粗糙集理论及其电力行业应用

孙秋野 黎明 著

机械工业出版社
CHINA MACHINE PRESS

粗糙集理论及其电力行业应用

孙秋野　黎　明　著

机 械 工 业 出 版 社

本书在介绍粗糙集的基本原理的基础上，立足于电力工程实际应用，重点对当前比较流行的经典粗糙集算法进行较为详尽的解读。同时根据工程应用的实际特点及程序实现中需要注意的问题提供取自于实际工程的典型案例进行详尽的解释，力求使读者通过阅读本书能够获得一条由粗糙集原理到实际电力工业应用的捷径，而这正是当前的各类书籍所没有涉足的领域。

本书第1、2章，系统介绍了粗糙集的基本理论及其在电力系统中的应用情况；第3章重点讨论了粗糙集在输配电系统故障诊断中的应用情况；第4章重点讨论了粗糙集在变压器油中溶解气体在线监测中的应用情况；第5章重点讨论了粗糙集在电力市场辅助分析决策中的应用情况。

本书力求清晰准确，以粗糙集成功工程项目为实例，旨在提供给读者一个具体形象的该方法的应用模型，等于架设起了一座沟通粗糙集理论与工程应用的桥梁。本书可以作为高等院校的高年级本科生和研究生教材或毕业设计及课题研究的辅助读物，或者作为工程技术人员的参考书。

图书在版编目(CIP)数据

粗糙集理论及其电力行业应用/孙秋野，黎明著. —北京：机械工业出版社，2009.1

ISBN 978-7-111-25308-2

Ⅰ.粗… Ⅱ.①孙…②黎… Ⅲ.粗糙集—应用—电力行业 Ⅳ.TM11

中国版本图书馆CIP数据核字(2008)第157857号

机械工业出版社(北京市百万庄大街22号 邮政编码100037)

策划编辑：张俊红 责任编辑：刘星宁 责任校对：张晓蓉

封面设计：马精明 责任印制：洪汉军

北京铭成印刷有限公司印刷

2009年1月第1版第1次印刷

184mm×260mm ·16.75印张·415千字

0001—3500册

标准书号：ISBN 978-7-111-25308-2

定价：35.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

销售服务热线电话：(010)68326294

购书热线电话：(010)88379639 88379641 88379643

编辑热线电话：(010)88379764

封面无防伪标均为盗版

序

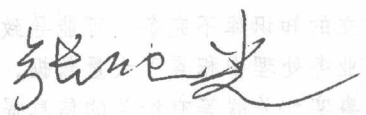
电力是人类社会活动中应用范围最广、对人们生活质量影响最大的能源。随着我国国民经济的持续发展，高度的社会信息化和生产自动化都在不断加深人们对电力的依赖性，这也促使电力行业加大对电力系统可靠性的管理力度，力求为用户提供更加充裕、持续、安全可靠的电力供应。

当今的电力工业由于自动化、计算机等方面技术的高速发展，已经脱离了原有的就地控制、手动控制模式，跨进了一个网络信息时代。原有的信息处理和控制手段已经无法有效处理当前电力工业出现的新问题。由此，大量的人工智能技术被广大研究者引入了电力系统的许多领域，并取得了引人注目的成果。这其中，由于粗糙集善于从海量强干扰数据中挖掘潜在的有价值信息，而且无需提供问题所需处理数据集合以外的任何先验信息，从而得到了众多电力科研工作者的青睐。

粗糙集理论是一种建立在熵集上的理论。一方面，它为归纳机器学习建立了理论基础；另一方面，从该理论中提出的独立约简与正区域的概念出发，可以演变为对实际应用有重要意义的一系列理论。粗糙集理论作为数据挖掘中的一个重要方法，在很多领域都起到了重要的作用。但遗憾的是，由于粗糙集理论需要应用者具有较艰深的数学基础，从而严重影响了其在电力系统中的广泛应用。

本书正是针对这一状况，将粗糙集原理与实际的电力系统工程应用相结合，力求理论研究来源于工程难题，然后以工程实际说明理论价值。近10年来，本书作者孙秋野博士、黎明博士和课题组的合作者们在粗糙集基础理论以及电力系统应用方面进行了一系列深入的研究，并取得了良好的研究成果。这些成果分别发表在中国电机工程学报、自动化学报、仪器仪表学报、自然科学进展等权威刊物上，并申请了多项国家发明专利。本书涵盖了作者近10年的研究成果，将粗糙集理论与电力系统工程应用进行了较好的结合，并在输配电网故障诊断、电力变压器油中溶解气体监测、电力市场决策支持等电力系统的前沿热点研究领域中提供了工程实例，是将粗糙集与电力系统工程应用有机结合的一本好书。

Automatica 副主编
IEEE Trans. SMC Part B 副主编
IEEE Trans. FS 副主编
教育部长江学者特聘教授



前 言

近年来，随着电力事业的不断发展，粗糙集技术在电力系统的各个领域中也得到广泛的应用，这主要是由电力系统和粗糙集的特点决定的。

在电力系统中，数据主要分为由分布于系统各处的各种装置实时采集的现场数据和由调度中心诸如监控与数据采集(SCADA)系统收集的中央数据以及管理信息系统、地理信息系统等在使用过程中产生的大量数据，数据的来源较多。另外，整个电力系统是用大规模的非线性微分方程和非线性代数方程来描述的，两者联立就组成了大规模奇异非线性动态大系统。在对其进行特征描述时往往涉及上千个状态变量。传统的处理方法是对系统进行降维或者简约化处理，这在一定程度上影响了最终结果的精度。

电力系统是一个标准的混杂系统，其上层(调度中心)给出的调度决策主要是逻辑性的操作指令，而下层控制(如发电机的励磁与调速控制)主要是连续性的。为了达到电力系统的多目标优化控制的目的，应将不同性质的上层和下层控制有机地结合起来。而且，其采集到的数据中包含着诸如噪声、数据缺失等不确定因素。当系统处于紧急状态，甚至瓦解状态时，必须制定实时在线快速决策，使系统重新回到正常状态。

电力系统的这些特征使得它迫切需要一种能快速处理海量数据的技术支持。

粗糙集技术可以在以下几个方面与电力系统结合：

1. 电力网络的故障诊断

当前的电力系统，特别是输配电系统，能够得到的可用信息越来越多，在应用于配电网故障诊断的众多信号中，也包含很多冗余信息。对于这些冗余信息，当前主要采用的故障诊断方法多数无法将其过滤掉，也就是说需要全部的信息才可以进行故障诊断。而对于目前的配电网信息情况传输来说，在故障条件下，需要传输的信息非常多，很多时候会产生数据风暴现象，这时候想得到完整的故障信息无疑是很困难的。因此，能够得到匹配度更高的规则系统无疑是很有意义的一项工作。而粗糙集约简技术正是一个完全由数据驱动的海量数据信息提取手段，据此可以获得最优的故障判断规则集。

2. 电力设备的状态监测

电力变压器作为电力系统中最重要的电气设备之一，维护其正常运行是整个电力系统可靠供电的基本保证。由于电力变压器多在室外露天工作，运行条件恶劣，故出现故障的几率较高。据统计，我国大型电力变压器的故障率高达8.5%/年。因而，研究电力变压器的故障诊断方法具有重要的现实意义。但是，获取完备的知识库是制约专家系统发展的瓶颈，如果建立的知识库不完备，可能导致专家系统推理混乱并得出错误的结论。电力设备的监测系统在业务处理中积累了大量数据，可以利用粗糙集将这些数据中蕴藏着的许多潜在的重要因素、事实和关联等有价值的信息提炼出来。

3. 电力市场用户特征分析

在电力市场中，用户和其他成员是平等的伙伴关系。对于开放用户市场的电力市场，用户可以选择供应方和贸易方式。在电力市场营销中，供电方作为商家为节省营销成本获得更

多的利润，应该通过收集、加工和处理用户的大量信息确定特定用户群体和消费需求，进而推断出用户群体下一步的消费行为。然后以此为基础，对识别出来的用户群体进行特定内容的定向营销。同样在电力市场中，供电方也必须在对用户负荷的特性充分了解的基础上，对用户的行为分门别类，从而在保证系统安全稳定运行的前提下，制定出有竞争力的供电策略。考虑到电力系统自身的特点，供电方还应制定有效的负荷管理策略，调整负荷曲线的形状，降低对峰荷的要求，节约能源。上述工作都可以采用粗糙集约简技术进行。

另外，在电力系统安全稳定性评估、电力系统规划设计、电力系统经济规划和电力系统调度运行等方面粗糙集技术都有广泛的应用前景。

当前粗糙集技术发展得如火如荼，但对于实际工程人员和大中专院校的本科生和研究生来说如何在较短的时间内将其应用于实际工程还是一个难题。当前，普遍的情况是介绍粗糙集理论的书仅仅是介绍理论本身，主要的内容都是深入的理论基础、相关的证明推导和最新的研究成果，而并不涉及算法的计算机语言实现问题。但是，作为大多数关心粗糙集发展的人来说，他们并不是粗糙集领域的专家，也不需要对粗糙集理论做多么重大的创造性工作，他们关心的并不是枯燥的理论推导，而是如何将这些先进的理论应用到自己的实际研究工作中去，更好地为应用服务。因此，对于这些人来说，粗糙集如何应用计算机语言实现远比算法是从何而来更有实际价值。

遗憾的是，由于粗糙集理论发展时间不长，理论及应用并不完善，因此关于算法实现及工程应用的书基本没有。本书正是针对当前这一状况，将粗糙集原理与实际的电力系统工程应用相结合，力求理论研究来源于工程难题，然后以工程实际说明理论价值。

本书立足于在较短的篇幅内提炼粗糙集算法的精髓，使读者能够将更多的注意力用于应用本身，力争使读者能够在最短的时间内完成从对于粗糙集不了解到能够完成初步的程序设计，并进而能够进行工程应用的三级跳。并提供一条切实可行的将粗糙集理论转化为程序的实现方案。本书以其成功工程项目为实例，旨在提供给读者一个具体形象的该方法的应用模型，等于架设起了一座沟通粗糙集理论与工程应用的桥梁。

本书可以作为高等院校的本科生和研究生教材或毕业设计及课题研究的辅助读物，或者作为工程技术人员的参考书。

本书由孙秋野、黎明著，全书由孙秋野统稿。在本书的编写过程中，得到了王志强博士、张铁岩教授、冯健副教授、季策副教授、李爱平副教授、闫世杰副教授、王占山副教授、王迎春博士的大力支持，同时刘国威、葛辉、杨伟志、杨东升、杨君、王智良、刘金海、刘秀冲、刘鑫蕊、罗艳红、梁绵鑫、赵琰、马铁东、魏庆来、张锐等也参加了本书部分内容的编写和素材提供工作。另外，作者在编写本书的过程中参考了不少专家和学者的著作、学术论文和经验总结等，在此对他们表示最诚挚的谢意！

限于作者的理论水平和实际开发经验，书中难免存在一些不足之处或者错误，恳望读者和相关专家批评指正。

作　者

目　　录

第 1 章　粗糙集理论的基本概念与应用

粗糙集(Rough Set,RS)理论是一种建立在熵集上的理论。一方面，它为归纳机器学习建立了理论基础；另一方面，从该理论中提出的独立约简与正区域的概念出发，可以演变为对实际应用有重要意义的一系列理论。如果考虑目前数据挖掘中最为典型的研究动机——关联规则所使用的方法，可以同样建立在熵集上，由粗糙集理论研究派生的研究成果与研究方法同样可以适用于关联规则的研究。事实上，关联规则在算法上所面临的问题与基于粗糙集理论设计的算法所面临的问题没有差别，都是海量数据的问题。尽管关联规则有其特殊的研究课题，但是在约简问题上与粗糙集理论没有本质的不同。而且，粗糙集理论似乎比关联规则方法有更大的研究空间。

粗糙集理论作为数据挖掘中的一个重要方法，在很多领域都起到了重要的作用。它的最主要的贡献体现在以下三个方面：

1）可以作为归纳机器学习的理论基础。

2）独立约简，这是一个有数学基础的结构目标，从而部分解决了数据挖掘平凡性问题。

3）在粗糙集理论中，正区域起着重要的作用，一个给定信息系统正区域的本质是在样本集合中无矛盾样本的数量。

1.1　粗糙集的产生与发展

粗糙集理论是一种刻画不完整性和不确定性的数学工具，能有效地分析和处理不精确、不一致、不完整等不完备信息，并从中发现隐含的知识，揭示潜在的规律。粗糙集理论是由波兰学者 Pawlak Z. 在 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学术界的重视，研究集中在东欧一些国家，直到 20 世纪 80 年代末期才逐渐引起各国学者的注意。1991 年 Pawlak Z. 出版了专著《Rough Set-Theoretical Aspects of Reasoning about Data》。该著作系统全面地阐述了粗糙集理论，奠定了严密的数学基础。1992 年第一届关于 RS 理论的国际会议在波兰召开，以后每年一次。1995 年，ACM Communication 将其列为新浮现的计算机科学的研究课题。1998 年，国际杂志《Information Sciences》出版了一期粗糙集理论研究的专辑。近几年以来，由于它在机器学习与知识发现、数据挖掘以及决策支持与分析等方面的应用，逐渐成为人工智能领域中一个新兴的学术热点。我国对粗糙集的研究虽然起步较晚，但发展迅速，并出版了一些专著，引起了越来越多的科研人员的关注。

粗糙集理论是建立在分类机制的基础上的，它将分类理解为在特定空间上的等价关系，而等价关系构成了对该空间的划分。粗糙集理论将知识理解为对数据的划分，每一个被划分的集合称为概念。粗糙集理论的主要思想是利用已知的知识库，将不精确或不确定的知识用已知知识库中的知识来(近似)描述。

该理论主要有以下几个特点：

1）粗糙集不需要先验知识。“Data speak themselves”是粗糙集研究的理念，它是完全由数据驱动的，不需要像概率统计方法要预先假定概率分布，也不需像模糊集理论要假设模糊隶属函数的结构。粗糙集分析方法仅利用数据本身提供的信息，无需任何先验知识。

2）粗糙集是强大的数据分析工具。它能表达和处理不完备的信息；能在保留关键信息的前提下，对数据进行化简并求得知识的最小表达；能够评估数据之间的依赖关系，揭示出概念的简单模式；能从经验数据中获取易于证实的规则知识。因此，基于粗糙集的数据分析(Rough Set Data Analysis,RSDA)方法成为粗糙集理论中的一个非常活跃的分支。

目前，对粗糙集理论的研究主要集中在：粗糙集模型的扩展、粗糙集数学性质的研究、问题的不确定性研究、粗糙逻辑和粗糙推理，以及与其他处理不确定性、模糊性问题的数学理论的关系等方面。

1. 粗糙集模型的扩展

Pawlak 粗糙集模型的推广一直是粗糙集理论研究的主流方向。目前主要有两种方法，分别是构造性方法和代数(公理化)方法。

Pawlak 提出的经典的粗糙集模型在应用于数据分析时，会遇到噪声、数据缺失、大数据量、连续属性离散化等具体问题，使得实际效果不是很理想。为此，人们提出了许多粗糙集模型的扩展模型，其中最典型的有可变精度模型和相似模型。

（1）可变精度模型

在数据集中存在噪声等干扰情况下，经典理论会由于对数据的过拟合而使其对新对象的预测能力大为降低。而在实际应用中，噪声是在所难免的。为增强粗糙集模型的抗干扰能力，Ziarko 提出了一种可变精度粗糙集模型。该模型通过引入分类精度，使模型具有一定的容错性。

（2）相似模型

经典粗糙集模型的基础是不可分辨关系，但是这个条件是很强的。在数据中存在缺失的属性值时(在数据库中很普遍)，按照不可分辨关系进行处理，往往约简效果很差。为扩展粗糙集的能力，Kryszkiewicz 提出了用相似关系来代替不可分辨关系作为粗糙集约简的依据。

2. 粗糙集数学性质的研究

粗糙集数学性质的研究，主要讨论粗糙集的代数结构、拓扑结构和收敛性等问题。其中，粗糙集的代数结构和拓扑结构，粗糙函数的一些性质，与模糊隶属函数相对应的粗糙隶属函数，以及实数粗糙离散化和实函数粗糙离散化等方面的问题都得到了众多专家的重视。

3. 粗糙逻辑与粗糙推理

粗糙逻辑是指定义在属性值为邻域的决策表上的一种逻辑。它在数据约简中有着广泛的应用前景。Pawlak 建立了粗糙逻辑的 5 个逻辑真值。Orlowska 提出了以等价关系 R 为基础的新谓词，扩充了经典的二值逻辑。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，并建立了基于两个算子的近似推理演绎系统。Liu 还提出了带 L 和 H 的粗糙逻辑近似推理模式和归结原理，并证明了它的归结完备性定理。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。

4. 粗糙集约简算法

约简是粗糙集用于数据分析的重要概念。然而最小约简的计算是 NP-hard 的。因此运用

启发信息来简化计算是最直接的思想。Hu 提出了一种基于属性重要性的约简算法，该算法以“核”为计算约简的出发点，将属性的重要性作为启发规则。Jakub 提出的遗传算法去寻找系统的最小约简。Kryszkiewicz 和 Rybinski 研究了在复合信息系统中寻求约简的问题。通过寻求子系统的约简最终求出复合系统的约简。其主要思想是将布尔函数的化简问题转化成集合空间中的边界搜索问题，从而在已知子系统约简的情况下，简化复合系统的搜索空间。Starzyk 等提出强等价的概念，进而发展为扩展法则，用于快速简化区分函数。Bazan 等提出动态约简方法，该方法能够有效地提高约简的抗噪声能力。

5. 粗糙集理论与其他方法的融合

目前粗糙集理论与其他处理模糊性或不确定性方法的理论研究，主要集中在概率统计、模糊数学、D-S 证据理论和信息论的相互渗透与补充。Duntsch 定义了粗糙集数据分析的统计特性，并依据该统计特性给出了一种数据过滤方法，并结合信息熵原理定义了两种衡量粗糙集数据分析质量的不确定性测度。而模糊集和粗糙集之间的关系，粗糙集证据理论之间的关系，粗糙集理论与证据理论中的信任函数之间的内在联系等都被专家进行了深入的讨论。这些研究成果都说明，粗糙集理论和它们都有交叉的部分，不能够互相取代，反而需要相互补充融合，揭示它们之间内在的联系和本质的区别是非常有意义的研究课题。

粗糙集理论在应用方面取得了令人瞩目的成就，实际上正是应用的不断发展才引起了人们的广泛注意，并极大地促进了理论的研究。粗糙集理论在数据减缩与规则生成、大数据集、信息检索、决策支持方面都取得了一定的成绩。基于粗糙集理论的数据分析软件的开发也取得了很大的进步，比较著名的有以下几种：

1）美国 Kansas 大学开发的基于粗糙集理论的学习(Learning from examples based on RS, LERS)系统。

2）波兰 Poznan 科技大学开发的 Rough DAS&Rough Class。

3）挪威科技大学和波兰华沙大学数学研究所开发的 Rosetta 系统。

4）英国 Ulster 信息与软件工程研究所开发的 Grobian。

这些软件的开发使得粗糙集理论作为一种工具被更多的人更广泛地使用。

1.2　知识的表示方法

1.2.1　知识的基本表示方法

知识是人工智能中一个非常重要的概念，解决复杂问题需要大量的知识以及处理这些知识的机构。知识在不同的范畴中有不同的含义。在粗糙集理论中，知识被看作是关于论域的划分，是一种对对象进行分类的能力。例如，医生给病人看病，可以依据病人的征兆判断出病情属于哪一类。这种根据事物的特征将其分门别类的能力，就是“知识”。

设 $U\neq\varnothing$ 是人们感兴趣的对象组成的有限集合，称为论域。任何子集 $X\subseteq U$，称为 U 中的一个概念或范畴。为规范化起见，认为空集也是一个概念。U 中的任何概念族称为关于 U 的抽象知识，简称知识。知识被认为是一种对抽象或现实的对象进行分类的能力。根据所讨论对象的特征差别，其分门别类的能力均可以看作是某种知识。一个划分 ε 的定义为 $\varepsilon=$

$\{X_1, X_2, \cdots, X_n\}$，$X_i \subseteq U$，$X_i \neq \varnothing$，对于 $i \neq j$，$j = 1, 2, \cdots, n$，$\bigcup_{i=1}^{n} X_i = U$。$U$ 上的一族划分称为关于 U 的一个知识库(Knowledge Base)。不可分辨关系是粗糙集理论的基石，揭示了论域中知识的颗粒结构，也是定义其他概念的基础。

定义 1.1 给定一个有限的非空集合论域 U，定义 R 代表论域 U 中的一种等价关系，称为 U 上的分类族，即知识库 $K = (U, R)$ 为一个近似空间。

定义 1.2 对于子集 X、$Y \subset U$，若根据关系 R、X 和 Y 的属性不可分辨时，用 $[X]_R$ 来表示，它代表子集 X 和 Y 同属 R 中的一个范畴。

定义 1.3 若 $P \subseteq R$，且 $P \neq \varnothing$，则 P 中全部等价关系的交集也是一种等价关系，称为 P 上的不可分辨关系，记为 $IND(P)$：

$$[X]_{IND(P)} = \cap [X]_R \qquad P \subseteq R$$

在不可分辨关系的基础上，引入粗糙集理论中关于知识的等价概念。

定义 1.4 令 P、$Q \subset R$、P、Q 都是关于论域上的知识，且 $K = (U, P)$ 和 $K_1 = (U, Q)$ 为两个知识库，当 $IND(P) = IND(Q)$ 时，称知识库 K、K_1 是等价的。

这个概念意味着可以用于不同的属性集对同一对象进行描述，以表达关于论域上完全相同的事实，即 K 与 K_1 具有相同的表达能力。

1.2.2 知识不确定性的表示与分类

粗糙集理论中知识的不确定性主要由两个原因产生：一个原因是直接来自于论域上的二元等价关系及其产生的知识模块，即近似空间本身。如果二元等价关系产生的每一个等价类中只有一个元素，那么等价关系产生的划分不含有任何信息。划分越粗，每一个知识模块越大，知识库中的知识越粗糙，相对于近似空间的概念和知识就越不确定。这时处理知识的不确定性的方法往往用香农信息熵来刻画，知识的粗糙性与信息熵的关系比较密切，知识的粗糙性实质上是其所含信息多少的更深层次的刻画。单从这个角度来看，粗糙集理论与信息论的关系是比较密切的。

粗糙集理论中知识不确定性的另一个原因来自于给定论域里粗糙近似的边界。当边界为空集时，知识是完全确定的，边界越大，知识就越粗糙或越模糊。

下面举一个具体的例子来解释上面的概念。

有 8 个物体，可以用颜色(C)、密度(P)、体积(L)和重量(W)四个属性来描述。于是，由表 1-1 可以构造信息系统 $I = <U, \Omega, V, f>$。

表 1-1 决策表

X	C	P	L	W	X	C	P	L	W	X	C	P	L	W
1	1	1	2	2	4	2	2	1	1	7	1	2	2	2
2	2	3	2	3	5	2	2	2	2	8	3	3	2	3
3	1	1	1	1	6	3	1	1	1					

其中，$U = \{x_1, x_2, \cdots, x_7, x_8\}$，$\Omega = \{C, P, L, W\}$，$C$、$P$、$L$ 为条件属性，W 为决策属性。$V_C = \{1(红), 2(黄), 3(绿)\}$，$V_P = \{1(低密度), 2(中密度), 3(高密度)\}$，$V_L = \{1(小), 2(大)\}$，$V_W = \{1(轻), 2(中), 3(重)\}$，$f$ 表示了用语言描述的属性值到数值描述的关系，如把“红色”用数值“1”表示。表 1-1 为该信息系统的决策表。由各个属性决定的等价类集合为

$$U/C=\{\{x_1,x_3,x_7\},\{x_2,x_4,x_5\},\{x_6,x_8\}\}$$
$$U/P=\{\{x_1,x_3,x_6\},\{x_4,x_5,x_7\},\{x_2,x_8\}\}$$
$$U/L=\{\{x_3,x_4,x_6\},\{x_1,x_2,x_5,x_7,x_8\}\}$$
$$U/W=\{\{x_3,x_4,x_6\},\{x_1,x_5,x_7\},\{x_2,x_8\}\}$$

下面利用粗糙集的分析方法来分析哪些条件属性对于得出物体“重量”这个结论是关键的。

$$R=\{C,P,L\},$$
$$X_1=[1]_W=\{x_3,x_4,x_6\},\ X_2=[2]_W=\{x_1,x_5,x_7\},\ X_3=[3]_W=\{x_2,x_8\};$$
$$U/R=\{\{x_1\},\{x_2\},\{x_3\},\{x_4\},\{x_5\},\{x_6\},\{x_7\},\{x_8\}\};$$
$$\underline{R}X_1=\{x_3,x_4,x_6\},\ \overline{R}X_1=\{x_3,x_4,x_6\},\ \alpha_R(X_1)=1;$$
$$\underline{R}X_2=\{x_1,x_5,x_7\},\ \overline{R}X_2=\{x_1,x_5,x_7\},\ \alpha_R(X_2)=1;$$
$$\underline{R}X_3=\{x_2,x_8\},\ \overline{R}X_3=\{x_2,x_8\},\ \alpha_R(X_3)=1;$$

所以，$\{C,P,L\}\rightarrow\{W\}$。

令 $R_C=R-\{C\}$，$U/R_C=\{\{x_1\},\{x_2,x_8\},\{x_3,x_6\},\{x_4\},\{x_5,x_7\}\}$。

$$\underline{R_C}X_1=\{x_3,x_4,x_6\},\ \overline{R_C}X_1=\{x_3,x_4,x_6\},\ \alpha_{R_C}(X_1)=1;$$
$$\underline{R_C}X_2=\{x_1,x_5,x_7\},\ \overline{R_C}X_2=\{x_1,x_5,x_7\},\ \alpha_{R_C}(X_2)=1;$$
$$\underline{R_C}X_3=\{x_2,x_8\},\ \overline{R_C}X_3=\{x_2,x_8\},\ \alpha_{R_C}(X_3)=1;$$

所以，$\{P,L\}\rightarrow\{W\}$。

令 $R_P=R_C-\{P\}$，$U/R_P=\{\{x_3,x_4,x_6\},\{x_1,x_2,x_5,x_7,x_8\}\}$。

$$\underline{R_P}X_1=\{x_3,x_4,x_6\},\ \overline{R_P}X_1=\{x_3,x_4,x_6\},\ \alpha_{R_P}(X_1)=1;$$
$$\underline{R_P}X_2=\varnothing,\ \overline{R_P}X_2=\{x_1,x_2,x_5,x_7,x_8\},\ \alpha_{R_P}(X_2)=0;$$
$$\underline{R_P}X_3=\varnothing,\ \overline{R_P}X_3=\{x_1,x_2,x_5,x_7,x_8\},\ \alpha_{R_P}(X_3)=0;$$

所以，$\{L\}\rightarrow\{W\}$不成立。

由以上分析可得$\{P,L\}$是该系统的一个约简，也就是说物体的密度和体积决定了物体的质量，这个结论是合理的。消除决策表中冗余的行后得到约简的决策表，见表 1-2。

不难验证$\{P,L\}$是该信息系统惟一的约简，所以$\{P,L\}$也是该信息系统的关键信息。

表 1-2　化简后的决策表

X	P	L	W	X	P	L	W	X	P	L	W
1	1	2	2	3，6	1	1	1	5，7	2	2	2
2，8	3	2	3	4	2	1	1				

1.3　粗糙集的基本概念

1.3.1　粗糙集的基本定义

粗糙集理论的主要思想就是在保持分类能力不变的前提下，通过知识约简，导出问题的决策或分类规则。目前，粗糙集理论已经被成功地应用于机器学习、决策分析、过程控制、

模式识别与数据挖掘等领域。

下面，为了讨论方便，对粗糙集理论中的一些重要概念作简单介绍。

1. 信息系统

一个信息系统 I 是一个四元组：$I=<U, \Omega, V, f>$。其中，U 为论域，是全体对象的有限集合，设有 n 个对象，则其可以表示成 $U=\{x_1,x_2,\cdots,x_n\}$；Ω 为有限个属性集合，设有 m 个属性，则其可以表示为 $\Omega=\{A_1,A_2,\cdots,A_m\}$；$V$ 为属性的值域集合，$V=\{V_1,V_2,\cdots,V_m\}$，V_i 是属性 A_i 的值域；f 为信息函数，$f: U\times\Omega\rightarrow V$，$f(x_i,A_j)\in V_j$。

2. 上近似、下近似、边界区

给定一个有限的非空集合 U，R 为 U 上的一族等价关系。R 将 U 划分为互不相交的基本等价类，记作 U/R，二元组 $K=(U,R)$ 构成一个近似空间。设 $X\subseteq U$，$Y\in U/R$，则

集合 X 关于 R 的下近似 $\underline{RX}$ 定义为

$$\underline{RX}=\cup\{Y\in U/R: Y\subseteq X\}$$

$\underline{RX}$ 是由那些根据知识 R 判断肯定属于 X 的对象所组成的最大集合，也称为 X 的正区，记作 $POS_R(X)$。

集合 X 关于 R 的上近似 $\overline{RX}$ 定义为

$$\overline{RX}=\cup\{Y\in U/R: Y\cap X\neq\varnothing\}$$

$\overline{RX}$ 是由所有与 X 相交非空的 R 的等价类的并集，是根据知识 R 判断所有可能属于 X 的对象组成的最小集合。

$NEG_R(X)=U-\overline{RX}$ 称为集合 X 的负区，它表示了根据知识 R 可以判断肯定不属于 X 的对象组成的集合。

集合 X 关于 R 的边界区 $BN_R(X)$ 定义为

$$BN_R(X)=\overline{RX}-\underline{RX}$$

边界区为集合 X 的上近似和下近似之差，它表示了根据知识 R 既不能判断肯定属于 X，也不能判断肯定不属于 X 的对象所组成的集合。如果 $BN_R(X)$ 是空集，则称 X 关于 R 是清晰的，反之，如果 $BN_R(X)$ 是非空的，则称集合 X 为关于 R 的粗糙集。

图 1-1 直观地描述了上近似、下近似和边界区的关系。设椭圆部分所围的区域为集合 X，整个矩形所围的区域为论域 U。

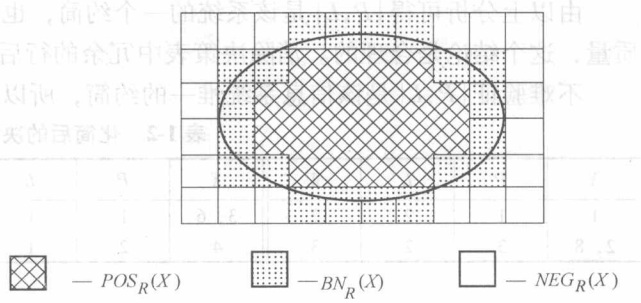


图 1-1 粗糙集示意图

3. 逼近精度

上近似、下近似和边界区等概念刻画了一个不能精确定义的集合的逼近特性。逼近精度作为集合不确定性的一种度量，定义为 $\alpha_R(X)=\dfrac{\mathrm{card}(\underline{RX})}{\mathrm{card}(\overline{RX})}$，其中，$\mathrm{card}(*)$ 表示集合 $*$ 的基数或势，对有限集合来说表示集合中所包含元素的个数。显然 $0\leqslant\alpha_R(X)\leqslant 1$，如果 $\alpha_R(X)=1$，则称集合 X 相对于 R 是清晰的；$\alpha_R(X)<1$，则称集合 X 相对于 R 是粗糙的。$\alpha_R(X)$ 可认为是在等价关系 R 下逼近集合 X 的精度。