数据空间的信息粒子的大小,而粗糙算子的定义却影响到从数据库中所提取规则的可靠性.Zhu[158]基于文献[151]中的覆盖粗糙集模型研究了覆盖元素的约简以及上近似算子和下近似算子之间的关系等问题[159-161].而Tsang与Chen[163]给出了覆盖粗糙集模型的实际应用背景并给出了基于覆盖粗糙集模型的属性约简方法.下面我们列出有关文献中关于覆盖粗糙集近似算子的基本内容.

定义 1.1[152] 设 U 是有限论域,C 是 U 的一簇子集.如果 C 中任意的子集 K 都非空且 $\bigcup_{K\in C} K = U$,则 C 称为 U 的一个覆盖.

定义 1.2[152] 设 U 是有限论域,C 是 U 的一簇子集.称序对 (U,C) 为一个覆盖近似空间.

定义 1.3[152] 设 (U,C) 为一个覆盖近似空间,$x \in U$.令

$$\mathrm{Md}(x) = \{K \in C : x \in K \wedge (\forall S \in C \wedge x \in S \wedge S \subseteq K \Rightarrow K = S)\}$$

则称 $\mathrm{Md}(x)$ 为 x 的最小描述.

关于覆盖粗糙集下近似算子的构造,大部分学者的思路是一致的.

定义 1.4[152] 设 (U,C) 为一个覆盖近似空间,$X \subseteq U$,令

$$X_* = \bigcup \{K \in C : K \subseteq X\}$$

则称 X_* 为 X 关于 (U,C) 的下近似.

关于覆盖粗糙集上近似算子的构造,学者们提出了五种定义方法.

Bonikowski[152]等人定义的覆盖粗糙集上近似算子如下:

$$X^* = X_* \cup \{K \in \mathrm{Md}(x) : x \in X - X_*\}$$

Pomykala[155]以及Mordeson[157]等人采用的是如下的方式来

定义覆盖粗糙集上近似算子:

$$\bar{s}(X)=\bigcup\{K\in \boldsymbol{C}:K\cap X\neq\varnothing\}$$

Chen[163] 等人指出以上两种定义方式都有不完善的地方,对于上近似算子 X^* 而言,在有些情况下可能会丢掉有用的信息;对于算子 $\bar{s}(X)$ 而言,计算的结果可能会包含一些不必要的信息.针对这两个算子的缺点,Chen[163] 提出了第三种上近似算子的定义如下:

$$\overline{X}=\bigcup\{K\in \mathrm{Md}(x):x\in X\}$$

根据这个定义,显然有 $X^*\subseteq\overline{X}\subseteq\bar{s}(X)$.

后来,Zhu[106] 等人又提出了两种定义覆盖粗糙集上近似算子的方式:

$$RH(X)=X_*\bigcup\{K\in \boldsymbol{C}:K\cap\{X-X_*\}\neq\varnothing\}$$

$$IH(X)=X_*\bigcup\{\mathrm{Neighbor}(x):x\in X-X_*\}$$

其中 $\mathrm{Neighbor}(x)=\bigcap\{K:x\in K\in \boldsymbol{C}\}$.

本书在第2章将给出描述覆盖结构的另外一种方式.基于这种覆盖结构,重新定义了覆盖粗糙集的上近似算子和下近似算子,以此为基础研究了基于覆盖粗糙集模型的属性约简方法[164].

1.2.3 广义粗糙集理论

将 Pawlak 粗糙集模型中的等价关系推广到任意的二元关系就得到了基于一般关系的粗糙集模型. Yao[67,70-72] 在这方面推广做了大量的工作.

定义1.5 设 U 是论域,$R\subseteq U\times U$ 是 U 上的二元关系.称关系 R 是串行的,如果对于任意 $x\in U$ 存在 $y\in U$ 使得 $(x,y)\in R$.称 R 是自反的,如果对于任意 $x\in U$ 使得 $(x,x)\in R$.称关系 R 是对称的,如果对于任意 $x,y\in U$,$(x,y)\in R$ 蕴涵 $(y,x)\in R$.称关系 R 是传递的,如果对于任意 $x,y,z\in U$,$(x,y)\in R$ 和 $(y,