信息科学技术学术著作丛书

应用粗糙计算

胡清华　于达仁 著

科学出版社

信息科学技术学术著作丛书

应用粗糙计算

胡清华　于达仁　著

安徽大学图书馆
藏书
www.lib.ahu.edu.cn

科学出版社

北京

内 容 简 介

本书系统总结了作者近几年在粗糙集理论、模型、算法和应用方面的研究成果，以分类决策中人们普遍使用的若干一致性假设为主线，论述了等价关系、邻域关系、模糊关系以及优势关系下的粒化和近似问题，进而分析了各种关系诱导出来的近似空间的不确定性度量问题。本书的特点是理论分析、算法设计和实际应用相结合，将粗糙集理论应用于模式识别、机器学习和数据挖掘的算法设计，形成了特征依赖性分析、特征选择、属性约简、样本约简以及规则学习等算法。

本书补充了集合论的基础知识，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，也可作为决策科学和信息科学领域的研究人员与工程人员的参考书。

图书在版编目(CIP)数据

应用粗糙计算/胡清华，于达仁著. —北京：科学出版社，2012
(信息科学技术学术著作丛书)
ISBN 978-7-03-034795-4

Ⅰ.①应… Ⅱ.①胡…②于… Ⅲ.①计算机算法 Ⅳ.①TP301.6

中国版本图书馆 CIP 数据核字(2012)第 125282 号

责任编辑：刘宝莉 / 责任校对：陈玉凤
责任印制：张 倩 / 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

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2012 年 6 月第 一 版 开本：B5(720×1000)
2012 年 6 月第一次印刷 印张：12 1/4
字数：232 000

定价：60.00 元

(如有印装质量问题，我社负责调换)

《信息科学技术学术著作丛书》序

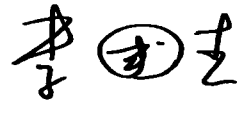
21 世纪是信息科学技术发生深刻变革的时代，一场以网络科学、高性能计算和仿真、智能科学、计算思维为特征的信息科学革命正在兴起。信息科学技术正在逐步融入各个应用领域并与生物、纳米、认知等交织在一起，悄然改变着我们的生活方式。信息科学技术已经成为人类社会进步过程中发展最快、交叉渗透性最强、应用面最广的关键技术。

如何进一步推动我国信息科学技术的研究与发展；如何将信息技术发展的新理论、新方法与研究成果转化为社会发展的新动力；如何抓住信息技术深刻发展变革的机遇，提升我国自主创新和可持续发展的能力？这些问题的解答都离不开我国科技工作者和工程技术人员的求索和艰辛付出。为这些科技工作者和工程技术人员提供一个良好的出版环境和平台，将这些科技成就迅速转化为智力成果，将对我国信息科学技术的发展起到重要的推动作用。

《信息科学技术学术著作丛书》是科学出版社在广泛征求专家意见的基础上，经过长期考察、反复论证之后组织出版的。这套丛书旨在传播网络科学和未来网络技术，微电子、光电子和量子信息技术，超级计算机、软件和信息存储技术，数据知识化和基于知识处理的未来信息服务业，低成本信息化和用信息技术提升传统产业，智能与认知科学、生物信息学、社会信息学等前沿交叉科学，信息科学基础理论，信息安全等几个未来信息科学技术重点发展领域的优秀科研成果。丛书力争起点高、内容新、导向性强，具有一定的原创性；体现出科学出版社“高层次、高质量、高水平”的特色和“严肃、严密、严格”的优良作风。

希望这套丛书的出版，能为我国信息科学技术的发展、创新和突破带来一些启迪和帮助。同时，欢迎广大读者提出好的建议，以促进和完善丛书的出版工作。

中国科学院计算技术研究所所长



前　言

从生产和科学实验的数据中提炼一般规律是知识发现的主要途径。当前许多大型生产和科研活动的信息都存储在数据库中，在一些领域出现了容量庞大的数据库系统，这些数据蕴涵了丰富的有用信息。如何从海量的、不一致的、不精确的、结构多样的数据中提炼有价值的、新颖的知识成为信息科学和智能科学面临的共同挑战。

模式分类是智能活动的主要形式，从数据中抽象分类模型是机器学习和数据挖掘研究的核心内容，而不一致性样本是分类学习面临的主要困难。粗糙集是波兰学者 Pawlak 于 20 世纪 80 年代初提出来的一种刻画分类问题中不一致性的数学工具。该方法在近十余年里得到了迅速发展，成为机器学习领域十分活跃的一个分支，是归纳学习的一类重要范式。

Pawlak 粗糙集理论的核心概念是基于等价关系的粒化和基于下、上近似的逼近，这两个概念构成了 Pawlak 粗糙集理论的基石。基于等价关系的粒化类似于人脑认知过程的概念生成，而用已知的概念去近似刻画另一类事物是人脑思维的另一个重要特点，我们往往采用近似而不是精确的方式用已知的概念去近似刻画另一类事物。

Pawlak 粗糙集理论模拟了人脑认知和思维中的粒化和近似两大特点，但这只是人脑思维的简单模型。想象一下我们的语言概念系统就会发现，大脑思维中的基本概念并不一定是由等价关系生成的互斥的信息粒，而是由属性相似、距离相近或者功能一致等复杂关系形成的交叠的和分层的概念集。基于等价关系的粗糙集模型仅仅是人脑思维的积木世界，在感知和现实世界里都存在十分复杂的粒化结构和近似形式，将粗糙集理论应用于复杂问题求解就必须拓展粗糙集理论中这些基本定义。

实际应用中的分类任务大都是由数值变量、时间序列或者图像等复杂数据形式描述的。本书将这些任务抽象为广义度量空间的分类建模问题，研究了度量空间的邻域粒化结构、模糊粒化结构和序结构，拓展了经典粗糙集中的信息粒化形式和相应的近似推理算子。进而抽象出人脑思维中存在的六种决策一致性假设，并基于粗糙集理论分别建立了这些一致性假设的数学模型，从而为复杂分类任务的一致性刻画建立了数学工具，拓展了粗糙集理论的应用范围。

本书注重从实际应用中抽象理论问题，进而基于理论模型开发实用的学习算

此为试读，需要完整PDF请访问：www.ertongbook.com

法，从而将理论和应用紧密结合起来。此外，本书将粗糙集理论放在机器学习和模式识别这个框架下进行讨论，为理解粗糙集模型提供了一些新的视角。希望这些内容能构成本书区别于现有的几本粗糙集相关专著的特色。

全书共分七章，第1章综述粗糙集研究的现状和存在的问题；第2章介绍集合论和模糊集合论的基本知识；第3～6章分别介绍经典粗糙集模型、邻域粗糙集模型、模糊粗糙集模型、优势关系粗糙集模型及其应用，建立了六类一致性假设的数学模型；第7章给出这六类一致性假设的信息熵模型。

本书的工作得到很多专家和同行的帮助。吴从炘教授将我们引入一个十分宽广的学术领域；郭茂祖教授、陈德刚教授、王熙照教授、吴伟志教授、梁吉业教授、李德玉教授、钱宇华教授和米据生教授在研究过程中给予了大力帮助；与姚一豫教授和 Witold Pedrycz 教授的合作加速了研究的进程；刘金福、谢宗霞、安爽、朱鹏飞和车勋建协助完成了部分研究工作。此外，本书还收录了其他作者发表在期刊、会议论文中的一些重要成果，在此一并表示感谢！

本书相关研究受到国家杰出青年科学基金(50925625)、国家自然科学基金面上项目(60703013、10978011、61105054)和国家博士后基金资助。

由于作者水平有限，书中难免存在不足之处，敬请读者批评指正。

目　录

第1章 绪 论

1.1 复杂数据的知识发现

知识是人类认识和改造客观世界的结果，也是推动人类社会进步的动力。从生产和科学实践的经验中发现和提炼一般规律是知识发现的主要途径。当前，由于信息技术的发展以及大型生产和科研活动的开展，在许多领域，如深空探测、基因分析、社会调查和工业过程监控等领域，出现了容量庞大的数据系统。这些数据蕴涵了丰富的有用信息。如何从海量的数据中获取有价值的、新颖的知识成为信息学科一个重要的研究课题。

显然，通过人工分析大规模数据是不现实的，研究利用计算机从海量数据中自动发现知识无论对科学研究还是社会生产生活都具有重要的价值。美国 Mjolsness 和 DeCoste 在 *Science* 杂志上系统分析了机器学习和知识发现技术在科学研究各个阶段扮演的角色，认为机器学习和知识发现技术能够在各个方面协助研究人员加速科研进程。自动知识发现技术在信息检索、图像理解、文本分类、工业过程监测和故障诊断等领域正在发挥越来越重要的作用。

计算机自动知识发现面临的主要困难是信息的多样性、不确定性和不一致性。其中，多样性表现为数据结构的多样性和数据值域的多样性。在工业监控和金融分析等领域存在大量时序数据库；在天文观测、资源探测、城市规划和交通管理等领域存在大量的空间数据库；在网络环境中存在海量文本、图像以及声像等非结构化数据。数据值域的多样性则表现为描述对象的属性的值域是复杂多样的，可分为名义值、整型值、实数值、模糊值、集值、区间值等。有时部分对象的某些特征的值还是缺失的。大量结构复杂、形式多样的数据给知识发现带来了挑战。本书将集中分析关系型数据库中多种类型变量共存时的分类知识发现问题。

分类学习是知识发现的一大类任务。在实际应用中，描述分类的属性往往不是单一类型的，而是多种类型的变量共存的。以美国加州大学机器学习与智能系统研究中心收集的分类学习测试数据(http://archive.ics.uci.edu/ml/index.html)为统计对象，发现无论在物理学、生命科学、医疗诊断领域还是在社会统计、金融分析、信息安全和设备健康监测等领域都存在大量由符号变量和数值变量共同描述的分类任务。

在医疗诊断中，如心律不齐、心脏病、乳腺癌、肝炎、甲状腺病、皮肤病的自动分类学习都涉及混合数据问题。由匈牙利心脏病研究所 Janosi、瑞士苏黎世医学院 Steinbrunn 等收集的 4 个心脏病数据包含 900 多个病例的 76 个特征，这些特征既有如性别、胸部疼痛位置、胸部疼痛类型、是否抽烟等符号属性，也包含年龄、血压、血糖浓度、最大心跳数、抽烟史等数值属性。

在物理学研究领域，由 Sterling 和 Buntine 提供的钢材退火数据集“Annealing”记录了 798 次退火试验，每次试验由 5 个数值变量和 33 个离散变量描述，其中部分属性值没有记录，标记为“?”。

在社会统计领域，由 Becker 从美国 1994 年人口统计数据中抽取的 48 854 人的年收入调查表中包含 5 个表示年龄、每周工作时间、资本收益情况等连续变量和 9 个表示工作类型、受教育程度、婚姻状况、职业、社会关系、人种、出生国等符号变量。由美国统计局收集，Lane 和 Kohavi 公布的人口收入调查数据集则包含 199 523 人的收入记录，每人由 7 个连续变量、33 个符号变量描述。

商业领域的信誉卡发放评估和产品营销数据库中往往数值属性和符号属性共存。例如，德国信誉卡数据集由 7 个数值属性，13 个定性属性描述；澳大利亚信誉卡分析数据则采集了 6 个连续属性和 9 个离散变量。

此外，在天文物理研究中的空间天气预报、地震预报分析、电力设备的故障诊断、股票市场分析等大型复杂决策问题中，所需处理的数据都混合了名义、数值、区间值和模糊值的变量。研究混合数据知识发现的模型和算法无论就知识发现的理论研究，还是许多领域的应用需求都具有重要的价值。

1.2 混合数据分类建模的不确定性分析

分类是人类智能行为的一种主要形式，从大量样本数据中发现分类知识、建立分类模型是知识发现的一大类任务。分类学习面临的主要困难是数据样本中的随机性、模糊性和不一致性给分类学习带来的不确定性因素。

1.2.1 数据类型及其信息结构分析

一般而言，给定的分类学习样本由一个数据矩阵描述。矩阵的一行$(a_{j1}, a_{j2}, \cdots, a_{ji}, \cdots, a_{jN}, d_j)$记录了一个学习样本，其中 a_{ji} 表示第 j 个样本在 i 个特征上的取值，d_j 则是第 j 个样本的决策。混合分类数据是指属性的值域以及决策的值域是多种形式的。广义上讲，属性的值域可以为图像、声音、文本、矩阵、时序、数值和字符；狭义上讲，属性的值域为名义型字符、有序型字符、数值量、区间值或者模糊值。本书研究的问题是指描述分类问题的属性为符号型、数值型、区间值或者模

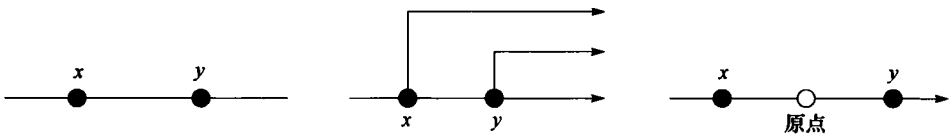
此为试读，需要完整PDF请访问：www.ertongbook.com

糊型的混合数据。

不同类型的数据中蕴涵了不同的信息结构，表达了样本之间不同侧面的信息。名义变量由若干个状态表示，这些状态之间既没有数量关系，也没有等级的序关系，如对象的性别、颜色。有序符号变量则由有序的若干等级来表示，变量的各值之间存在全序或者偏序结构，如国民的受教育程度，由符号描述的年龄、身体状态等。名义变量和有序符号变量有时候统称为符号变量、离散变量或者定性变量等。数值变量在实数或者实数的一个子集上取值，如描述病人状态的变量：血压、血糖浓度、体温等和描述设备状态的变量：温度、流量、压力、流速等。在某些特殊的场合，人们还会采用区间值和模糊值来描述分类问题，如股票每日的开盘价、收盘价所构成的区间、最低价和最高价构成的区间，每日气温的最低温度和最高温度构成的区间等。

在分类学习中，当决策的类别为名义变量时，称之为一般的分类问题。在某些情况下，决策值之间存在序结构，则称之为有序决策问题或者排序问题，如投稿的决策：录用、修改和拒稿。

显然，不同的属性中蕴涵的信息结构和可以实施的运算是不同的。就单个属性而言，可以用数轴等价的分析属性的信息结构，如图 1.1 所示。



(a) 名义属性描述的样本　(b) 有序离散属性描述的样本　(c) 数值属性描述的样本

图 1.1 不同属性空间的信息结构

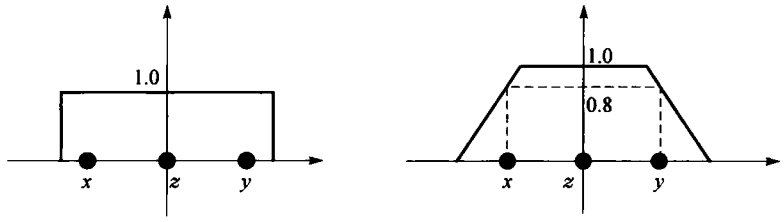
在名义属性描述的信息空间中，对象 x 和对象 y 要么相等，要么不等，不存在任何过渡的情况，也就是说，样本的取值不是连续的。在名义属性空间中，可以定义离散距离函数

$$\Delta(x,y)=\begin{cases}1, & x\neq y\\ 0, & x=y\end{cases}$$

在有序离散属性空间中的对象，不仅可以知道对象 x 和对象 y 是否相等，而且还知道 x 和 y 的大小关系。在某些情况下，将有序变量按照序结构转化为一串整数，称之为整型变量，那么实数域上的某些运算也是有意义的。

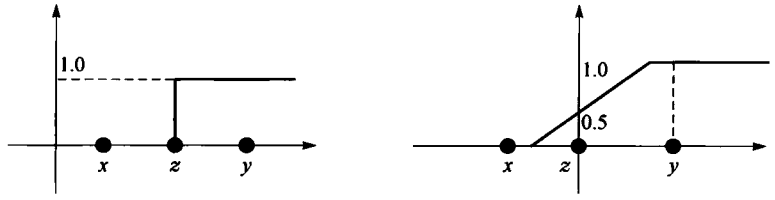
在数值属性上可以实施实数域上的各种数学运算，实数域上的运算也将诱导数值空间的各种结构。首先，实数是有序的，我们可以比较不同对象在某一特征上的大小，因此实数域中的对象可以建立序结构；其次，实数域是连续的，对象之

间的距离可以在实数域上取任一值，实数域的连续性使得对象之间存在邻域结构。由于实数轴的连续性，数轴上的各点的邻域构成的邻域簇相互交叠、关联，形成了实数空间的覆盖。如果在同一属性空间上允许不同对象的邻域大小不同，在不同的维度空间上也采用不同的邻域大小，那么可以形成多维空间中十分复杂的信息结构。最后，实数域的连续性也导致了信息结构的模糊性，我们说对象 x 和对象 y 在某一数值特征上取值相近或者说对象 y 比对象 x 要大得多，这种描述都是模糊的，可以在实数空间中建立模糊邻域和模糊序结构，如图 1.2 和图 1.3 所示。



（a）z的邻域　　（b）z的模糊邻域

图 1.2　对象 z 的邻域



（a）比z大的集合　　（b）远比z大的模糊集

图 1.3　对象的序结构

图 1.2 给出一维实数空间中与对象 z 近邻的经典子集和模糊子集。如果采用不对称模糊邻域和不同维度上大小不一致的模糊邻域，那么可以得到形式多样、结构复杂的邻域系统。

图 1.3 给出一维数值空间中比对象 z 大的经典子集和远比对象 z 大的模糊子集。在图 1.3(b)中对象 y 远比 z 大的隶属度为 1，而对象 x 远比 z 大的隶属度为 0。基于不同的模糊隶属度函数，可以得到一系列远比 z 大的模糊集。

当名义变量、有序变量和数值变量共存于某一分类问题中，那么不同类型变量蕴涵的信息结构的组合就更丰富了。这些信息结构在不同的领域、不同的应用中反映了问题的不同侧面，全面理解数据中蕴涵的信息需要构造不同的学习算法以发现混合数据中蕴涵的信息结构。

1.2.2 混合数据分类的不确定性分析

处理不确定性是人的认知和推理的重要能力，也是人工智能、机器学习、知识发现和智能决策所面临的主要困难。数据的随机性、信息的不完备性和决策的不一致性是导致这一困难的主要原因。

随机性是指发生在数据采集、传输和存储过程中的偶然因素引起数据偏离其真实值的现象。在实验分析和数据录入过程中，由于记录或打字错误，最终分析的数据中难免会存在错误信息。在基于传感器系统的对象监测中，由于环境的干扰、传感器分辨精度有限或者传感器失效，也会在测量的真实值中引入噪声信息。由此看来，不仅描述对象的符号信息是不精确的，对象的数值属性描述也不一定反映对象的真实状态。由精确数值描述的对象尽管看起来似乎得到了精确描述，但由于数据测量的随机性使得记录的数值在一定邻域内的值都可能是该对象的真值。随机性是知识发现和分类建模中引起不确定性的重要原因。

信息的不完备性体现在三个方面：其一，由于某种原因，部分对象的某些特征值没有被测量和记录，因此在数据矩阵中出现了空缺；其二，描述对象分类的某些重要特征没有采集和存储，使得利用现有的特征不足以区分各类对象，从而出现类重叠区域；其三，反映分类模型的某些模式没有被激发，使得样本信息不足以完整反演出分类模型。由于分类学习往往是基于现有的样本信息估计分类模型，因此不考虑第二和第三类信息不完备性。

不一致性是分类建模复杂性的又一重要因素。回想一下人的决策过程，不难发现在面向不同类型属性描述的决策问题时人采用了不同的一致性分析策略。

对于由名义型属性描述的分类问题，用户自然希望特征相同的对象归属于相同的决策类别，否则这些对象将不可区分。因此在分类建模时，使用了一个潜在的假设，即条件属性取值相同的对象，其决策应该相同，否则决策是不一致的，称为第一决策一致性假设。

当对象由数值属性描述时，不同的对象获得完全相同的属性描述的概率很低，精确的属性值在决策中将会被自然地泛化到一个区间，而不再以数值进行推理。此时人们将第一分类一致性假设泛化为：数值属性取值相近的对象，其分类应该相同，否则分类是不一致的。我们称之为第二决策一致性假设。当数值和符号属性同时存在时，人们希望符号属性取值相同，数值属性取值相近的对象应该被归于同一类别，否则分类是不一致的，称之为第三决策一致性假设。

在模糊理论中，不仅需要考察对象的分类是否一致，而且还需要计算对象分类的一致性程度。此时，在考察对象的分类一致性时以对象模糊邻域内的其他对象为参考，如果模糊邻域内的对象分类相同则分类是一致的，否则对象分类的不

一致性程度由该对象到最近的异类对象的距离决定，距离越大，则分类一致性程度越高，称之为第四决策一致性假设。

以上四个分类一致性假设都是针对一般的分类问题而言的，面对有序决策问题时，人们使用了另外一个一致性假设：当两个对象其他属性完全相同时，如果对象 x 就属性 a 而言比对象 y 占优，那么 x 的决策应该至少不比 y 差，否则决策是不一致的。我们称之为第五决策一致性假设。举例而言，当投稿 x 在论文写作质量方面与投稿 y 相当，但 x 在原创性水平方面比 y 高，如果投稿 x 被录用，那么投稿 y 自然也应该被录用，否则违背了论文评判的一致性。

在决策建模中，所面临的问题往往是不一致的。此时不仅需要知道对象的决策是否是不一致的，而且还需要精确刻画有序决策中的不一致性程度。当两个对象其他属性完全相同时，如果对象 x 就属性 a 而言比对象 y 占优，但是 x 的决策比 y 差，显然，此时决策是不一致的。不一致的程度由 x 就属性 a 而言比对象 y 占优的程度决定，占优的程度越大，决策的不一致性也就越大。我们称此为第六决策一致性假设。

不一致性是分类学习面临的主要挑战，大量的学习算法所解决的就是分类不一致情况下的最优决策问题，如贝叶斯最优决策，决策树叶子节点上样本类别不一致时的处理，软间隔支持向量机。

知识发现就是要从随机的、不完备的和不一致的信息中提炼一般规律。随机性、不完备性和不一致性是干扰分类学习算法、影响分类建模精度和泛化能力的主要因素，这三类不确定往往同时存在于分类数据中，因此混合数据的知识发现模型必须能够刻画和处理数据中的不确定性。

1.3 基于粗糙集的分类不确定性刻画

从分类不确定分析中可知，不一致信息是分类学习面临的主要困难，粗糙集理论是描述和处理分类不一致的有效方法。人在认识分类的不一致性时需要考察的是对象的粒子，而不是对象个体。知识是建立在粒化和概念的基础上的，信息粒化使得人具有在部分精确(随机性)、部分已知(不完备性)、部分一致的(不一致性)情况下做出合理决策的能力。

粗糙集正是模拟人的思维的这一特点而发展起来的粒计算模型。该理论是波兰学者 Pawlak 于 20 世纪 80 年代初提出用以刻画由不精确信息描述的分类问题中不一致性的数学工具。该方法在近十余年里得到迅速发展，成为机器学习领域十分活跃的一个分支，在属性依赖性分析、特征子集选择和约简、分类知识发现等方面取得了成功。

1.3.1 粗糙计算模型的发展

由符号数据描述的分类问题中，训练样本被条件属性和决策划分为两组等价类。分类可以理解为由条件属性形成的等价类去描述由决策属性生成的决策概念。在此过程中，有时决策概念可以写成一组条件属性生成的等价类的并。此时称决策概念是一致的，或者说可定义的。但某些决策概念不能被条件属性形成的等价类精确描述，部分等价类内的样本来自多个决策。此时，分类是不一致的，称此决策概念是粗糙的。对于粗糙的决策概念，粗糙集理论定义了一对下、上近似集合近似描述此决策。下近似是被该决策概念完全包含的等价类，而上近似是与决策概念交集不为空的等价类的并集。上、下近似的差称为决策的边界。边界就是那些条件属性值相同，但决策不同的等价类的并。显然，边界上的样本违背了第一分类一致性假设。

粗糙集理论以精确的数学形式刻画了分类问题的不一致性，并且认为分类边界是造成分类复杂的原因。这一观点与其他学习理论和方法是一致的。粗糙集定义保持分类正域不变的最小特征子集为系统的约简。约简在保持系统分类一致性不变的前提下获得分类模型的最小表达。粗糙集对分类不确定性的度量是通过分类样本直接计算出来的，不需要任何先验信息，因此粗糙集理论被认为是一种客观的处理分类不确定性的数学工具。同时，粗糙集分析得到的结果为产生式规则，容易被理解和使用，因此得到应用数学和人工智能领域研究人员的广泛关注。

Pawlak 粗糙集理论的核心概念是基于等价关系的粒化和基于下、上近似的逼近。这两个概念构成了 Pawlak 粗糙集理论的基石。基于等价关系的粒化相当于人的认知过程的概念生成，人在认识客观世界时，根据对象的特性将对象分成不同的子集，并将每个子集标记为某一单词，从而形成抽象的概念。单词抽象了对象的共性，忽略了对象的细节，因此基于单词的描述具有鲁棒性和泛化能力。随着人的认知能力的增强，人类描述对象的词汇也越来越丰富、对对象的刻画也越来越精确。这可类比于粗糙计算中随着属性的增加，划分变得更精细这一过程。逼近是人类思维的另一个重要特点。由于认知能力有限和为了描述的鲁棒性，人们往往采用近似的方式而不是精确的方式用已知的概念去刻画另一类事物。

Pawlak 粗糙集理论模拟了人思维中的粒化和近似两大特点，但是该模型只是人们思维的简单模型。回想一下我们的语言概念系统就会发现，我们思维中的基本概念并不一定是由等价关系生成的互斥的对象子集，而是由属性相似、距离相近或者功能一致等复杂关系形成的交叠的和分层的对象集合，其次语言中的概念

往往是模糊的，而不是清晰的。基于等价关系的粗糙集模型仅仅是人思维的积木世界，在我们思维和现实世界里存在十分复杂的粒化结构和近似形式，将粗糙集理论应用于复杂问题分析就必须拓展粗糙集理论中某些基本概念。

粗糙集模型的泛化也存在粒化和近似两个维度。首先讨论在近似方式方面的拓展。1992 年，Yao 和 Wong 等提出决策理论粗糙集，该模型不再以决策正域为评价机制，而是综合考虑各类决策错误代价定义属性重要度和约简。1993 年，Ziarko 发现 Pawlak 粗糙集模型中上、下近似的定义缺乏对随机性进行刻画的机制，从而出现不能容忍噪声的问题。经典的下、上近似要求条件属性形成的等价类要么完全包含于粗糙集，要么与粗糙集的交集为空。但现实应用中的数据由于各种不确定性因素不可避免地会存在一定程度的噪声。于是 Ziarko 引入了可变精度粗糙集(variable precision rough set，VPRS)模型。该模型通过包含度阈值允许一定程度的噪声存在，只需要等价类绝大部分被粗糙集包含或者绝大部分不被包含。

尽管可变精度粗糙集模型通过包含度阈值引入噪声容忍机制，但这只是一种工程化的处理手法。1999 年和 2005 年，VPRS 被进一步泛化为概率粗糙集模型和贝叶斯(Bayes)粗糙集模型，从而将随机性和粗糙性两种处理不确定信息的数学工具结合起来。

从粒化方式方面，1990 年 Dubois 和 Prade 首先将 Pawlak 粗糙集中的等价关系拓展为模糊等价关系，提出模糊粗糙集和粗糙模糊集的概念，实现了用模糊集逼近模糊集的推理方式。模糊等价关系要求生成模糊粒化结构的关系满足自反性 $R(x,x)=1$、对称性 $R(x,y)=R(y,x)$和传递性 $R(x,z)\geqslant\min(R(x,y),R(y,z))$。在实际应用中，从数据中计算模糊等价关系并不是一个简单的任务，而且可能不符合数据中的信息结构。1998 年，Morsi 和 Yakout 引入模糊 T 等价关系、三角范数 T 及其诱导的剩余蕴涵算子 θ 拓展了模糊粗糙集的定义。2004 年，Mi 和 Zhang 等利用蕴涵算子 θ 和其诱导的 δ 算子给出广义的模糊粗糙集定义。2005 年，Yeung 和 Chen 等将这些模糊下、上近似算子进行了归纳，并给出公理化描述。2008 年，Li、Leung 和 Zhang 等再一次泛化了模糊粗糙集模型。在新模型中，无需计算对象之间的关系，只需生成论域的模糊覆盖，这一泛化开拓了模糊粗糙集理论的新应用领域。

此外，在经典集合的框架下，Pawlak 粗糙集也被进行了多种拓展。首先，为了处理数据中的遗失值，1997～2000 年研究了相似关系粗糙集，样本的遗失值被认为可能取任意的特征值，因此该样本被分到所有的相似类中。为了处理数值数据，Lin 于 1989 年提出邻域系统的概念用于数据库的近似检索，Lin、Liu 和 Huang 等于 1990 年提出邻域粗糙集的概念。1998 年，Yao 研究了邻域近似算子

此为试读，需要完整PDF请访问：www.ertongbook.com