中国计算机学会学术著作丛书

粗糙集理论、算法与应用

苗夺谦　李道国　著

清华大学出版社

中国计算机学会学术著作丛书

粗糙集理论、算法与应用

Rough Sets Theory, Algorithms and Applications

苗夺谦　李道国　著

清华大学出版社
北京

内 容 简 介

本书主要介绍粗糙集理论、算法与应用。粗糙集理论是关于不精确、不相容、不完备数据处理的数学理论，是经典集合论的重要发展，为真实世界数据的知识表示、学习、归纳和挖掘等方面的研究提供了一种有效的处理技术和方法。由于它无需提供所处理数据之外的任何先验信息，因此在智能信息处理研究中发挥着越来越重要的作用。

本书共分三部分。其中，理论部分简要介绍了经典集合论与模糊集合论的一些相关背景知识、粗糙集理论的一般方法，讨论了粗糙集的代数结构与数学分析性质，初步分析了粗糙集与模糊集的融合；算法部分介绍了现有的知识约简算法，对各种算法的复杂性、完备性作了比较分析；应用部分主要讨论了粗糙集在机器学习(Monk 问题求解)和自然语言处理中的应用研究，如基于粗糙集的词性标注、信息检索、文字识别和文本分类等。

本书适用于高等院校计算机、自动化、信息科学、管理工程和应用数学等专业的师生阅读，尤其是对高年级本科生、硕士生和博士生从事相关研究有所裨益。同时，对相关学科领域的科技工作者和工程技术人员也有一定的参考价值。

图书在版编目(CIP)数据

粗糙集理论、算法与应用/苗夺谦，李道国著. —北京：清华大学出版社，2008.4
(中国计算机学会学术著作丛书)
ISBN 978-7-302-16552-1

Ⅰ. 粗…　Ⅱ. ①苗… ②李…　Ⅲ. 数值计算　Ⅳ. O241

中国版本图书馆 CIP 数据核字(2007)第 184261 号

责任编辑：赵彤伟
责任校对：刘玉霞
责任印制：王秀菊
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京鑫丰华彩印有限公司
装 订 者：三河市李旗庄少明装订厂
经　　销：全国新华书店
开　　本：175×245　**印　张**：21.5　**字　数**：417 千字
版　　次：2008 年 4 月第 1 版　　**印　　次**：2008 年 4 月第1 次印刷
印　　数：1～3000
定　　价：53.00 元

清华大学出版社计算机学术著作出版基金

评审委员会

中国计算机学会学术著作丛书

名誉主任委员：张效祥
主任委员：唐泽圣
副主任委员：陆汝钤
委　　员：（以姓氏笔画为序）
王　珊　吕　建　李晓明
林惠民　罗军舟　郑纬民
施伯乐　焦金生　谭铁牛

序

Preface

第一台电子计算机诞生于20世纪40年代。到目前为止，计算机的发展已远远超出了其创始者的想像。计算机的处理能力越来越强，应用面越来越广，应用领域也从单纯的科学计算渗透到社会生活的方方面面，从工业、国防、医疗、教育、娱乐直至人们的日常生活，计算机的影响可谓无处不在。

计算机之所以能取得上述地位并成为全球最具活力的产业，原因在于其高速的计算能力、庞大的存储能力以及友好灵活的用户界面。而这些新技术及其应用有赖于研究人员多年不懈的努力。学术研究是应用研究的基础，也是技术发展的动力。

自1992年起，清华大学出版社与广西科学技术出版社为促进我国计算机科学技术与产业的发展，推动计算机科技著作的出版，设立了“计算机学术著作出版基金”，并将资助出版的著作列为中国计算机学会的学术著作丛书。时至今日，本套丛书已出版学术专著近50种，产生了很好的社会影响，有的专著具有很高的学术水平，有的则奠定了一类学术研究的基础。中国计算机学会一直将学术著作的出版作为学会的一项主要工作。本届理事会将秉承这一传统，继续大力支持本套丛书的

出版，鼓励科技工作者写出更多的优秀学术著作，多出好书，多出精品，为提高我国的知识创新和技术创新能力，促进计算机科学技术的发展和进步做出更大的贡献。

中国计算机学会

2002 年 6 月 14 日

前言

Foreword

21世纪是信息化社会的时代，计算机与网络信息技术的日新月异，使得人们在各个领域获取的数据和信息急剧膨胀，并且由于人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)。面对如此丰富的可利用数据与信息资源，人类却陷入了乏力于获取知识的困境，因为目前我们仍然缺乏有效的、能够利用计算机及信息技术来帮助人类从海量数据和信息中获取有用的信息和知识的方法。国际数据公司的一份调查报告称，一个大型企业的数据库中，仅有约7%的数据和信息得到了较好的利用。因此，如何处理这些模糊的、不精确的、不完整的海量信息，从中获取潜在的、新颖的、正确的、有利用价值的知识，是对智能信息处理提出的严峻挑战。由此产生了人工智能和信息科学研究的一个崭新领域——智能信息处理。

粗糙集理论与方法能有效地处理复杂系统中的数据和信息，它业已成为一种处理模糊和不精确问题的新型数学工具。它与模糊集方法、证据理论方法和概率方法等其他处理不确定性问题理论的显著区别在于无须提供所处理数据之外的任何先验信息。粗糙集理论与模糊集合论、概率论和证据理论的关系，既非相互冲突，也不可相互替

代，而是具有很强的互补性。

粗糙集理论是波兰科学院 Z. Pawlak 院士于 1982 年提出的一种关于数据分析和推理的理论。最初关于粗糙集理论的研究主要集中在东欧国家，当时并没有引起重视。直到 20 世纪 80 年代末 90 年代初，由于粗糙集理论在数据挖掘、决策分析、模式识别、机器学习和智能控制等方面的成功应用，才得到国际人工智能与模式识别研究学者的广泛关注，并开始注重其应用研究。1991 年，Z. Pawlak 的专著*Rough Sets—Theoretical Aspects of Reasoning about Data* 问世，标志着粗糙集理论与应用的研究进入了活跃时期。1992 年在波兰召开了关于粗糙集理论的第一届国际学术会议，以后每年举行一届，到目前已举办十余届。粗糙集理论现已成为人工智能和信息科学最为活跃的研究领域之一，且对于认知科学也十分重要。粗糙集理论的主要任务是近似分类、知识约简（属性和属性值约简）、属性相依性分析、根据决策表产生最优或次优决策控制算法等。对它的研究主要集中在两个方面：一是理论研究方面，发表了关于粗糙集代数、粗糙集拓扑及其性质、粗糙逻辑和近似推理等方面的系列文献，建立起处理不完整、不精确、不确定问题的理论体系；二是应用领域方面，研究粗糙集方法在医学、管理学、金融学、气象学、图像处理、语音识别、字符识别和决策分析等众多领域的应用。

我国对该理论的研究起步较晚，始于 20 世纪 90 年代初期。1994 年，作者苗夺谦到中国科学院自动化研究所攻读博士学位，有幸师从戴汝为院士、王珏研究员，在国内较早开展了关于粗糙集理论与应用的研究。1997 年完成了国内第一篇关于粗糙集理论的博士学位论文《粗糙集理论及其在机器学习中的应用研究》，对粗糙集理论的知识表示、知识约简算法以及在机器学习中的应用进行了较为系统的研究。自 1999 以来，作者获得了三项国家自然科学基金项目“基于粗糙集理论的知识获取算法研究”（69805004）、“粗糙集与模糊集的融合研究及其在数据挖掘中的应用”（60175016）和“信息粒计算的表示、建模和推理方法的研究”（60475019）的资助，发表相关学术论文 30 余篇，获得省部级科技进步奖两项。本书系统总结了上述研究工作，并对国内外有关研究成果进行了归纳，系国家自然科学基金资助项目的研究成果。本书的主要目的是介绍粗糙集的基本理论与方法及该理论的研究发展状况。为了读者阅读方便，我们在书中对国内外已发表的相关粗糙集理论的文献进行了精选和系统化处理，突出了重点，规范了一些常用的记号，在统一的框架下阐述了粗糙集理论，初步探索了粗糙集与模糊集的融合，介绍了粗糙集在机器学习和自然语言处理中的应用。期望能为从事粗糙集理论、信息科学、自动控制、人工智能、粒度计算和模式识别相关研究的研究人员提供帮助。

本书在阐述基本概念和方法时，力求概念清晰、内容组织合理、论证严谨、

深入浅出、通俗易懂，充分体现出内容广泛、学术思想浓厚和学术观点新颖的特点。

全书共有 11 章。第 1 章介绍了经典集合论的一些相关背景知识；第 2 章详述了粗糙集的基本理论；第 3 章对粗糙集的代数结构进行了探讨，包括粗糙代数、粗糙群和粗糙环；第 4 章推广了粗糙函数的定义，讨论了粗糙函数的微积分性质，对粗糙微分方程作了简要介绍；第 5 章给出了粗糙集的三种新的知识表示，包括信息表示、差别矩阵表示和差别表表示；第 6 章从改变知识表示入手，提出了信息系统的若干种知识约简的启发式算法；第 7 章对决策表的知识约简进行了较深入的探讨；第 8 章提出了一种能够处理连续与离散属性的统一的粗糙集理论框架；第 9 章对粗糙集与模糊集的融合作了初步探讨；第 10 章介绍了粗糙集在 Monk 问题求解中的应用，并对粗糙集方法、AQ 方法及 ID 方法在理论与实验上进行了比较分析；第 11 章介绍了粗糙集理论在语言信息处理中的应用研究，如基于粗糙集的词性标注、信息检索、文字识别和文本分类等。

本书的编写和出版得到了国家自然科学基金、同济大学研究生教材出版基金以及清华大学出版社的大力支持，在此一并表示诚挚的谢意。

本书由同济大学苗夺谦审定。第 1、2 章由杭州电子科技大学李道国执笔，第 3 章由韩素青、苗夺谦执笔，第 4 章由李道国执笔，第 5、6、7 章由苗夺谦、李道国执笔，第 8 章由苗夺谦执笔，第 9 章由范世栋、李道国、苗夺谦执笔，第 10 章由侯丽珊、苗夺谦执笔，第 11 章由支天云、王素格、李道国执笔。本书写作过程中虽经多次修改，但仍可能存在一些纰漏，热忱欢迎广大同仁批评指正！

苗夺谦　李道国

2007 年 8 月于同济大学

目 录

Contents

第1章 经典集合论知识简介

经典集合是近代数学最基本的概念。为了与模糊集(fuzzy set)和粗糙集(rough set)相区别,经典集合(classical set)也称为普通集合(common set)或清晰集合(crisp set),它可以表达清晰的概念,例如奇(偶)数、三角形、抛物线等,再比如在知识表达系统中,经典集合表示清晰的可定义的信息范畴或信息粒。从认知科学的角度讲,一个概念可以用它的内涵和外延来刻画:符合某概念对象的全体构成此概念的外延;区别于其他概念的全体本质属性就是此概念的内涵。因此人们表达一个概念时,一般有两种方法:一是指出概念的内涵——内涵法;二是指出概念的外延——外延法。从集合论的观点看,一个概念可利用枚举式的外延法来表示,也可利用描述式的内涵法来表示,因而集合论成为描述客观世界中千差万别事物的理论工具。

本章首先介绍本书所需要的经典集合论的一些相关知识,例如,集合的概念和集合的运算,尤其是智能信息处理中常使用的集合的积与商运算;其次着重阐述集合论中关系的一般概念和基本性质,主要有等价关系、序关系和函数关系;最后简述粗糙集、模糊集和经典集合论之间的联系和各自的特点,且规范了一些常用的术语和记号,以便广大读者参阅。

1.1 经典集合论基础

19世纪末，德国数学家格奥尔格·康托尔(Cantor，1845—1918)创立了朴素集合论，但该理论在定义集合的方法上会导致悖论。为了消除这些悖论，罗素等一批数学家共同努力，在20世纪初创建了更严密、更精致的集合论——公理化集合论，它是微积分理论体系的基础，对现代数学和逻辑学的发展产生了巨大的影响。然而上述经典集合论无法处理模糊的信息和知识，1965年美国的控制论专家扎德(L. A. Zadeh)提出了模糊集合的概念，标志着模糊集合论的诞生。模糊集合论利用隶属度函数的数学方法来认识和处理模糊性，以适应现代控制论、信息论、系统论以及计算机科学发展的需要。当然，模糊集合论也有其局限性。近年来，在研究不完整数据及不精确知识的表达、学习、归纳等方法的基础上，波兰华沙理工大学的科学家帕拉克(Z. Pawlak)基于“知识(人的智能)就是一种分类能力”的观点，于1982年开创性地提出了粗糙集理论(rough set theory 或 rough sets)。粗糙集理论具有很强的定性分析能力，能够有效地表达不确定的或不精确的知识，善于从数据中获取知识，并能利用不确定、不完整的经验知识进行推理等，因此在知识获取、机器学习、规则生成、决策分析、智能控制等领域获得了广泛应用，特别是在数据挖掘领域，获得了巨大成功。与模糊集合论相比，粗糙集理论有自己独特的优势。模糊集合论和粗糙集理论极大地促进了集合论的发展，丰富了集合论的内容，使集合论成为我们求解问题不可或缺的理论工具。

为了更好地理解粗糙集理论，本节首先介绍经典集合论的一些相关的基本概念，诸如集合、空集、子集、子集簇、幂集等，然后介绍集合上的代数运算(交、并、补、差、积、商等)和一些定律，最后着重描述特征函数(隶属函数)的集合表示法。

1.1.1 经典集合论的基本概念

集合(经典集合)是集合论中一个未给予严密数学定义的最基本的概念。为了描述它，我们可以说一个**集合**就是将具有某种共同属性且彼此不同的对象放在一起，视为一个整体，其中组成这一整体的对象称为该集合的元或元素。例如，所有具有“2的倍数”这一特性的整数组成的集合，它表达了“偶数”这一概念。正如集合论的创始人康托尔所指出的“所谓的集合，可以理解为由我们的知觉或思维确定的、能明确区分开的对象 m_i 聚集成的一个整体 M，这些对象 m_i 叫作 M 的元素”。对于经典集合而言，元素和集合的关系只能是属于或不属于，非此即彼，且集合具有确定性、互异性和无序性，它可形式化地描述为 $\{x \mid x$ 具有性质 $P\}$。

集合举例：

(1) 太阳系的八大行星＝{金星，火星，木星，水星，土星，地球，海王星，天王