Rough集及Rough推理

刘清 著

科学出版社
Science Press

国家科学技术学术著作出版基金资助出版

Rough 集及 Rough 推理

刘 清 著

科 学 出 版 社
2001

内 容 简 介

Rough 集理论是一种处理含糊和不精确性问题的新型数学工具。

本书共分七章，分别介绍了 Rough 集的研究现状和发展趋势、Rough 集的基本概念及其理论基础、数据约简的各种方法、数据推理原理和各种推理模式、Granule-软计算、Rough 逻辑及其推理系统等。本书内容新颖，取材于国内外最新资料，总结了作者近年来的研究成果，反映了 Rough 集理论及其应用研究的现状和研究的新水平。每章后面的思考题既可帮助读者理解概念，领会内容，又可供进一步深入研究作参考。

本书可用作计算机及相关专业的科研人员和高校教师开展 Rough 集理论和应用研究的主要参考书之一；也可用作计算机及相关专业研究生的教材或本科高年级学生选修课教材。

图书在版编目(CIP)数据

Rough 集及 Rough 推理/刘清著. －北京：科学出版社，2001
ISBN 7-03-009580-4

Ⅰ.R… Ⅱ.刘… Ⅲ.粗糙集 Ⅳ.O144

中国版本图书馆 CIP 数据核字(2001)第 041395 号

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
丽源印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2001 年 8 月第 一 版 开本：850×1168 1/32
2001 年 8 月第一次印刷 印张：8 1/8
印数：1—1 500 字数：203 000

定价：20.00 元

(如有印装质量问题，我社负责调换〈北燕〉)

Preface

It is my great pleasure to welcome the initiative of Professor Qing Liu to present a book on rough set theory in Chinese language.

Rough set theory is a new mathematical approach to uncertain and vague data analysis. It is, no doubt, one of the most challenging areas, beside fuzzy set theory, of modern computer applications nowadays and a new, very important and rapidly growing area of research and applications. The application of rough set theory for knowledge discovery, data reduction, decision support, classification, pattern recognition, control and others have proved to be a very effective new mathematical approach. The theory has found many, interesting real-life applications in medicine, banking, engineering and others.

Many international renowned conferences and seminars included rough sets in their programs. So far about two thousand papers and several books have been published on rough sets and their applications worldwide.

The book written by Professor Qing Liu presents the basic concepts of rough set theory through clear and well-organized way and outlines some applications of this idea. The book can serve as a valuable introduction to rough set based methods of data analysis and as a guide for this new fascinating and fast spreading discipline.

I am deeply convinced that the book will play an important role in pursuing further development and applications of rough sets in China.

Professor Qing Liu needs to be congratulated for his valuable work.

Zdzislaw Pawlak
May 29,2001

前　言

一、Rough 集提出的背景

经典逻辑中只有真、假二值，但实际上有大量含糊现象存在于真和假二值之间。因此，长期以来许多逻辑学家和哲学家就致力于研究含糊概念。早在 1904 年，谓词逻辑的创始人 G. Frege 就提出了含糊（德文 Vague）一词，并把它归结到边界线区域，也就是说在全域上存在一些个体既不能在其某个子集上被分类，也不能在该子集的补集上被分类。例如，“高个人”概念，在人类全域的某个子集——中国人的集合上不能被分类出究竟多高的人才归于“高个人”？同样，在中国以外的其他国家的人的集合上也不能被分类出“高个人”的类集。于是“高个人”是含糊概念。

20 世纪 60 年代初，L. A. Zadeh 提出了模糊（用英文词 Fuzzy 翻译德文 Vague）集，不少理论计算机科学家和逻辑学家，试图通过这一理论解决 G. Frege 的含糊概念，但遗憾的是，模糊集是不可计算的，没有给出数学公式描述这一含糊概念，故无法计算出它的边界线上的具体的含糊元素数目。如，模糊集中的隶属函数 μ 和模糊逻辑中的算子 λ 都是如此[125]。时隔 20 年后的 80 年代初，Z. Pawlak 针对 G. Frege 的边界线区域思想提出了 Rough（波兰人对 Vague 的译文）集[1]，他把那些无法确认的个体都归属于边界线区域，而这种边界线区域被定义为上近似集和下近似集之差集。由于上近似集和下近似集都可以通过等价关系给出确定的数学公式描述，所以含糊元素数目可以被计算出来，即在真假二值之间的含糊程度可以计算，从而实现了 G. Frege 的边界线思想。Rough 集理论主要兴趣在于它恰好反映了人们用 Rough 集方法处理不分明问题的常规性，即以不完全信息或知识去处理一些不分明现象的能力，或依据观察、度量到的某些不精确的结果而进行分类数据

的能力。

二、Rough 集及其应用的现状

Rough 集理论是一种处理含糊和不精确性问题的新型数学工具，比起模糊集，对于当今现代计算机的应用来说，这种理论无疑是最有挑战性的领域之一。它自问世以来，无论是在理论或应用上都是一种新的、最重要的并且是迅速发展的一门既有理论又有应用的研究领域。对于人工智能和认知科学似乎也是十分重要的，尤其在机器学习、知识获取、决策分析、数据库的知识发现、专家系统、决策支持系统、归纳推理、矛盾归结、模式识别、模糊控制及其他各个方面的应用，Rough 集理论都为之提供了一种很有效的新的数学方法。Rough 集自提出以来就一直得到模糊数学的创始人 Zadeh 的重视，并给予很高的评价，把它列入他新提倡的软计算的基础理论之一。由此可见，Rough 集理论及其广泛应用越来越受到重视。

Rough 集概念在某种程度上与其他为处理含糊和不精确性问题而研制的数学工具有相似之处，特别是和 Dempster-Shafer(DS)证据理论相似。两者之间的主要区别在于：DS 理论利用置信和似然推理函数作为主要工具；而 Rough 集理论利用下近似集和上近似集。另一种关系存在于模糊集理论和 Rough 集理论之间，从它们之间的比较可以看出，这两种理论不是互相冲突，而是互相补充的。总之，Rough 集理论和模糊集理论对于不完全的知识来说它们有各自独立的方法。

当前许多国际重要学术会议和学术研讨班都把 Rough 集理论的研究列入会议和讨论班的主要内容之一。至今大约有 2000 多篇 Rough 集方面的研究论文发表于国际重要期刊和国际会议刊物上，这些研究成果总结并编著成书，国际上虽则不少，但国内这类书甚少，如此不利于我国学者研究 Rough 集理论和我国科技人员对 Rough 集应用软件的开发。在同仁和朋友的提议下，产生了撰写一本 Rough 集专著的念头，故撰写本书的目的在于进一步

促使 Rough 集理论及其应用在我国的发展。这种理论在许多重要的实际生活中都有应用，利用 Rough 集理论处理的主要问题包括数据库中的数据约简、数据相关性的发现、数据意义的评估、由数据产生决策控制算法、数据的近似分类、数据中的相似性或差异性的发现、数据中范式的发现以及因果关系的发现。特别地，Rough 集方法在医学、药学、银行、商业、金融、市场研究、工程设计、气象学、振动分析、开关函数、冲突分析、图像处理、声音识别、并发系统分析、决策分析、字符识别及其他领域都有重要的应用。

自 Rough 集理论提出以来，大致从两个方面研究 Rough 集理论及其应用。一方面是对 Rough 集的理论研究，发表了 Rough 集代数、Rough 集拓扑及其性质、Rough 逻辑及处理近似推理的逻辑工具等论文；在这些论文中充分论述了 Rough 集与 Fuzzy 集、证据理论与 Rough 集理论之间的关系，它们也建立了 Rough 集与概率逻辑、Rough 集与模态逻辑等的统一框架。

另一方面，Rough 集理论的研究者们很重视它的逻辑研究，发表了一系列的 Rough 逻辑方面的论文。比如，Z. Pawlak 在论文[76]中建立了五个逻辑真值；E. Orlowska 在论文[67]中提出以等价关系 R 作为新的谓词，以扩充经典二值逻辑；Lin 和 Liu 基于拓扑学观点定义了类似于下和上近似的算子 L 和 H，并建立了带这两个算子的近似推理的逻辑演绎系统[32]，Liu 在文献[25～29,31,35,88]中提出了带算子 L 和 H 的 Rough 逻辑的近似推理模式和归结原理，并证明了它的归结完备性定理，如此等等。所有这些研究都为经典逻辑在近似推理中的应用开辟了新途径，也为国际上的大逻辑学家王浩先生 40 年前提出的“Approximate Proof”的实现迈出了好的第一步，更为 Rough 集理论在近似推理中提供了具体的应用途径。另外，有关 Rough 集方法的函数研究方面，近年来出现了不少 Rough 数及 Rough 隶属函数的研究，发表了一系列关于实数 Rough 离散化和实函数 Rough 离散化方面的论文[2,75,87,114]。

三、Rough 集及其应用的发展前景

不精确性是 Rough 集理论的关键词，它涉及集合论定义中的许多实质性内容。集合的近似定义是现代数学中的重要概念之一，而与布尔逻辑非常相关的经典集合论又是数字计算机运算的核心。众所周知，许多实践问题不能满足现存计算机的求解条件，特别是机器学习、模式识别以及某些控制问题等，这种困难常常使得不能建立描述个体的算法。而 Rough 集理论及其扩充对于建立此类个体的近似描述，提供了一种精确的数学技术。Rough 集方法对于处理这类问题提供了一种通用的由精确数学语言支持的哲学框架。目前，Rough 集理论有许多高水平的理论研究论文，但应用方面好的成果还不多。本书列举了一些应用实例，从这些例子充分说明了 Rough 集理论已经构成了 KDD(Knowledge Discovery in Database，数据库中的知识发现)的一个完备基础，说明了它也是分布式和多-Agent 系统中数据挖掘的新方法。近来对于大型数据库中的数据挖掘的 Rough 集方法已经提出来了。在这里将进一步强调一下 Rough Mereology 方法是很有前景的研究领域，它是 Rough 集理论的扩充，是一种包含关系，记成 $x\ \mu_r\ y$，意味着 x 是 y 的以程度为 r 的部分。Rough Mereology 提供了智能 Agent 在分布式环境下对个体进行综合和分析的一种方法学。近来 Rough Mereology 已经被用作研究信息 Granule 计算的基础，试图朝着用词计算的公式化方向发展。Rough Mereology 的应用前景是寻找从数据提取逻辑结构的算术方法。例如，寻找对应于关联特征的关联规则的提取、综合缺省规则、综合近似推理模式，以及建立从数据提取更高层的知识等方法。这些方法的研究与开发是进一步发展 Rough 集的关键，也是 KDD 研究的中心问题之一[4,73,93,94,116]。

可以预言，Rough 集方法将在数据挖掘和软计算，特别是处理大型数据库和复杂问题等方面，显示出“英雄有用武之地”的气魄。

具体归纳为以下几个内容：

1. 概念近似

在传统的方法中，概念近似是由被列在给定的数据表中的信息构造的，但在许多应用的实例中，这个给定的表提供给我们的只是一部分有效信息，因此，我们必须获取关于个体及其分类的另外信息。从这另外的信息，可以得到扩充原数据表的一种新数据表。

2. 动态信息系统

已经发现，当动态过程被应用于实践时，方能得到比较好的分类结果，如动态约简。动态信息系统的一般思想是通过信息系统来标记一个偏序集，而其子表的偏序确定方法为：如果 y 的子向量扩充了 x 的子向量，则称这个表 x 优先于表 y。

3. 多-Agent 系统

上面已提到，综合一个信息的全局 Granule 是由局部信息 Granule 的合成得到的。这个综合过程是分布式或多-Agent 系统中的一种近似推理。从这个观点看，它似乎要求把现有的算术 Rough 集方法扩充到分布式和多-Agent 环境。

4. 基本研究工具的开发

(1)逻辑。在近似推理中应用逻辑的方法将体现在以下方面：

- 对于近似推理模式采用综合方法。
- 对于合成基本相关性提取关联规则。

(2)代数。Rough 集理论中研究的代数结构包括各种集合类型。可通过提取这些必然的结构来研究所推测的结构是否是合适的。

(3)拓扑。个体空间上的拓扑表示通常隐藏在数据中，并且为了近似化一个决策函数，通常要显示它们在数据中的存在性。这里的重要问题是从数据中提取适当的函数，它将为构造决策函数

的近似提供相关的启发信息。

• 从数据中提取信息 Granule 的逻辑结构，这是属于探索逻辑中是否存在一种拓扑模型的表达方式；

• 构造各种知识结构中的界面，这种问题的组成形式是与 Granule 计算相关的；

• 从数据中提取相似性测度函数，我们希望有一系列的测度方法，以此构造这种拓扑模型。

上述提到的研究课题对开发 KDD 问题也有重要影响。特别地，对于从数据中提取各种格式的关联规则的研究尤其重要，如多-Agent环境中的关联规则。在数据挖掘和知识发现中，可利用各种知识源和各种知识结构的有利条件探讨混合系统中新的算术方法，也就是说，Rough 集方法将和其他方法结合起来使用，所谓其他方法，即模糊集、神经网络、进化计算、统计推理、证据理论、置信网络等。

四、本书的主要内容

本书共分七章。第 1 章为阅读本书的基础；第 2 章介绍 Rough 集的基本概念、近似分类、广义 Rough 集、Rough 实函数；第 3 章介绍了数据约简的各种方法；第 4 章介绍数据推理原理和各种数据推理模型、各种决策规则提取的方法以及决策规则产生器的研究；第 5 章讨论了 Granule-软计算，它是 Rough 集理论更深入研究和发展的新动向，着重介绍了当前研究得比较多的理论，如分布式环境中的 Rough 集、Agent 及多-Agent 系统；第 6 章介绍 Rough 逻辑和它的公理化系统、归结推理及其他几个 Rough 系统；第 7 章介绍 Rough 集理论的广泛应用的一些实例，所选实例来自各个专业领域，有社会科学、自然科学、工业控制和实际生活等。每章后面都配有思考题，有的思考题对巩固概念和领会内容大有帮助，有的则对进一步研究其理论和拓广其应用极有启发价值。

五、本书的撰写特点

本书取材除了综述国内外现有关于 Rough 集方面的文献中的精华内容之外，其余大部分是作者本人近年来的研究成果和国际上的最新研究动态。

书中首先介绍了研究 Rough 集理论的基础知识，着重阐述了 Rough 集的基本概念，力求概念清晰，内容组织合理。在此基础上描述了用 Rough 集方法处理各种不精确或含糊数据分析的特点，力求使这些特点组合成能在机器上实现的算法；在论述决策表与决策算法等价时，又突出阐述了决策算法的优化或称极小化，作为实现数据约简的目标；特别地，本书指出了 Rough 集的研究新动向、新发展，尤其是对 Rough 逻辑及其在近似推理中的应用价值的介绍，对真值的研究和实函数的离散化等新课题的介绍，对国内学者进一步深入研究 Rough 集理论颇有启发意义，如第 5 章中的分布式环境中的 Rough 集和多-Agent 系统中的 Granule 以及用 Granule 概念来研究 Rough 集方法中的分类等；又如第 6 章的 Rough 逻辑中的归结原理和 Rough 系统等对经典二值逻辑的非单调化研究很有启发意义。本书除了一般介绍 Rough 集应用外，还专门用一章介绍 Rough 集方法的更广泛应用，而且是来自工、农、经、商、社、医各个领域的实例，这对促进 Rough 集的应用极有意义。以上都说明了本书所具有的特色，既内容广泛，又有浓厚的学术思想和颇为新颖的学术观点。其写作上注重概念准确，论证严谨，有深入浅出、通俗易懂之感。

六、本书的主要读者

可以是计算机及其相关专业的科研人员和高校教师、研究生、本科高年级学生、从事工、农、医、商、经营、通信等诸行业的科技工作的科技人员。

七、致谢

在撰写本书过程中得到波兰科学院院士、Rough 集创始人 Z. Pawlak 教授的直接指导，无论在材料来源、内容组织上，他都给予了具体的建议，并特意为本书撰写了英文序。美国 San Jose 州立大学 T. Y. Lin 教授、加拿大 Regina 大学 Y. Y. Yao 教授也给予了许多帮助，并多次建议我写一本中文的 Rough 集方面的书。清华大学石纯一教授、中国科学院计算技术研究所史忠植教授、原电子工业部信息中心刘锡荟教授等国内同仁也给予了极大支持，对书稿提出了许多宝贵意见。江顺亮博士仔细阅读了本书手稿，并提出了宝贵意见。对以上国际国内的学者和专家，谨在此表示真诚的感谢。

本书得到国家科学技术学术著作出版基金和国家自然科学基金资助，科学出版社对本书的出版也给予了大力支持，在此一并表示真诚的谢意。

由于作者水平有限，对 Rough 集及 Rough 逻辑的研究工作还很粗浅，有的内容甚至是阶段性的研究成果，不妥、错误之处恳请读者批评指正。

刘　清

2001 年 8 月 20 日

目　录

第1章　集合与关系

§1.1　集合及其运算

自从19世纪康托(Cantor)创立了集合论以来，集合论就已成为现代数学的基础。集合的表示通常有三种。其一是列举法，也就是集合的元素全部枚举出来，这只有元素数目少的情况下用这种方法。例如，中国的直辖市={北京，天津，上海，重庆}。其二是性质法，就是用集合中的元素具有某种共性来描述集合。例如，具有 $x<10$ 的特性的奇数的集合，$A=\{x:x<10\wedge x$ 是奇数$\}$。其三是特征函数法，即集合中的元素与特征值0和1对应起来表示。例如，一个学习小组有6个人，分别用 x_1,x_2,x_3,x_4,x_5 和 x_6 表示，如果 x_i 是男，则记成 $1/x_i$，若 x_i 是女，则记成 $0/x_i$，于是可将这个集合记成 $A=\{0/x_1,1/x_2,0/x_3,1/x_4,1/x_5,1/x_6\}$。

一个集合通常用大写字母表示，而集合的元素用小写字母。某个元素 a 属于集合 A，或不属于 A，分别记作 $a\in A$ 和 $a\notin A$。

一个集合的全部元素数目称做该集合的基数，记成 $|A|$ 或 $K(A)$，$\mathrm{card}(A)$。其基数可以是无限的。例如，{北京，天津，上海，重庆}和$\{x:x$ 为奇数$\}$分别为有限集和无限集。对于无限集，通常不能用数字来写出它的基数。

如果集合 B 中的元素全部都能在集合 A 中找到，则集合 B 被称为集合 A 的子集合，简称子集，记成 $B\subseteq A$ 或 $A\supseteq B$。显然，对任何一个集合 A，都有 $A\subseteq A$。用符号$\not\subset$表示不包含于。若存在一个 $x\in B$，但 $x\notin A$，则 $B\not\subset A$。

有两个特别的子集空集和全集。如果一个集不包含任何元素，则称之为空集，用$\varnothing$表示，或$|\varnothing|=0$。对任一集合 A，都有$\varnothing\subseteq A$。在一定范围内，如果所有子集均为某一集合，则称该集合