模糊粗糙集理论与方法

陈德刚 著

科学出版社

模糊粗糙集理论与方法

陈德刚　著

科学出版社

北京

内 容 简 介

本书系统总结作者近十年来在模糊粗糙集理论方面的研究成果，以决策系统中条件属性与决策属性之间的不一致性为主线，论述基于模糊相似关系的模糊集合的上、下近似及数学结构，模糊粗糙集的数字特征，基于模糊粗糙集的属性约简，最后重点论述模糊粗糙集与核方法的内在联系. 本书的特点是首先为模糊粗糙集理论建立完备坚实的数学理论框架，在此基础之上设计属性约简和分类的算法，实现了理论分析、算法设计和实际应用的结合.

本书的内容自成体系，既可作为应用数学和信息科学的高年级本科生和研究生的教材，也可作为相关领域的研究人员的参考书.

图书在版编目(CIP)数据

模糊粗糙集理论与方法/陈德刚著. —北京：科学出版社，2013.11
ISBN 978-7-03-039017-2

Ⅰ. ①模… Ⅱ. ①陈… Ⅲ. ①模糊集理论-研究 Ⅳ. ①O159

中国版本图书馆 CIP 数据核字 (2013) 第 256589 号

责任编辑：徐园园 赵彦超 / 责任校对：刘亚琦
责任印制：赵德静 / 封面设计：陈 敬

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

中国科学院印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2013 年 11 月第 一 版 开本：720×1000 1/16
2013 年 11 月第一次印刷 印张：11 1/4
字数：218 000
定价：58.00 元
(如有印装质量问题，我社负责调换)

前　言

模糊粗糙集理论与方法是 1990 年由 D. Dubois 和 H. Prade 提出的处理数值型数据中存在的不一致性的数学理论. 经过 20 余年的发展, 模糊粗糙集无论是在理论上还是在应用方面都取得了相当丰富的研究成果, 建立了比较完备的理论框架, 并且在机器学习与数据挖掘等领域有着广泛的应用. 目前关于粗糙集的专著无论是英文的还是中文的都已经出版了很多种, 在这些专著中虽然有些章节专门介绍了模糊粗糙集理论, 但是限于篇幅有限或者侧重点不同等原因, 其论述一般都没有全面反映模糊粗糙集理论的发展历史以及研究现状, 也没有覆盖及总结目前模糊粗糙集主要的研究方向, 因而也没有为模糊粗糙集的未来发展趋势进行合理的科学预测, 从而不能满足专门研究者的需要. 随着国内外相关领域研究人员特别是一些博士或硕士研究生对模糊粗糙集越来越关注, 对模糊粗糙集的发展历史进行全面细致的回顾与总结, 并对其研究现状进行梳理分类从而为研究者们提供预备知识以供参考的需求就越来越迫切了. 本书正是基于这样的目的来完成的.

作者在 2000 年 7 月于哈尔滨工业大学数学系基础数学专业获得理学博士学位之后, 经过短暂的暑假休整, 于 9 月赴西安交通大学理学学院数学研究中心博士后流动站工作, 合作导师为张文修教授. 在此之前, 尽管在电话里张老师说过要我从事粗糙集方面的研究, 但是直到见到张老师之前作者对什么是粗糙集一无所知. 张老师把他与吴伟志、梁吉业和李德玉几位博士研究生 (当时在读) 刚刚完成的专著《粗糙集理论与方法》的校对稿带给我, 要我在新学期的讨论班上进行讲解, 当时的讨论班补充了新入学的几位博士研究生, 如米据生和魏立励等. 由于作者此前从事的都是基础数学方面的研究, 缺乏数据挖掘与机器学习的研究背景, 导致作者在最初始终很难触及到粗糙集理论的核心思想. 这一点可以从作者当时完成的两篇与格和群有关的论文看出来, 在这两个研究中没有从粗糙集的应用背景出发而片面追求数学上的抽象, 从而也缺乏继续发展的动力.

在西安交通大学工作的两年对作者来说一件幸运的事情是有幸聆听了张文修老师多次具有深刻思想性的讲解; 另一件幸运的事情是作者结识了当时刚刚从事粗糙集研究不久并且后来构成西安交通大学粗糙集研究团队核心力量的几位同行, 特别是现已为教授的吴伟志和米据生把他们当时关于模糊粗糙集刚刚完成尚未发表的研究成果无私地介绍给我, 使得作者能够在此基础上于 2002 年在香港理工大学访问期间完成了第一篇关于模糊粗糙集的研究论文, 也开启了作者从事这一领域研究的大门. 后来作者于 2002 年在清华大学自动化系 CIMS 研究中心从事博士后

研究工作, 并且每年都有机会赴香港理工大学电子计算学系短期访问, 开始大量地接触与计算机科学相关的研究领域, 也慢慢地领悟了一些粗糙集理论研究与基础数学研究的不同之处, 意识到了模糊粗糙集的生命力在于其强烈的应用背景, 因而逐渐地改变了自己的研究侧重点, 更多地把自己的研究出发点定位于模糊粗糙集在数据处理方面的应用.

本书的主要内容取材于作者近十年来在模糊粗糙集领域内完成的一些研究工作, 包括经典粗糙集理论介绍、模糊粗糙集的构造性与公理化方法、模糊粗糙集的数学结构、模糊粗糙集的度量、基于模糊粗糙集的属性约简理论与方法、核函数框架下的模糊粗糙集理论与方法以及变精度模糊粗糙集理论与方法. 除第 1 章对模糊集合的基本概念和理论进行了简单的介绍以外, 其余所有的章节绝大部分内容都是作者的研究成果, 因而本书称为一本专著是合适的. 正因为本书过于强调原创性, 因而也必然带来内容选择上的局限性. 虽然本书中大部分内容能够反映模糊粗糙集目前的研究水平, 但是也有少部分内容相对于目前的研究现状略显滞后. 并且由于个人能力问题, 作者的研究工作也不可能全面涉及与涵盖模糊粗糙集全部的研究方向, 因而本书中列出的研究领域也具有一定的局限性, 这一点请读者谅解. 好在书后的参考文献罗列了绝大多数作者认为在模糊粗糙集的领域中重要的文献, 读者可以按照自己的兴趣与需要自行选择参考.

本书作者在十年的粗糙集理论研究历程中得到了来自方方面面的关心和帮助. 张文修教授把作者带入到粗糙集的研究领域中并且给予了启蒙的指导; 吴伟志教授、米据生教授在作者的研究刚刚起步时无私地提供了最新的研究成果; 与梁吉业教授、李德玉教授、钱宇华教授、魏玲教授、李同军教授、邵明文教授等多次的讨论拓宽了作者的视野; 与杨苏教授、王熙照教授、邝得互教授、曾祥财教授、胡清华教授等的合作加速了研究工作的进程; 王长忠博士、赵素云博士、何强博士等协助完成了部分研究工作. 作者在此表示衷心的感谢!

本书的相关研究得到了国家自然科学基金 (71171080, 61170107) 和华北电力大学文理振兴计划的资助, 在此一并表示感谢!

由于作者水平有限, 书中难免存在各种不足之处, 敬请各位读者不吝赐教.

陈德刚

2013 年 6 月

目　　录

第0章　绪　　论

人类社会发展历史有明确记载的已经有几千年, 在这漫长的发展历史过程中人类通过生产实践和社会实践等活动在不停地认识和改造客观世界. 随着时间的推移, 人类社会的生产力在飞速发展, 科学技术水平在日新月异地不断提高, 人类对客观世界的认识也越来越深入, 这些认识一般以知识的形式被人类代代相传并且不断地丰富完善. 一般来说, 人类首先通过生产实践和社会实践活动采集各种各样的数据, 通过对这些数据的分析以获得潜在的知识. 现代社会随着计算机技术的发展以及许多大规模生产、科研、军事、经济等活动的开展, 人类收集数据、存储数据的能力得到了极大的提高, 在许多领域中都积累了大量的数据, 如何有效地利用这些数据来获取潜在的知识为人类造福就成为几乎所有领域的共同需求. 由于这些数据确实数量庞大、结构复杂, 与生产力相对落后时收集到的数据无论是从规模上还是从复杂程度上都远远不可同日而语, 因而再像从前那样基本靠人脑去从数据中归纳知识发现客观规律显然不现实. 因而人们开始研究利用计算机对海量数据进行分析以发现数据中潜在的知识. 在这样的大趋势下, 智能数据分析技术如机器学习和数据挖掘等方法受到了广泛的关注.

利用智能数据分析技术对海量数据进行分析面临着许多困难, 其中有些困难从理论上讲是可以通过开发更先进的技术手段克服的, 比如, 海量数据的存储问题和不完备数据的完备化等. 但是有些困难不能单纯从技术的角度去解决, 比如, 数据中所存在的各种不确定性的处理就需要从理论的角度去进行研究, 以掌握其基本规律, 再以此为基础去发展相应的技术手段来处理这些不确定性. 早期谈到不确定性的时候一般指的是随机性, 人们提出了概率论与数理统计的数学方法作为处理随机性的工具. 从 20 世纪中期开始人们就自觉地开始利用统计的方法来研究智能数据分析的理论与方法, 最具有代表性的理论就是所谓的统计学习理论. 经过近半个世纪的发展, 统计学习理论已经建立起了完备的理论框架, 基于统计学习理论人们提出了多种有效的处理海量数据的技术手段, 并且在很多领域里得到了很好的应用. 在目前的机器学习领域中, 统计学习理论作为基本的理论已经广为接受, 被认为是主流的研究方向之一. 在统计学习理论中一个最基本的假设就是人们收集到的样本是独立同分布地服从某个未知的分布函数, 这些样本不是用来估计未知的分布函数而是用来直接构造某种学习机器. 尽管所构造的学习机器可能是个 “黑匣子”, 但是仍然具有很好的推广性. 这实质上已经背弃了传统的归纳演绎的认知体系, 而是实现了由数据直接到预测的跳跃. 事实上, 这种跳跃也是无奈之举, 是由所面临问题

的大规模和复杂性所导致的一种妥协与折中. 尽管统计学习理论中对算法的推广性有着完善的理论基础, 但是仍然改变不了算法难以解释的事实. 由于统计学习理论对问题的叙述相对抽象, 所涉及的数学理论不仅有概率论与数理统计, 而且还涉及泛函分析、函数逼近论等其他的现代数学理论, 并且具有相当的技巧性, 对于不具备相应数学背景的研究者想要充分地掌握该理论具有相当的难度.

由于基于统计的智能数据分析方法在机器学习的研究中一直占有主导的地位, 导致了对数据中其他类型的不确定性的研究没有得到机器学习领域广泛的重视. 然而这并不能否定数据中确实存在着各种与随机性不一样的不确定性, 并且随着对统计学习理论的认识的逐渐加深, 人们也意识到统计学习理论自有其局限性. 比如, 样本是独立同分布地服从某个分布函数的假设就过于理想化, 在很多现实问题中这一假设过于勉强. 事实上这个假设的目的更多的是为了理论推导的需要. 从这个假设也可以看出统计学习理论并不明显包含处理其他类型的不确定性的理论机制, 因而其并不能有效地处理其他类型的不确定性. 事实上, 海量数据中确实存在着不同于随机性的其他种类的不确定性, 比如, 模糊性和突变性等. 这里不准备对海量数据中存在的各种不确定性进行详细的分析与综述, 我们下面仅就对本书中重点涉及的一种不确定性的研究进行简单的回顾, 这种不确定性就是所谓的条件属性与决策属性之间的不一致性.

在许多实际问题中常常会观察到这样的现象, 两个对象具有相似甚至相同的描述, 但是却对某一决策问题具有完全不同的结论. 特别是对一些缺乏先验知识的突发事件这种现象往往更为突出, 比如, 许多突发性的流行病, 人们往往一时找不到导致这种传染病的根本原因所在, 就会导致对类似症状的患者无法做出准确的诊断. 如同我们在 2.1 节中分析的那样这种不确定性的存在根源于人类对客观世界认识能力的局限性. 我们以最基本的二分类问题为例, 如果收集到的数据不同类别之间的差别明显, 那么从这样的数据中学习到的分类规则就不会有很显著的歧义性. 反之如果相当数量属于不同决策类的样本之间没有明显的区别, 那么学习到的分类规则的推广性就要大打折扣. 事实上, 后一种情况无论使用什么学习方法都不会取得像前一种情况那样的分类效果. 学习问题的难点之一就是如何处理这种不一致性. 在统计学习理论中尽管隐含着具有处理这种不一致性的机制 (如软间隔的支撑向量机), 但是其一般把这种现象归结为误差或者边界点的存在, 并没有从其产生的根源上去分析.

为了能够准确地评估且处理数据中存在的这种条件属性与决策属性之间的不一致性, 1982 年波兰数学家 Z. Pawlak 提出了经典的粗糙集理论[1]. 正如模糊集合不是为了把清晰的集合变得模糊一样, 粗糙集理论也不是为了把明确的集合变得粗糙. 这里的粗糙意指数据中决策类不能由条件属性所准确刻画, 反映的是条件属性与决策属性对训练样本进行分类时产生的不协调. 对于那些矛盾的样本粗糙集理论

不去进行任何预处理而是承认其存在的客观性, 这样表示出来的知识就有某种不确定性, 这个特点是粗糙集理论区别于其他的智能数据处理方法的基本特征. 经典的粗糙集理论同时为符号值的关系数据库提供了知识表示的基本数学框架, 在粗糙集理论中把与数据相关的一些基本概念, 如样本集合、属性以及决策规则等利用等价关系和划分等数学概念来表示出来, 使得数据中所蕴涵的知识可以被抽象成数学概念并进行抽象的运算, 从而得到了以集合论为基本框架的知识表示体系, 同时也为把粗糙集理论与其他的方法相结合奠定了数学基础.

从数学的角度来看, 粗糙集理论不像大多数现代数学理论那样具有高度复杂性和抽象性, 掌握粗糙集理论也不需要过多的现代数学方面的预备知识. 但是简单易行正是粗糙集理论的优点所在, 对于符号值的关系数据库, 粗糙集理论中涉及的数学工具已经足以完成表示和挖掘知识的任务, 以此为基础开发出的许多算法在处理实际问题时表现出了优异的性能. 根据奥坎姆剃刀原则简单实用是第一位的, 由于粗糙集理论的出发点就是去挖掘和表示数据中蕴涵的知识, 并不是为了建立高深漂亮的数学理论, 与其他的理论与方法相结合也是为了更好地完成这一任务, 对于经典的粗糙集理论有机地借鉴其他领域的思想和方法 (而不是生拉硬拽地为了结合而结合) 一直是其发展的动力源泉.

受到冷战的影响以及语言的限制, 粗糙集理论在刚刚提出来的时候主要是一些东欧国家的学者对其进行了研究. 1991 年, Pawlak 出版了《粗糙集——关于数据推理的理论》这一专著, 接着 1992 年 R. Slowinski 主编出版了首部以粗糙集为主题的论文集[2], 推动了国际上对粗糙集理论与应用的深入研究. 1992 年在波兰召开了第一届粗糙集方面的国际研讨会, 着重讨论了集合近似定义的基本思想及其应用和粗糙集环境下的机器学习基础研究, 从此每年都会召开一次以粗糙集理论为主题的国际研讨会, 从而推动了粗糙集理论的拓展和应用. 我国对粗糙集理论的研究起步较晚, 所能检索到的最早发表的论文时间是 20 世纪 80 年代. 南昌大学的刘清教授、四川理工大学的曾黄麟教授、同济大学的苗夺谦教授以及河北大学的王熙照教授等在粗糙集领域中率先进行了有价值的探索, 开始了我国在这一方向的研究工作. 直到 20 世纪末, 尽管对粗糙集理论的研究在国际上已经吸引了相当数量研究者的注意力. 例如, 1995 年 ACM 将粗糙集理论列为新兴的计算机科学的研究课题, 但是与同时期出现的以支撑向量机方法为代表的统计学习理论相比较, 仍然处于计算机科学研究的边缘地位, 这一点可以从发表文章的数量和层次以及引用情况看出来. 尽管如此, 这一阶段的理论研究具有明显的原创性, 开拓了若干重要的研究方向, 这些研究方向直到目前仍然是粗糙集理论研究的核心内容. 在这些研究中, 值得重点强调的有以下几个方面的成果.

第一个就是 1992 年 A. Skowron 和 C. Rauszer 提出的关于属性约简的辨识矩阵理论[3], 在此之前人们计算属性约简都是利用启发式算法, 对约简的结构和原理

缺乏清晰的认识. 辨识矩阵方法的提出奠定了基于粗糙集理论的属性约简的数学基础, 为开发新的算法指明了全新的方向, 具有重要的理论意义和应用价值. 而且辨识矩阵的方法使用了离散数学中的逻辑运算的方法, 其理论从数学上看也不是平凡的, 辨识矩阵的方法是整个经典的粗糙集理论数学特点最突出的部分. 第二个是 1993 年 W. Ziarko 提出的变精度粗糙集理论[4], 它克服了经典粗糙集理论中上、下近似定义中对集合的包含过于苛刻的要求, 能够处理数据中的噪声, 使得粗糙集的方法具有更好的容错性, 因而也就有着更好的应用价值. 第三个是 Y. Y. Yao 关于经典粗糙集基本性质和结构的研究[5,6], 通过构造性和公理化的方法对经典粗糙集进行了细致完备的刻画, 把近似算子的性质依据其重要性进行了分类, 揭示了近似算子的本质特征, 其研究方法对其他类型的粗糙集的研究具有重要的指导意义, 直到目前仍然是研究新提出的粗糙集模型的首选方法. 第四个是以 W. Zakowski 于 1983 年提出的覆盖粗糙集模型为代表的广义粗糙集模型[7], 把经典粗糙集中的划分替换为覆盖虽然是一个自然又简单的思想, 但是这种替换带来了对粗糙集理论全新的认识, 并且极大地拓宽了粗糙集的研究方向, 对覆盖粗糙集和广义粗糙集[8] 的研究直到目前仍然是粗糙集理论的研究热点之一. 第五个是 1990 年 A. Skowron 开始的关于粗糙集与证据理论之间的关系的研究[9], 通过利用证据理论中的信任函数和似然函数来等价刻画粗糙集中的近似算子开创了对粗糙集数字特征的研究. 最后一个就是由 D. Dubois 和 H. Prade 于 1990 年正式提出的模糊粗糙集理论[10], 通过引入模糊相似关系把模糊集合与粗糙集相结合对模糊概念进行了近似刻画, 从而把粗糙集的应用范围从符号值数据拓宽到实数值数据.

进入 21 世纪以来, 随着互联网的飞速发展人们获得信息的渠道大大拓宽, 速度也得到了极大的提高, 也使得粗糙集理论在全世界范围内得到了更多的关注, 越来越多的研究人员开始从事这方面的研究, 取得了大量的研究成果. 截至 2013 年 4 月 30 日, 在 Google Scholar 上输入 “rough set” 进行搜索, 可以得到 2600000 多条记录, 尽管这里面有许多是重复的, 但是仍然说明目前对粗糙集理论的研究成果的丰富程度. 在 21 世纪初开始对粗糙集理论的研究工作中, 中国学者的研究工作占有举足轻重的地位, 形成了所谓的粗糙集研究的中国学派, 其中以西安交通大学张文修教授指导的粗糙集研究团队最具有代表性. 事实上, 从 20 世纪末开始在西安交通大学理学院由梁吉业教授 (当时为在读博士生) 提议就开始了有关粗糙集理论的讨论班, 主要参加者有吴伟志教授和李德玉教授等, 2000 年以后随着米据生教授、魏玲教授、李同军教授等的加入, 这个研究团队越来越壮大, 尽管每年都有成员毕业离开, 但是每年又都会有新鲜的血液补充输入进来. 这种局面一直持续了大约 10 年, 培养出几十位粗糙集研究方面的专家, 取得了一大批重要的研究成果. 其中代表性的研究有梁吉业教授利用包含度理论对粗糙集数值特征的研究, 吴伟志教授和米据生教授关于模糊粗糙集理论构造性和公理化方法的研究以及李德玉教授关

于信息系统映射性质的研究等. 尽管西安交通大学研究团队随着张文修教授退休和各位成员的毕业离校而解散, 但是团队成员仍然坚持每年至少进行一次学术活动来报告各自的最新研究成果, 讨论当前的研究动态. 在西安交通大学粗糙集研究团队之外, 国内还有一些其他的团队在从事这方面的研究, 也取得了许多重要的研究成果, 这里就不一一列举. 总之, 目前国内对以粗糙集为代表的粒计算理论的研究正在如火如荼地开展, 每年一次的粒计算国内会议都吸引了大量的研究人员参与, 每年也都能获得若干项国家自然科学基金的资助.

由于本书主要涉及模糊粗糙集理论, 因此我们对粗糙集研究领域的其他内容不进行详细介绍. 作者在 Google Scholar 上以 "fuzzy rough set" 为主题词进行了搜索, 共得到搜索条目 187000 多条, 这也堪称是一个海量数据集. 显然以作者的个人力量无法把这些文献一一综述, 只能依据作者这些年对模糊粗糙集的了解来回顾一下模糊粗糙集理论的研究历史和研究现状, 并且展望一下未来的研究方向. 这里主要综述与本书内容密切相关的研究方向, 所引用的文献基本上是作者比较熟悉的和在该方向比较有代表性的, 对于其他不熟悉的模糊粗糙集方面的重要研究作者不敢妄加不负责任的评论, 因而所综述的内容必然有相当大的局限性, 此乃受到作者的研究水平所限, 也请相关的专家与读者见谅. 在以下的介绍中会涉及一些专业术语, 对于那些在本书后面的章节中会给出术语的详细解释, 为了避免重复我们这里就先不进行解释.

在绝大多数关于模糊粗糙集的文献中都会综述说模糊粗糙集的概念第一次是由 D. Dubois 和 H. Prade 在文献 [10] 中利用模糊等价关系针对三角范数 $T_M = \mathrm{Min}$ 提出的, 事实上这种说法不够准确. 基于 $T_M = \mathrm{Min}$ 的模糊粗糙集最早由 L. Farinas del Cerro 和 H. Prade 于 1986 年在文献 [11] 中引入的, 文献 [10] 把这种特殊的模糊粗糙集利用可能性理论的方法推广为基于一般三角范数的模糊粗糙集, 并且在下近似算子中首次采用了 S-蕴涵算子和 R-蕴涵算子. 下面我们按照文献 [10] 中给出的方式和符号原封不动地列出这个定义.

设 X 是一个非空论域, R 是一个 X 上的模糊 T-相似关系, 对任意一个模糊集合 F, 其上、下近似分别定义为

$$\forall x \in X, \quad \mu_{\omega(R^*(F))}(x) = \sup_y \mu_F(y) * \mu_R(x, y),$$

$$\forall x \in X, \quad \mu_{\omega(R_*(F))}(x) = \sup_y \mu_R(x, y) \to \mu_F(y).$$

这里 $*$ 在文献 [10] 中用来表示一个三角范数, "$\to$" 定义为 $a \to b = 1 - a * (1 - b)$ 称为一个 S-蕴涵算子. 在文献 [10] 中还定义了 R-蕴涵算子 "$\Rightarrow$" 为 $b \Rightarrow c = \sup\{x : x * b \leqslant c\}$, 利用 R-蕴涵算子 "$\Rightarrow$" 定义了另一种下近似为 $\forall x \in X, \mu_{\omega(R_*(F))}(x) = \sup_y \mu_R(x, y) \Rightarrow \mu_F(y)$.

尽管文献 [10] 中给出的符号比较特殊, 也比较晦涩难懂, 但是其实这里的上、下近似算子就是目前常见的模糊粗糙集的近似算子定义的等价表示形式, 也是本书中所重点讨论的对象. 在文献 [10] 中对上近似算子和两种下近似算子的性质进行了详细的刻画, 现在所熟知的模糊粗糙集的基本性质都在文献 [10] 中得到了证明, 这些性质在本书第 4 章中也特别罗列出来.

由以上的综述可以看出文献 [10] 确实是模糊粗糙集理论的奠基性文献. 借鉴了文献 [10] 中提出的对模糊集合进行近似逼近的思想之后, 在早期关于模糊粗糙集的研究中许多学者提出了各种不同的把模糊集合与粗糙集相结合的方法, 这些方法由于缺乏完备的数学定义和具体的应用背景因而没有体现出生命力, 基本上是在提出之后缺乏后续的深入研究而湮没. 这里列出其中的几篇参考文献 [12]~[15], 有兴趣的读者可以自行了解一下. 早期其他的关于模糊粗糙集的研究还包括在一些实际问题上的应用, 这里不再一一列举, 有兴趣的读者可以自行到网上查询.

在粗糙集的发展历史上 1998 年是一个值得提起的年代, 许多重要的研究结果纷纷在这一年发表出来. 比如, 在我们前面提到的 Y. Y. Yao 关于粗糙集近似算子的构造性和公理化刻画[6,7], 以及 Z. Bonikowski 关于不完备粗糙集的研究[16] 等都是在这一年发表. 对模糊粗糙集理论, 1998 年 N. N. Morsi 和 M. M. Yakout 发表在 *Fuzzy Sets and Systems* 上的文献 [17] 是其发展历史上的一个里程碑. 在文献 [17] 中针对文献 [10] 中提出的第二对近似算子利用公理化的方法进行了详细的刻画, 几乎把与这一对近似算子相关的性质都发掘出来, 所得到的结果具有相当的深度, 使用的方法具有代表性且被后来的研究者们所广泛采用. 文献 [17] 首先把文献 [10] 中晦涩的符号改进为现在通用的表示模糊粗糙集近似算子的定义, 我们仍然原封不动地如下给出文献 [17] 中的模糊集合的近似算子.

设 U 是一个非空论域, R 是一个 U 上的模糊 T-相似关系, 对任意一个模糊集合 μ, 其上、下近似分别定义为

$$\overline{A}_R\mu(x) = \sup_{u\in U} T(R(u,x),\mu(u)), \quad \underline{A}_R\mu(x) = \inf_{u\in U} \vartheta(R(u,x),\mu(u)),$$

这里 T 是一个三角范数, ϑ 是 T 的 R-蕴涵算子. 文献 [17] 的另一个重要的贡献是罗列出了 R-蕴涵算子 ϑ 的 18 条性质, 这些性质对模糊粗糙集的后续研究非常重要, 这一点可以从本书的内容中看出来.

对模糊粗糙集的发展另一篇重要的文献就是 [18]. 在文献 [18] 中考虑利用一般的 border 蕴涵算子 $\Im$ 来定义下近似算子, 所采用的模糊关系 R 是模糊等价关系. 这里 $\Im$ 是一个二元函数定义为 $\Im : [0,1]\times[0,1] \to [0,1]$, 满足 $\Im(1,0) = 0$, $\Im(1,1) = \Im(0,1) = \Im(0,0) = 1$ 和 $\Im(1,x) = x$. 上、下近似算子在文献 [18] 中定义如下:

$$\underline{F_{AS}}_{\Im}(A)(x)=\inf_{y\in U}\Im(R(x,y),A(y)),\quad \overline{F_{AS}}^{T}(A)(x)=\sup_{y\in U}T(R(x,y),A(y)).$$

文献 [18] 中还特指了三种 border 蕴涵算子称为 S-蕴涵, R-蕴涵和 QL-蕴涵, 这里的 S-蕴涵和 R-蕴涵与文献 [10] 中的定义是一致的. 由于在近似算子 $\underline{F_{AS}}_{\Im}$ 和 $\overline{F_{AS}}^{T}$ 的定义中采用的是模糊等价关系而不是一般的模糊 T-相似关系, 因而 $\underline{F_{AS}}_{\Im}$ 和 $\overline{F_{AS}}^{T}$ 的定义中就存在着某种不协调性, 这也导致这些算子的性质不是很令人满意. 在文献 [18] 的最后还提出了一个 T_L-模糊粗糙集公理化的问题, 利用算子 $\underline{F_{AS}}_{\Im}$ 和 $\overline{F_{AS}}^{T}$ 无法完成对 T_L-模糊粗糙集的公理化刻画, 但是用文献 [17] 中的方法就可以完全解决这个问题.

2003 年和 2004 年吴伟志教授分别在文献 [19] 和 [20] 中研究了文献 [10] 中提出的第一对算子的结构, 所采用的方法包括构造性和公理化的方法; 2004 年米据生教授为文献 [10] 中提出的第二种下近似算子定义了一个与其对偶的上近似算子 $\overline{R}(A)(x)=\sup_{y\in U}\sigma(1-R(x,y),A(y))$, 并同样地利用公理化的方法进行了研究[21].

本书的作者在西安交通大学数学研究中心博士后流动站工作期间结识了吴伟志和米据生两位教授并且得到了他们的无私帮助, 很幸运地提前阅读了当时他们刚刚完成还未发表的文献[19−21], 以此为基础开始了对模糊粗糙集理论的研究. 在 2002 年于香港理工大学访问期间把文献 [17], [19], [20], [21] 中的关于模糊集合的上、下近似算子总结推广为以下四种算子[22]:

(1) T-上近似算子: $\overline{R_T}A(x)=\sup_{u\in U}T(R(x,u),A(u))$.

(2) S-下近似算子: $\underline{R_S}A(x)=\inf_{u\in U}S(N(R(x,u)),A(u))$.

(3) σ-上近似算子: $\overline{R_\sigma}A(x)=\sup_{u\in U}\sigma(N(R(x,u)),A(u))$.

(4) ϑ-下近似算子: $\underline{R_\vartheta}A(x)=\inf_{u\in U}\vartheta(R(x,u),A(u))$.

文献 [22] 首先在无穷论域上利用构造性和公理化的方法对以上四种算子进行了总结性的研究, 指出了即便是在有限论域模糊粗糙集与经典的粗糙集在基本性质上也有着本质区别, 继而对其格结构和拓扑结构进行了刻画.

对模糊集合的上、下近似算子的研究在相当长的时间内是模糊粗糙集的主流研究方向. 在以上提到的工作的基础上, 许多学者在这一问题上又进行了推广性的研究, 这些研究包括对公理化方法中公理集合的刻画[23−27], 把论域推广到更宽泛的框架, 如格和双论域等[28−42]. 由于这些研究基本上没有超出前述的研究工作的范畴, 这里就不再详细综述了, 有兴趣的读者可以参考相关的文献.

以上这些对模糊粗糙集的研究都是利用模糊集合的隶属函数作为表达近似算子的工具, 尽管这些算子绝大多数都以经典粗糙集中的近似算子为特例, 但是模糊粗糙集到底如何推广了经典粗糙集并不是十分清楚, 模糊粗糙集是否具有经典粗糙集那样的粒结构是一个值得研究的问题. 文献 [43] 和 [44] 对这个问题给出了肯定的回答, 指出了在模糊粗糙集的框架内哪种模糊集合可以起到经典粗糙集中与等价

类相似的作用, 证明了前述的四种算子具有与经典粗糙集类似的粒结构, 说明了模糊粗糙集从粒结构的角度同样推广了经典粗糙集.

以上我们简要回顾了模糊粗糙集基本模型的提出和发展历程. 事实上在模糊粗糙集的诞生开始除对各种模糊粗糙集模型的研究之外, 对其他方面的课题研究也逐渐开展起来. 其中得到最多关注的就是基于模糊粗糙集的属性约简的研究. 第一次把模糊粗糙集用于特征选择是 1992 年首次在文献 [46] 中提出的. 文献 [46] 利用弱模糊划分定义了模糊正域、负域和边界域, 并且以此为基础提出了衡量属性集合重要度的指标, 根据这个指标提出了特征选择的思想. 但是由于在当时模糊粗糙集刚刚诞生, 并且这个思想不是基于模糊相似关系的模糊粗糙集给出的, 因而没有引起充分的关注. 目前关于这方面的研究的绝大多数文献中都把基于模糊粗糙集的属性约简的首次提出都归结于 2004 年发表的文献 [45], 事实上这个说法是不尽准确的, 不过一个不争的事实是基于模糊粗糙集的属性约简受到广泛关注确实是从文献 [45] 开始的, 目前几乎所有的关于基于模糊粗糙集的属性约简的研究也都是以文献 [45] 为起点. 下面我们仍然采取原封不动的方式来介绍文献 [45] 中的属性约简的算法.

设 U 是非空有限论域, P 和 Q 是 U 上的模糊等价关系, Q 的 P 正域定义为 $\mu_{\mathrm{Pos}_P(Q)}(x)=\sup_{X\in U/Q}\mu_{\underline{P}X}(x)$, $x\in U$, Q 相对于 P 的依赖函数定义为

$$\gamma'_P(Q)=\frac{\left|\mu_{\mathrm{Pos}_P(Q)}(x)\right|}{|U|}=\frac{\sum_{x\in U}\mu_{\mathrm{Pos}_P(Q)}(x)}{|U|},$$

这里 $\underline{P}$ 定义为 $\mu_{\underline{P}X}(x)=\sup_{F\in U/P}\min\{\mu_F(x)\}$, $\inf_{y\in U}\max\{1-\mu_F(y),\mu_X(y)\}$. 尽管在文献 [45] 指出这个定义引自文献 [10], 但是事实上它与文献 [10] 中的下近似算子并不一致, 可以举出反例来说明 $X\subset\underline{P}X$ 是可能成立的. 对于两个模糊相似关系集合 $P=\{a,b\}$, U/P 定义为 $U/\mathrm{IND}\{a\}$ 和 $U/\mathrm{IND}\{b\}$ 的笛卡儿乘积. 比如, 若 $U/\mathrm{IND}\{a\}=\{N_a,Z_a\}$, $U/\mathrm{IND}\{b\}=\{N_b,Z_b\}$, 则 $U/P=\{N_a\cap N_b,N_a\cap Z_b,Z_a\cap N_b,Z_a\cap Z_b\}$. 利用依赖函数 γ' 可以来计算所有属性集合 C 的约简 R. 如果对任意的 $x\in C-R$ 有 $\gamma'(R)=\gamma'(R\cup\{x\})$, 则 R 是 C 的一个约简. 这个思想显然是从经典粗糙集的属性约简的方法中借鉴过来的. 从这个思想出发文献 [45] 设计了一个快速算法来计算属性约简, 并且以此为基础发表了一系列文章[47−52]. 由于文献 [45] 中的方法完全是从形式上把经典粗糙集中相应的方法甚至符号照搬过来, 对基于模糊粗糙集的属性约简的本质没有清楚地认识, 因而所提出的算法有很多问题. 一个最关键的问题就是由于采用了一个不合理的模糊集合的下近似的定义, 导致了依赖函数 γ' 随着属性个数的减少不具有单调性, 直接的后果就是设计的属性约简算法不收敛. 另一个问题就是 $U/\mathrm{IND}\{a\}$ 和 $U/\mathrm{IND}\{b\}$ 的笛卡儿乘积并不能保证其是一个模糊等价关系的等价类的集合, 尽管这一点对经典的等价关系是成立的. 由此可见, 单纯地简单模仿经典粗糙集中的方法而不去探究属性约简的本质特征肯

定会带来许多问题.

事实上, 文献 [45] 中的算法提出之后不久就有研究者发现其不收敛的问题. 文献 [53] 和 [54] 在通过实验发现这个问题之后, 提出了一种紧邻域的改进策略, 提高了算法的运行效率. 但是他们仍然采取了相同的下近似的定义, 因而没有解决依赖函数不单调的问题. 文献 [55] 和 [56] 采取了合适的模糊粗糙集下近似定义提出了若干种度量以设计计算属性约简的启发式算法, 保证了算法的收敛性. 这些属性约简的方法都是针对文献 [11] 中基于 T_M 的模糊粗糙集提出的, 都是以设计启发式的算法为主, 许多关于属性约简的重要课题没有涉及, 比如没有提出核心的概念.

在从理论角度对文献 [45] 的算法进行了仔细分析之后, 文献 [57] 从数学角度对基于 T_M 的模糊粗糙集的属性约简给出了严格的定义, 并且利用模糊粗糙集的粒结构研究了约简的结构, 设计了基于辨识矩阵的计算约简的算法, 文献 [59] 又利用文献 [60] 中的样本对选择的思想设计了计算约简的快速算法. 需要指出的是文献 [45] 的作者在文献 [58] 中对其算法所存在的问题进行了辩解, 他们回避了对文献 [10] 中概念的引用错误, 认为依赖函数不单调是因为去除了噪声所致. 但是在其所有的关于属性约简的文献中都没有明确地提出其算法具有去除噪声的理论机制, 包括如何判断那些条件属性是噪声以及如何去除噪声等, 更没有对其不收敛的算法进行改进.

对基于一般的模糊粗糙集的属性约简的研究是从文献 [61] 开始的, 通过模糊粗糙集的粒结构引入了辨识矩阵的方法来计算属性约简; 文献 [62] 研究了基于不同的模糊粗糙集模型的属性约简的关系, 文献 [63] 研究了回归问题的属性约简理论与方法, 文献 [64]~[68] 也从不同的角度研究了基于模糊粗糙集的属性约简的算法与应用. 可以不夸张地总结说基于模糊粗糙集的属性约简是模糊粗糙集主要的应用领域, 同时也为模糊集理论提供了一个新的应用面向.

自从粗糙集理论诞生以来, 尽管与一些其他的理论如证据理论和模态逻辑等进行了结合性的研究, 但是实际上其发展一直是相对独立于主流的机器学习理论与方法, 对于其他的机器学习方法粗糙集更多地是作为数据预处理的方法被独立使用, 很少有真正成功的交叉结合研究. 相同的局面也发生在模糊粗糙集的研究中, 这极大地阻碍了其与主流机器学习领域之间的相互交流和借鉴, 同时也限制了模糊粗糙集研究内容的深入和拓展. 比如, 在 2005 年之前相当长的一段时间内对模糊粗糙集的研究集中在模糊近似算子的拓展方面, 这些研究大多数都是为了拓展而拓展, 得到的结果大同小异. 由于这些拓展没有明确的目的性和应用背景, 对所提出的算子缺乏合理的解释, 因而也很难深入地研究下去. 尽管基于模糊粗糙集的属性约简拓宽了模糊粗糙集理论的研究范围, 但是仍然没有走出粗糙集的研究范畴.

2006 年 B. Moser 同时发表了两篇文章[69,70], 指出了在单位区间上取值且对角线取值为 1 的核函数都是特殊的模糊 T-相似关系, 并且常用的三种模糊相似关系

都是半正定的, 这样就建立了模糊相似关系与核函数之间的紧密联系. 由于模糊相似关系是模糊粗糙集理论中最基本的概念, 因而尽管文献 [69] 和 [70] 与模糊粗糙集没有直接的关系, 其仍然启发我们可以从核函数的角度去研究基于核函数的模糊粗糙集. 对这方面的研究是从把 Gauss 函数作为特殊的模糊相似关系开始的. 文献 [71] 首先提出了基于 Gauss 函数的模糊集合的近似算子, 并研究了基于 Gauss 函数的属性约简的理论与算法, 文献 [72] 和 [73] 研究了基于 Gauss 函数的模糊粗糙集的性质, 文献 [74] 利用基于 Gauss 函数的模糊粗糙集改进了硬间隔的支撑向量机算法. 由于一般的模糊相似关系不能保证其正定性, 因而很难从理论上把模糊粗糙集理论完全地纳入到核方法的框架之内. 借鉴了文献 [75] 中关于非正定核函数特征空间几何解释的思想, 文献 [76] 首次把模糊相似关系归结为 Krein 核函数, 在 Krein 空间中对模糊集合的近似算子进行了几何解释, 并且改进了软间隔的支撑向量机. 文献 [71]~[74] 和 [76] 中的工作从核函数的角度对模糊粗糙集中的基本概念进行了数学解释, 初步建立了模糊粗糙集与核方法之间的联系, 拓宽了模糊粗糙集的研究思路.

模糊粗糙集的另一个重要的研究内容就是模糊粗糙集的数字特征. 根据作者的知识, 文献 [45] 中提出的依赖函数 γ' 是最早提出的模糊粗糙集的数字特征. 文献 [77] 定义了基于 T_M 的模糊粗糙集的信息熵和条件熵, 文献 [78] 对基于 T_M 的模糊粗糙集定义了 Zadeh 意义下的模糊集合的信任函数与似然函数. 对一般模糊粗糙集的数字特征的研究是从文献 [79] 开始的, 通过在概率空间上定义了合理的模糊集合的测度, 引入了刻画模糊集合近似算子的信任函数和似然函数, 以此为基础定义了模糊粗糙集的依赖函数、近似精度、信息熵等概念. 文献 [80] 讨论了模糊信任结构与证据理论之间的关系.

根据文献 [76] 中的分析, 在把模糊粗糙集应用于分类问题时, 需要计算样本对所在的决策类的下近似值, 这个下近似值取决于该样本到异类样本的距离的最小值. 如果样本中有噪声存在, 往往会导致存在异类样本与该样本距离很小, 显然这个距离不是鲁棒的. 也就是说在处理分类问题时模糊粗糙集对噪声尤其是类别标记错误的样本十分敏感. 从模糊粗糙集的粒结构来看, 在计算某个模糊集合的下近似时只考虑完全被该模糊集合包含的基本颗粒, 这样的要求显然从实用的角度来说过于严格. 为了解决这个问题, 文献 [81] 和 [83] 提出了变精度模糊粗糙集的概念, 通过在模糊集合的上、下近似的粒结构中放松对基本颗粒完全包含的要求, 重新定义了含参数的模糊粗糙集, 并以此为基础构造了分类器[82], 实验表明这种做法确实使得模糊粗糙集的抗噪性大大提高. 文献 [84]~[86] 借用了机器学习中关于算法鲁棒性的思想, 从改造隶属度的计算入手, 在计算点到集合的距离时忽略一些距离很小的样本, 在忽略的样本数与增大的距离之间取得一个折中, 提出了软距离的概念, 同样使模糊粗糙集具有了较强的抗噪性.