菅利荣 著

面向不确定性决策的杂合粗糙集方法及其应用

科学出版社
www.sciencep.com

面向不确定性决策的杂合粗糙集方法及其应用

管利荣 著

科学出版社

北京

内 容 简 介

本书介绍了粗糙集理论、方法与应用，并针对实际应用领域中知识表示系统可能包含的多种不确定性情况，较系统地介绍了粗糙集理论与其他相关软技术理论的杂合方法与应用。书中融入了国内外学者研究的许多最新成果。全书内容分为八章，包括绪论，粗糙集理论，粗糙集与概率论的杂合，粗糙集与优势关系的杂合，粗糙集与模糊集的杂合，粗糙集与灰色系统的杂合，变精度粗糙集、模糊集与神经网络的杂合，以及杂合粗糙集方法的应用分析等。

本书可作为高等院校经济管理类专业及应用数学、信息科学、自动控制等专业的高年级本科生及研究生教材，也可作为人文、社会科学及其他相关学科的参考书，还可作为相关企事业单位管理人员、科研机构及工程技术人员等广大研究人员与实际工作者的参考书。

图书在版编目(CIP)数据

面向不确定性决策的杂合粗糙集方法及其应用/菅利荣著. —北京：科学出版社，2008

ISBN 978-7-03-021132-3

Ⅰ.面… Ⅱ.菅… Ⅲ.粗糙集-研究 Ⅳ.O144

中国版本图书馆 CIP 数据核字(2008)第 022864 号

责任编辑：余 丁 / 责任校对：陈玉凤

责任印制：刘士平 / 封面设计：耕者

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

新蕾印刷厂印刷

科学出版社发行 各地新华书店经销

*

2008 年 3 月第 一 版 开本：B5(720×1000)

2008 年 3 月第一次印刷 印张：12 1/4

印数：1—3 000 字数：231 000

定价：38.00 元

(如有印装质量问题，我社负责调换〈环伟〉)

序

现代科学技术在高度分化的基础上高度综合的大趋势，导致了具有方法论意义的系统科学学科群的出现。系统科学揭示了事物之间更为深刻、更具本质性的内在联系，极大地促进了科学技术的整体化进程，许多科学领域中长期难以解决的复杂问题随着系统科学新学科的出现迎刃而解；人们对自然界和客观事物演化规律的认识也由于系统科学新学科的出现而逐步深化。20 世纪 40 年代末期诞生的系统论、信息论及控制论，产生于 20 世纪 60 年代末、70 年代初的耗散结构理论、协同学、突变论、分形理论以及 70 年代中后期相继出现的超循环理论、动力系统理论、泛系理论等都是具有横向性、交叉性的系统科学新学科。

在系统研究中，由于内外扰动的存在和认识水平的局限，人们所得到的信息往往带有某种不确定性。随着科学技术的发展和人类社会的进步，人们对各类系统不确定性的认识逐步深化，不确定性系统的研究也日益深入。20 世纪，在系统科学和系统工程领域，各种不确定性系统理论和方法的不断涌现形成一大景观。如由美国心理学家 McCulloch 和数理逻辑学家 Pitts 于 40 年代首先提出的神经网络，Zadeh 教授于 60 年代创立的模糊集理论，邓聚龙教授于 80 年代创立的灰色系统理论，Pawlak 教授于 80 年代创立的粗糙集理论(rough sets theory)等，都是不确定性系统研究的新兴学科。这些新学科皆源于人类社会实践的迫切需求，从不同角度、不同侧面论述了描述和处理各类不确定性信息的理论和方法，既各具特色，又彼此互补，为人们研究不确定性系统提供了丰富的方法和工具。

不确定性多属性决策是现代决策科学的一个重要组成部分，它在社会、经济、管理及工程设计等众多领域中有着广泛的实际应用背景。一般来说，在不确定性决策中，面对某一特定的应用问题，研究方法和工具的选择是一大难题。将多种技术巧妙杂合，优势互补，建立一种杂合方法是化解这一难题的一个重要的途径，杂合系统的应用效果通常优于单一方法，多种方法的杂合是解决不确定性问题的一个有广阔应用前景的发展方向。菅利荣教授撰写的《面向不确定性决策的杂合粗糙集方法及其应用》，在粗糙集理论与灰色系统、人工神经网络、概率统计、模糊集理论等杂合方法的研究与应用方面做了大量开拓性工作，建立了多种功能强大的杂合软决策方法。其中基于变精度粗糙集的分层知识粒度构造方法、基于优势关系偏好多属性概率决策分析的扩展 VPRS 模型、变精度粗糙模糊集模型和模糊变精度粗糙集模型、灰色变精度粗糙集模型和基于杂合 VPRS 与神经网络的知识发现及预测方法等属于创新性成果。这些方法拓宽了粗糙集理论的适用范围，为经

济、管理中的许多不确定性决策问题提供了多种科学、规范的决策工具。该书立足于实践，旨在解决经济、管理决策实践中可能遇到的多种不确定性问题。书中不仅详细论述了不同软技术杂合思想，而且包含大量从管理决策研究成果中提炼的实际案例。

我有机会与菅利荣教授合作共事，深深地为她奋发向上、不懈追求、吃苦耐劳、热心学术研究的韧劲和精神所感动。数十篇高水平学术论文和如期完成的一批重要科研课题及国家博士后基金项目应是最好的奖励。这本以其博士学位论文和博士后出站报告为基础写成的《面向不确定性决策的杂合粗糙集方法及其应用》一书，融入了菅利荣同志大量心血和汗水。在此，我愿意向从事不确定性多属性决策研究和应用的学者、实际工作者和在读研究生推荐这本有理论、有实践、有创新的著作，相信该书的出版对促进不确定性多属性决策问题研究和应用具有积极意义。

国家有突出贡献的中青年专家
南京航空航天大学特聘教授、博士生导师、经济与管理学院院长
刘思峰
二〇〇七年十二月于南京

前　言

管理活动是由一系列决策组成的，在市场竞争非常激烈的今天，无论企业或个人都经常会遇到复杂的不确定性决策问题，需要分析与处理决策问题中多种不确定性情况，如随机性、模糊性、偏好性、粗糙性和灰性等，且需要快速做出决策。不确定性决策问题普遍存在于管理科学、信息科学、系统科学、计算机科学、知识工程及可靠性技术等众多领域中。粗糙集理论、灰色系统理论、模糊集理论、遗传算法和神经网络等软计算技术及其优势互补地杂合，能够柔性地处理现实生活的不确定性情况，其目标是开采容错不精确性、不确定性、近似推理、部分正确性，以便获得可处理的、功能强大的、低成本的且与人类决策非常类似的解决方法，其主要思想是通过搜寻不确定问题的一种近似解，来设计可接受的低成本解决方案的计算方法。

由波兰科学家 Pawlak 于 1982 年最初提出的粗糙集理论(rough set theory)是处理不精确、不确定与不完全数据的理论，能有效地分析不精确、不一致、不完整等各种不完备的信息。粗糙集理论作为数据推理的强大工具已被成功地应用于知识获取、决策分析、预测、专家系统及数据库中的知识发现等许多领域。粗糙集理论与灰色系统、人工神经网络、概率统计、模糊集等软技术理论有较强的互补性，将粗糙集理论与其他相关的软技术理论两个或多个优势互补地杂合，构建功能更加强大的杂合软决策方法，可以拓宽粗糙集理论的适用范围，为不确定性决策问题提供多种科学、规范的决策方法。

本书介绍了粗糙集理论的方法与应用，并在分析粗糙集理论在实际应用中的优势、劣势及其适用范围的基础上，针对实际应用领域中知识表示系统可能包含的多种不确定性情况，较系统地介绍了粗糙集理论与其他相关软技术理论优势互补的杂合，书中融入了粗糙集理论、方法与应用的最新成果，其中许多内容是作者长期科研与教学成果的凝练。

本书主要特色是强调软计算技术的应用性，尽量减少繁琐数学推导的介绍，在阐述杂合方法理论思想的基础上，用案例来说明杂合方法的具体应用，努力突出粗糙集等软计算技术的理论思想和在实际案例中的应用。

本书可作为高等院校经济管理类专业及应用数学、信息科学、自动控制等专业的高年级本科生及研究生教材，也可作为人文、社会科学及其他相关学科的参考书，还可作为相关企事业单位管理人员、科研机构及工程技术人员等广大研究人员与实际工作者的参考书。

本书的出版得到了中国博士后科学基金和南京航空航天大学精品课程建设基金资助，在此表示衷心地感谢！

粗糙集等软计算技术是新兴的学科，仍处于不断发展、完善过程之中，再加上作者知识水平的局限性，书中难免存在不足之处，恳请有关专家和广大读者批评指正。

目　录

第一章 绪 论

数据挖掘与知识发现已成为当前非常活跃的研究领域，粗糙集理论作为数据推理的强大工具已被成功地应用于知识获取、决策分析、预测、专家系统及数据库中的知识发现等许多领域。粗糙集理论与灰色系统、人工神经网络、概率统计、模糊集等理论有较强的互补性，将粗糙集理论与其他软技术两个或多个优势互补地杂合，构建功能更加强大的杂合软决策方法，可以拓宽粗糙集理论的适用范围，为不确定性决策问题提供多种科学、规范的决策方法。

1.1 软计算技术产生的时代背景和意义

随着 Internet 和数据库技术的迅猛发展和广泛应用，数据库中存储的数据量以惊人的速度增加，庞大的数据量渗透到社会生活和生产的各个领域，其结果导致传统的统计技术及数据管理工具不再适用于分析这些巨量的数据集。海量的数据被描述为“丰富的数据，贫乏的知识”。人们需要采用自动化程度更高、效率更高的数据处理方法来处理大量数据，并提供有用的知识。从金融业到制造业，越来越多的公司正依赖于巨量数据的分析获得竞争优势，知识已成为社会生活和生产的第一推动力。为了帮助人们智能化地分析海量数据，自动地分析一些事例，出现了新一代的技术和工具，这些技术和工具主要用于数据挖掘(data mining，DM)和知识发现(knowledge discovery in database，KDD)领域。KDD 指从大型数据库中自动提取知识，目标是发现数据中隐藏的、以前未知的、潜在有用的知识，本质上是在大的数据集合中寻找数据间的规则及普遍模式。数据挖掘可以视为用来发现这些规则和模式的方法。KDD 的流程如图 1.1 所示。

1. *数据挖掘的分析方法*

从功能上可以将数据挖掘的分析方法划分为自动预测趋势和行为、关联分析、聚类分析、概念描述和偏差检测五种。

(1) 自动预测趋势和行为

数据挖掘自动在大型数据库中寻找预测性信息，以往需要进行大量手工分析的问题如今可以迅速直接由数据本身得出结论。一个典型的例子是市场预测问题，数据挖掘使用过去有关促销的数据来寻找未来投资中回报最大的用户，其他可预测的问题包括预测破产以及认定对指定事件最可能作出反应的群体等。

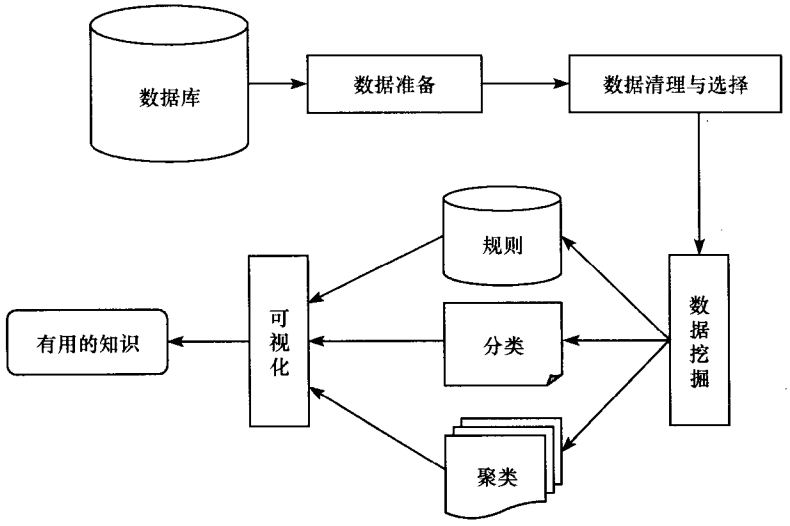

图 1.1 KDD 的流程

(2) 关联分析

数据关联是数据库中存在的一类重要的可被发现的知识。若两个或多个变量的取值之间存在某种规律性,就称为关联。关联可分为简单关联、时序关联、因果关联。关联分析的目的是找出数据库中隐藏的关联网。有时并不知道数据库中数据的关联函数,即使知道也是不确定的,因此关联分析生成的规则带有可信度也称为支持度,以表示这种规则发生的概率。例如在一个商场中,某天共有 1000 笔业务,其中有 100 笔业务同时买了微波炉和微波炉专用器皿,则购买微波炉的同时购买微波炉专用器皿的关联规则的支持度为 10%。

关联规则挖掘的一个典型例子是购物篮分析。市场分析员要从大量的数据中发现顾客放入其购物篮中的不同商品之间的关系。如果顾客买牛奶,他也购买面包的可能性有多大? 什么商品组或集合顾客多半会在一次购物时同时购买? 例如,买牛奶的顾客有 80%也同时买面包,或买铁锤的顾客中有 70%的人同时也买铁钉,这就是从购物篮数据中提取的关联规则。分析结果可以帮助管理人员设计不同的商店布局。一种策略是:经常一块购买的商品可以放近一些,以便进一步刺激这些商品一起销售,例如,如果顾客购买计算机又倾向于同时购买财务软件,那么将硬件摆放离软件陈列近一点,可能有助于增加两者的销售。另一种策略是:将硬件和软件放在商店的两端,可能诱发购买这些商品的顾客一路挑选其他商品。

(3) 聚类分析

数据库中的记录可被划分为一系列有意义的子集,即聚类。聚类增强了人们对客观现实的认识,是概念描述和偏差分析的先决条件。聚类技术主要包括传统

的模式识别方法和数学分类学。20世纪80年代初，Mchalski提出了概念聚类技术，其要点是：在划分对象时不仅考虑对象之间的距离，还要求划分出的类具有某种内涵描述，从而避免了传统技术的某些片面性。

(4) 概念描述

概念描述就是对某类对象的内涵进行描述，并概括这类对象的有关特征。概念描述分为特征性描述和区别性描述，前者描述某类对象的共同特征，后者描述不同类对象之间的区别。生成一个类的特征性描述只涉及该类对象中所有对象的共性。

(5) 偏差检测

数据库中的数据常有一些异常记录，从数据库中检测这些偏差很有意义。偏差包括很多潜在的知识，如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。偏差检测的基本方法是，寻找观测结果与参照值之间有意义的差别。

2. 数据挖掘所发现的知识

数据挖掘所发现的知识最常见的有以下五类：

① 广义知识(generalization)。广义知识指类别特征的概括性描述知识。根据数据的微观特性发现其表征的、带有普遍性的、具有较高层次概念的知识，反映同类事物共同性质，是对数据的概括、精炼和抽象。

② 关联知识(association)。关联知识反映一个事件和其他事件之间依赖或关联的知识。如果两项或多项属性之间存在关联，那么其中一项的属性值就可以依据其他属性值进行预测。

③ 分类知识(classification)。分类是反映同类事物共同性质的特征型知识和不同事物之间的差异型特征知识。

④ 预测型知识(prediction)。预测型知识指根据时间序列型数据，由历史的和当前的数据去推测未来的数据，也可以认为是以时间为关键属性的关联知识。

⑤ 偏差型知识(deviation)。偏差型知识是对差异和极端特例的描述，揭示事物偏离常规的异常现象，如标准类外的特例，数据聚类外的离群值等。所有这些知识都可以在不同的概念层次上被发现，并随着概念层次的提升，从微观到宏观，以满足不同用户不同层次决策的需要。

KDD属于典型的交叉学科领域，为了解决现实世界的问题，在KDD中需要计算机科学、统计、认知科学等相关领域的知识。KDD一词首先出现在1989年8月在美国底特律举行的第十一届国际联合人工智能学术会议上，随后在1991年、1993年和1994年都举行KDD专题讨论会，汇集了来自各个领域的研究人员和应用开发者，集中讨论数据统计、海量数据分析算法、知识表示、知识运用等问题。

KDD 组委会于 1995 年把专题讨论会更名为国际会议，并在加拿大蒙特利尔召开第一次 KDD 国际会议。1998 年在美国纽约举行的第四届知识发现与数据挖掘国际学术会议不仅进行了学术讨论，并且有 30 多家软件公司展示了它们的数据挖掘软件产品，不少软件已在北美、欧洲等国得到应用。KDD 和数据挖掘方法已成为当前人工智能和数据库技术的一个活跃的研究领域，依托数据库所进行的不确定性多属性决策已成为现代决策科学的一个重要组成部分，它广泛应用于社会、经济、管理及工程设计等众多领域。一些研究结果已被直接转换成商业计划，极大地改善了公司的决策质量。从金融业到制造业，越来越多的公司正依赖于巨量的数据分析获得竞争优势，知识已成为社会生活和生产的第一推动力。例如，应用数据挖掘(data mining)技术帮助解决商场货物摆放决策问题，著名的“啤酒与尿布”案例便是一个成功的例子。不同的数据挖掘方法有不同的目标，总的来说，数据挖掘方法可以分为两类：验证和发现。验证限于证明用户的假设，发现用于搜寻新的模式。发现又可分为预测和描述，通过系统发现的预测模式有助于指导未来的行为，描述是为了以一种可理解的形式将模式提供给用户。数据挖掘通常需要处理不完备、不确定的大容量数据，它与应用科学的其他学科有着相似性。面对复杂的不确定性决策问题，试图完全用数学模型进行精确刻画几乎是不现实的，即使对某些问题可行，但求解与分析也是非常困难的。为了帮助人们智能化地分析数据，自动地分析一些事例，出现了新一代的软计算工具，如粗糙集理论、灰色系统理论、模糊集理论、遗传算法和神经网络等，这些软计算技术及其优势互补地杂合旨在开采人们决策过程中的不精确性、不确定性、近似推理及部分正确性，以便获得易处理的、功能强大的、低成本的、且与人类决策极其类似的解决方法。

数据挖掘与知识发现中的许多问题可以凝练为管理活动中的不确定性决策问题。管理活动是由一系列决策组成的，在市场竞争非常激烈的今天，无论企业或个人都经常会遇到复杂的不确定性决策问题，需要分析与处理决策问题中多种不确定性情况，如随机性、模糊性、偏好性、粗糙性和灰性等，且快速做出决策。不确定性决策问题普遍存在于管理科学、信息科学、系统科学、计算机科学、知识工程及可靠性技术等众多领域中。由于现实世界中的许多不确定性问题是太复杂了，以至于不能用任何解析的且精确的模型来描述，精确的度量和控制的硬技术方法在处理这样复杂的问题时不是总是有效的，需要借助于人类的直觉，人类的直觉包含基于人们思维的直觉和已实现的主观思想，例如在噪声背景下专家对人脸的识别。粗糙集、灰色系统、概率论、模糊集和神经网络等软计算技术与人类思维的推理与学习的非凡能力相匹配，充分利用了人类的直觉知识，是解决不确定性决策问题的有效方法。与处理精确的、确定的且严格的硬技术方法相比较，软计算技术在获取不精确的或子最优的，但却经济的解决方法方面是有效的，且可与硬计算技术匹敌。由于软计算技术的独特功能，软计算技术已吸引了各种学术团队越来越浓厚

的研究兴趣。

1.2 粗糙集理论的特点与研究现状

在许多KDD问题中，学习分类是受到广泛研究的一类问题。粗糙集理论(rough set theory，RST)是处理不精确、不确定与不完全数据的理论，能有效地分析不精确、不一致、不完整等各种不完备的信息，是关于数据推理的一种强大工具。RST最初是由波兰科学家Pawlak于1982年提出的。由于最初关于粗糙集理论的研究大部分是用波兰语发表的，因此当时没有引起国际计算机学界和数学界的重视，研究地域也仅局限在东欧一些国家，直到20世纪80年代末才逐渐引起各国学者的注意。1992年，第一届关于粗糙集理论国际学术会议在波兰召开。1995年，ACM Communication将其列为新浮现的计算机科学的研究课题。1998年，国际信息科学杂志(Information Sciences)还为粗糙集理论的研究出了一期专辑。关于粗糙集理论的研究及应用现已经遍及于数学、计算机及各个专业领域。

1. 粗糙集理论的特点

粗糙集理论的成功主要是由于具有以下优点：仅依赖于原始数据，而不需要任何外部信息；RS方法不仅适用于分析质量属性而且适用于分析数量属性；约简冗余的属性，且约简算法较为简单，由RS模型导出的决策规则集给出了最小的知识表示；不修正不一致性，将生成的不一致规则划分为确定性规则和可能性规则；由RS方法导出的结果易于理解等。

由现实世界采集到的数据可能包含各种噪声，存在许多不确定因素和不完备信息有待处理。传统的不确定信息处理方法，如模糊集理论、证据理论和概率统计理论等因需要数据的附加信息或先验知识，在处理大数据量的数据库方面显得力不从心。作为一种软计算方法，RST与其他处理不确定和不精确问题理论的最显著的区别是它无需提供问题所需处理的数据集合之外的任何先验信息，如统计学中的概率分布、模糊集理论中的隶属度等，所以对问题的不确定性的描述或处理可以说是比较客观的。

RST是由实践需求驱动的用于数据分析与数据挖掘的一种新的数学方法，RST和KDD关系密切，它为KDD提供了一种新的方法和工具，理由如下：

① KDD研究的实施对象多为关系数据库，关系表可被看作为RST中的决策表，这给RS方法的应用带来极大的方便。

② 现实世界中规则有确定性的，也有不确定性的。从数据库中发现不确定性的知识，为RS方法提供了用武之地。

③ 从数据中发现异常，排除知识发现过程中的噪声干扰也是RS方法的特长。

④ 运用 RS 方法得到的知识发现算法有利于并行执行，可以极大地提高发现效率。对于大规模数据库中的知识发现来说，这是非常重要的。

⑤ 利用 RS 方法进行预处理，去掉多余属性，可提高发现效率，降低错误率。

⑥ 与模糊集方法或神经网络方法相比，由 RS 方法得到的决策规则及推理过程更易于被证实和解释。

2. 粗糙集理论的研究现状

RST 在人工智能中的研究主要可以分为两大类：有决策的分析与无决策的分析。

(1) 有决策的分析

有决策的分析主要应用于监督学习与决策分析。粗集理论在监督学习中的应用可以分为两个方面：对学习的训练集做预处理，应用 RS 方法获取规则。

对学习的训练集做预处理是考虑到实际测量中所获得的训练集，通常包含多余的属性，应用粗糙集的属性约简可有效地去除冗余的属性。例如，对豌豆疾病的数据进行粗集处理，使得原有属性数从 35 个约简到 9 个，而对美国 1984 年众议院的投票数据的分析，则使属性从原有的 16 个减少到 9 个；每个属性的值域也会有冗余，应用粗集方法中的约简技术可以删除某些属性的多余值；此外，应用粗集方法对神经网络的数据做预处理后，条件属性的个数及属性的取值都得到了化简。

应用 RS 方法获取规则是典型的监督学习，即利用粗集中提供的值约简方法由实例集直接获取规则。由于从决策表中直接获取所有的值约简已经被证明为一个 NP 完全问题，因此利用邻域独立的启发式算法求取最小值约简是一种常用的方法。在决策分析的应用中，则是利用粗糙集理论的属性约简、值约简及核等概念，对被决策的数据进行约简和寻求对于决策最有用的信息。

(2) 无决策的分析

无决策的分析主要包括：数据压缩、约简、聚类、模式发现与机器发现等。无决策的数据分析主要是利用属性约简去除不必要的属性，利用值约简压缩数据，进行数据的聚类分析。由于无决策的数据约简问题也是一个 NP 完全问题，因此仍然需要利用启发式知识求取最小约简。属于这类应用的典型人工智能分支是机器学习，特别是从大型数据库进行知识发现，粗糙集被认为是一种非常有效的方法。

目前，已开发了不少基于粗糙集的知识发现系统，其中具有代表性的有美国 Kansas 大学开发的 LERS(learning from examples based on rough sets)，在该系统中有两种不同的方法用于规则获取：一种是使用机器学习方法计算足够多的规则集；另一种是由知识获取方法计算所有的规则集。波兰 Poznan 工业大学计算科学研究所智能决策支持系统实验室研制的 ROSE 系统，该系统除了提供 RST 所有基础的运算外，还提供了避免数据离散的几种近似技术，如相似关系和优势关