Rough集理论与知识获取

王国胤 编著

西安交通大学出版社

Rough 集理论与知识获取

王国胤　编著

西安交通大学出版社

图书在版编目(CIP)数据

Rough 集理论与知识获取 / 王国胤编著. —西安:西安交通大学出版社,2001.5
ISBN 7-5605-1409-X

Ⅰ. R… Ⅱ. 王… Ⅲ. 计算机应用-信息处理-方法
Ⅳ. TP391

中国版本图书馆 CIP 数据核字(2001)第 022116 号

*

西安交通大学出版社出版发行
(西安市兴庆南路 25 号 邮政编码:710049 电话:(029)2668316)
长安县第二印刷厂印装
各地新华书店经销

*

开本:850mm×1 168mm 1/32 印张:7.5 字数:184 千字
2001 年 5 月第 1 版 2001 年 5 月第 1 次印刷
印数:0 001~1 000 定价:15.00 元(平) 20.00 元(精)

作者简介

王国胤，男，1970 年 3 月生，重庆市人，汉族，工学博士，重庆邮电学院特聘学者、计算机科学与技术学院副院长、计算机科学与技术研究所所长、教授。1992 年毕业于西安交通大学计算机软件专业，获工学学士学位（陕西省优秀大学毕业生、西安交通大学优秀大学毕业生）；1994 年毕业于西安交通大学计算机软件专业，获工学硕士学位；1996 年毕业于西安交通大学计算机组织与系统结构专业，获工学博士学位（西安交通大学优秀博士毕业生）；1998 年至 1999 年以访问学者身份工作于美国 University of North Texas 计算机科学系（1 年）；1999 年以访问学者身份工作于加拿大 University of Regina 计算机科学系（2 个月）。现为 IEEE 会员，国际科学技术开发协会（IASTED International Association of Science and Technology for Development）信息学技术委员会委员，重庆市科技进步奖专业评审委员会委员。1992 年以来，一直从事智能信息系统的理论及应用研究，在 Rough 集理论、并行神经网络体系结构、逻辑神经网络、神经网络自动知识获取、集成智能系统、数据压缩、网络信息系统等领域开展了研究。作为主研人员参与完成了 2 项国家自然科学基金项目和 1 项高技术产品研制开发项目，作为项目负责人主持完成了 5 项省、部级科研项目，现正主持国家自然科学基金、国家 863 计划、攀登计划、高等学校骨干教师资助计划、教育部回国留学人员科研启动基金和重庆市应用基础研究基金等共 6 项科研项目的研究工作，先后多次赴香港、美国、加拿大、日本等地出席国际学术会议和进行学术访问，在国内、外主要学术刊物和学术会议上发表了 40 余篇学术论文，其中多篇被国际权威检索刊物《工程索引》（SCI），《科学引文索引》（EI），《国际科技会议索引》（ISTP），《英国科学文摘》（INSPEC）等收录。

内容简介

Rough 集理论是一种研究不完整、不确定知识和数据的表达、学习、归纳的理论方法，近年来在理论模型、算法研究、工程应用中取得了好的成果和应用。本书重点在于阐述 Rough 集理论的模型、算法以及基于 Rough 集理论的知识获取技术。全书共分 11 章。第 1 章介绍了集合论基础；第 2 章介绍了信息表知识表达系统；第 3 章介绍了 Rough 集基础理论；第 4 章介绍了知识获取的基本问题；第 5 章介绍了知识系统不确定性的表示与处理问题；第 6 章介绍了数据预处理技术；第 7 章介绍了信息表属性约简的理论与算法；第 8 章介绍了信息表值约简的理论与算法；第 9 章介绍了逻辑推理方法；第 10 章介绍了几个典型的 Rough 集工程实例；第 11 章介绍了几个 Rough 集演示软件系统。

本书的目的就是要向计算机学科、人工智能学科、智能信息处理学科、机器学习学科、自动化学科等研究领域的研究人员系统介绍 Rough 集理论这一新的理论工具及其应用技术。本书可以作为计算机、自动化等专业高年级本科生、硕士生和博士生的学习参考用书，同时对相关学科领域的科技工作者和工程技术人员也有重要的使用和参考价值。

前　言

智能信息处理是当前信息科学理论和应用研究中的一个热点领域，随着过去几十年中人们在专家系统、知识工程、人工神经网络、模糊集合等众多领域的不断实践和探索，取得了很多很好的成绩。随着信息时代的到来，信息量不断增长，对信息分析工具的要求也越来越高，人们希望自动地从数据中获取其潜在的依赖模型。这样，大量的数据就无须人的处理，甚至无须人的观察。因此，研究能够从大量信息中形成实际概括（归纳）的系统就显得越来越重要。虽然已经有很多对数据进行分析的简单统计技术，但高级的智能数据分析技术还远没有成熟。因此，数据信息的产生和对它的理解之间的差距越来越大。

Rough 集（Rough Sets，有的也称粗集、粗糙集）理论是由波兰华沙理工大学 Pawlak 教授于 20 世纪 80 年代初提出的一种研究不完整、不确定知识和数据的表达、学习、归纳的理论方法，近年来得到国际上众多学者的重视。我国也在国家自然科学基金、国家 863 计划和一些省、市科学研究基金的支持下开展了一定的研究工作，逐渐取得了一些研究成果。

Rough 集的研究对象是由一个多值属性（特征、症状、特性等）集合描述的一个对象（观察、病历等）集合，对于每个对象及其属性都有一个值作为其描述符号，对象、属性和描述符是表达决策问题的 3 个基本要素。这种表达形式也可以看成为一个二维表格，表格的行与对象相对应，列对应于对象的属性；各行包含了表示相应对象信息的描述符，还有关于各个对象的类别成员的信息。通常，关于对象的可得到的信息不一定足以划分其成员类别。换句话说，这种不精确性导致了对象的不可分辨性。给定对象间的一个等价关系，即导致由等价类构成的近似空间的不分明关系，Rough

集就用不分明对象类形成的上近似和下近似来描述。这些近似分别对应了确定属于给定类的最大的对象集合和可能属于给定类的最小的对象集合。下近似和上近似的差是一个边界集合，它包含了所有不能确切判定是否属于给定类的对象。这种处理可以定义近似的精度和质量。Rough 集方法可以解决重要的分类问题，所有冗余对象和属性的约简包含属性的最小子集，能够很好地近似分类，得到可以接受质量的分类。而且，它还可以用决策规则集合的形式表示最重要属性和特定分类之间的所有重要关系。

本书是在课题组几年来进行多项相关科研项目研究所取得成果的基础上总结而成的，对国内、外有关的研究成果也进行了归纳总结并融入各章节的内容中。本书从基本理论概念到实际应用分析，从理论模型到算法实现和应用系统，都进行了详尽的讨论，全书分为 11 章。

第 1 章对集合论的基础知识进行了介绍。讨论了集合论的基本概念、集合代数运算以及集合关系。

第 2 章讨论了信息表知识表达系统这种 Rough 集理论的特殊处理对象，讲述了知识的分类概念、决策表等基本概念。

第 3 章介绍了 Rough 集的基本理论基础，如近似集合的概念、粗糙度与分类质量、Rough 集的代数性质和 Rough 集关系、不完备信息系统中 Rough 集理论的扩充等概念，这是以后各章节内容的基础。

第 4 章对知识获取的基本问题进行讨论，对知识获取的模型、可辨识矩阵、属性重要性以及决策规则等内容进行了介绍和分析。

第 5 章讨论知识系统不确定性表示与处理问题，对知识表示的基本方法、概率模型、可信度模型、证据理论、模糊推理等不确定性推理模型和决策表的不确定性度量问题进行了研究分析。

第 6 章讨论数据预处理问题，介绍了决策表的几种补齐算法和离散化算法。

第 7 章对信息表属性约简的理论和算法进行了介绍，讨论了

属性约简的集合观念和信息熵观念，给出了几种可行的属性约简算法。

第 8 章讨论信息表值约简问题，介绍了几种值约简算法和缺省规则获取算法。

第 9 章介绍逻辑推理系统，对逻辑推理的几种推理方法和知识系统中的不一致性问题以及不一致情况下的推理策略进行了分析讨论。

第 10 章分析介绍了几个典型的应用 Rough 集理论来解决实际问题的实例系统，如水资源调度、临床医疗诊断、客户行为预测和文本分类等。

第 11 章介绍了世界各国研究人员所开发的几个演示系统。

本书的完成，是与课题组长期的辛勤努力工作分不开的，在此要特别感谢的是课题组吴渝博士，她在项目研究工作中做了大量的工作；还有多位研究生，如常犁云、刘锋、侯利娟等。本书是大家辛勤劳动的结果。

本书作者 1998 年至 1999 年在美国 University of North Texas 作访问学者期间，该校计算机科学系教授 Paul S Fisher 博士为作者的科研工作提供了大力的支持和帮助；1999 年在加拿大 University of Regina 作访问学者期间，该校计算机科学系教授 Y. Y. Yao 博士与作者进行了很多有益的学术交流和讨论，并给作者提供了大量的参考文献资料，这无疑对本书的写作起到了很好的促进作用。南昌大学刘清教授也与作者进行过很多交流讨论并提供了大量的参考文献资料。西南交通大学靳蕃教授、重庆大学曹长修教授、重庆大学程代杰教授、中国科学院软件研究所王驹教授等也给予了作者很多鼓励和帮助。在此，对他们的支持和帮助，表示衷心的感谢。本书中还引用了许多国内外同行专家的一些研究成果，在此也对他们表示深深的谢意。

另外，还要特别感谢我的导师施鸿宝教授（现为同济大学教授）。我在西安交通大学攻读硕士学位和博士学位期间，得到了他

悉心的指导，是在他多年精心培养下，我才具有了现在研究工作的能力，走上了科学研究的道路，他对我的教导和影响也必将在我以后的科研工作中起到很大的作用。

在此，我还要感谢我的妻子何晓行女士，是她的大力支持才使我能够完成研究工作和本书的写作。我的父母亲也给予了我无微不至的关心，这些都是我完成本书的基础。

还要感谢西安交通大学出版社为本书出版给予的帮助。是大家的共同努力才使得本书能够最终出版，与读者见面。

课题组的研究工作、本书的写作完成和出版，得到了国家自然科学基金（编号：69803014）、国家 863 计划（编号：863-317-04-18-99）、攀登计划、教育部高等学校骨干教师资助计划、教育部留学回国人员科研启动基金以及重庆市应用基础研究基金的部分资助，在此一并表示诚挚的谢意。

由于作者水平有限，时间仓促，而且部分内容还是课题组所取得的阶段性研究成果，不妥、错误之处在所难免，希望能够得到读者的批评指正。

王国胤

2000 年 9 月于重庆

本专著得到下列基金资助：

- 攀登——特别支持费
- 国家自然科学基金(编号：69803014)
- 国家863计划(编号：863-317-04-18-99)
- 教育部高等学校骨干教师资助计划
- 教育部留学回国人员科研启动基金
- 重庆市应用基础研究基金

目　录

第1章　集合论基础

集合是现代数学和逻辑学的基本概念之一。本世纪以来，关于集合的理论——集合论，对现代数学和逻辑学的发展产生了巨大影响，今天它已成为数学和逻辑学的一种基础理论。集合论的创始人是康托尔(G. Cantor，1845～1918)，他所做的工作一般称为朴素集合论，由于在定义集合的方法上缺乏限制，会导致悖论。为了消除这些悖论，经过许多数学家的努力，20世纪初又创建了更精致的理论——公理化集合论，集合论至今仍在发展中。出于一些处理问题的需要，扎德(L. A. Zadeh)教授1965年提出了模糊集合的概念，模糊集理论在很多控制领域取得了很大的成功。近年来，波兰华沙理工大学坡那克(Z. Pawlak)教授等一批科学家提出了Rough集理论，研究不完整数据及不精确知识的表达、学习、归纳等方法。为了很好的理解Rough集理论，我们在本章首先对集合论进行简单的介绍。

我们首先介绍集合的基本概念，如集合、空集、子集，然后介绍定义在集合上的运算——集合代数，包括集合的基本运算(并、交、差、补)和集合运算的一些定律，最后对定义在集合上的关系进行介绍。

1.1　集合论的基本概念

用集合论的创始人康托尔曾经解释过的话来说：所谓集合，可以理解为由我们的知觉或思维确定的、能明确区分开的对象 m_i 聚集成的一个整体 M，这些对象 m_i 叫做 M 的“元素”。一般地说，集合就是把直观上或思想上的一些确定的、彼此不同的对象作为一

个整体，组成该整体的对象叫做该集合的元素。此时我们便说，这些元素组成该集合，这些元素属于该集合。

如下列集合：

1. 中华人民共和国的直辖市（北京市、上海市、天津市、重庆市）构成一个 4 元素的集合。

2. 所有三角形构成三角形集合。

3. 坐标满足方程 $x^2+y^2\leqslant R^2$ 的全部点构成（如图 1.1 所示）的点集。

通常用大写字母 $A,B,C,\cdots$ 代表集合；用小写字母 $a,b,c,\cdots$ 代表元素。

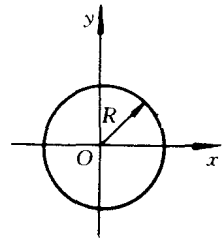

图 1.1

如果 a 是集合 A 的一个元素，则记为

$$a\in A。$$

如果 a 不是集合 A 的一个元素，则记为

$$a\notin A。$$

任一元素，对某一集合而言，或属于该集合，或不属于该集合，二者必居其一，也只居其一。

通常，集合有两种表示方法。

第一种为列举法，就是把集合中的元素一一列举出来，写在花括号内。

例 1.1 所有小于8的正整数组成的集合 A 可写成：

$$A=\{1,2,3,4,5,6,7\}。$$

例 1.2 全体自然数所组成的集合 $\mathbf{N}_+$ 可写成：

$$\mathbf{N}_+=\{1,2,3,\cdots,n,\cdots\}。$$

虽然集合 $\mathbf{N}_+$ 的元素是列举不尽的，但是例 1.2 已经列出了其中有代表性的元素，省略号表示可以继续顺次地写出它的元素。

第二种为描述法，就是用描述集合元素的共同性质的方法来表示这个集合。这种方法又叫做特征法。

例 1.3 所有小说组成的集合 A 可写成：