大数据与网络安全研究丛书

大数据下并行知识约简与知识获取

钱进 张楠 徐菲菲 著

科学出版社

大数据与网络安全研究丛书

大数据下并行知识约简与知识获取

钱 进 张 楠 徐菲菲 著

科学出版社

北 京

内 容 简 介

本书针对大数据的数据体量大、数据类型繁多、处理速度快、价值密度高等特点，以粒计算方法为理论基础，以经典粗糙集模型和区间值信息系统为研究对象，以 Hadoop 开源平台为实验环境，构建大数据下知识约简计算模型及知识获取方法。本书主要介绍大数据下 Pawlak 模型知识约简、区间值信息系统知识约简、层次粗糙集模型知识约简及知识获取的理论、模型和方法，并力求展现大数据下粒计算的最新研究成果。

本书可供计算机、自动化、应用数学等相关专业的研究人员、高校师生和工程技术人员使用。

图书在版编目(CIP)数据

大数据下并行知识约简与知识获取/钱进，张楠，徐菲菲著. —北京：科学出版社，2017.12

大数据与网络安全研究丛书

ISBN 978-7-03-055842-8

Ⅰ. ①大… Ⅱ. ①钱… ②张… ③徐… Ⅲ. ①知识获取-研究 Ⅳ. ①TP18

中国版本图书馆 CIP 数据核字(2017)第 300395 号

责任编辑：邹 杰 / 责任校对：郭瑞芝

责任印制：吴兆东 / 封面设计：迷底书装

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

北京九州逐驰传媒文化有限公司 印刷

科学出版社发行 各地新华书店经销

*

2017 年 12 月第 一 版 开本：787×1092 1/16

2017 年 12 月第一次印刷 印张：11 3/4

字数：279 000

定价：88.00 元

(如有印装质量问题，我社负责调换)

前　言

随着大数据时代的到来，各行各业不断释放出大规模复杂结构数据。数据已经不能简单地集中存储，而是分布存储在不同网络节点上。同一个样本的描述数据也不再局限于单个数据源，可能存在于多个不同数据源中。而这些复杂数据往往具有海量性、多源异构性、动态性、不确定性和知识稀疏性等特征，这给传统数据挖掘方法带来了新的机遇和挑战。如何有效地从大规模复杂数据中进行高效知识约简并发现有价值的知识已成为当前人工智能领域急需解决的科学问题。

现实世界中，数据往往是不精确、不确定、不完整的，甚至包含了大量噪声，人们对大规模多源异构不确定性数据进行整合和分析的需求也在与日俱增。要快速协同地分析这些不确定性数据，以及从不同层次为用户提供更为准确有效的层次性知识，这就必须研究一种新的面向复杂数据的智能数据分析理论、模型和方法。粒计算是当前计算智能研究领域中模拟人类思维和解决复杂问题的新方法，强调从多视角、多层次来理解和描述现实世界，通过人类粒化认知机理，对复杂问题进行不同粒度层次的抽象和处理。粒计算逐渐成为不确定性问题求解的重要理论，多粒度分析已经成为人类认知能力的重要特征。作为粒计算的三大模型之一，粗糙集理论是一种刻画不精确、不确定性的数学工具，主要利用上下近似逼近概念，通过知识约简能够直接从给定的数据中提取出简洁、易懂且有效的决策规则。这将为大数据下不确定性问题的近似建模与分析推理提供重要的理论依据。

传统基于串行计算技术的数据挖掘算法已经无法满足数据处理的时效性需求，并行计算技术可能是解决数据挖掘效率瓶颈问题的一个途径。Google 公司提出了分布式文件系统和并行编程模式 MapReduce，这为大数据挖掘提供了基础设施，同时给传统的数据挖掘研究提出了新的挑战。近几年，关于大数据下粒计算的研究引起了国内外学者的广泛关注，相继召开了 2015 年大数据与多粒度计算学术研讨会、2016 年大数据决策高峰论坛、2017 年大数据与粒计算学术研讨会等。与此同时，成立了一些大数据组织和研究机构，如中国计算机学会大数据专家委员会、大数据研究院等。此外，国际著名期刊 *Information Sciences* 还组织了 *Granular computing based machine learning in the era of big data* 专辑。如何利用大数据技术来优化和提升粒计算理论模型和方法受到众多研究者的广泛关注，成为粒计算学术界重视的一个研究方向。

本书旨在利用粒计算和粗糙集理论在复杂问题求解中的优势，使用 MapReduce 并行计算技术，研究大数据下不确定性问题的求解，开发高效的知识约简计算模型和实现知识获取方法。本书的研究工作将有助于完善粗糙集理论体系，促进粒计算研究的发展，同时有助于丰富并行编程模型，为从大数据中挖掘各种潜在的、有价值的层次性知识提供新方法、新手段。

全书共 8 章。第 1 章介绍粒计算、粗糙集理论和大数据的基本知识和研究现状；第 2 章介绍基于计数排序的高效 Pawlak 知识约简方法；第 3 章介绍区间值信息系统知识约简方法；第 4 章介绍大数据下 Pawlak 粗糙集模型知识约简方法；第 5 章介绍大数据下区间值信息系统的知识约简方法；第 6 章介绍大数据下层次粗糙集模型知识约简方法；第 7 章介绍大数据下层次粗糙集模型知识获取方法；第 8 章总结知识约简研究工作，并展望其研究趋势。本书第 1、2、4、6～8 章由钱进撰写，第 3 章由张楠撰写，第 5 章由徐菲菲撰写。

本书的出版得到了国家自然科学基金项目(项目编号：61403329)、江苏省自然科学基金项目(项目编号：BK20141152)、教育部人文社会科学基金项目(项目编号：15YJCZH129)、江苏省“青蓝工程”人才类项目、江苏理工学院科研项目的资助。在这里，对国家自然科学基金委员会、江苏省科技厅、教育部社会科学司、江苏理工学院表示诚挚的感谢。本书是三位作者在导师苗夺谦教授的指导下共同努力的结果。没有这些支持，本书不可能出版。同时，也要感谢姚一豫教授、王国胤教授、王熙照教授、梁吉业教授、李德玉教授、钱宇华教授、胡清华教授、李天瑞教授、吴伟志教授、米据生教授、周献中教授、陈德刚教授、刘文奇教授、张燕平教授、黄兵教授、徐伟华教授、李金海教授、闵帆教授、邵明文教授、王长忠教授、陈红梅教授等专家的指导和帮助，感谢同济大学 501 室的兄弟姐妹，感谢江苏理工学院计算机工程学院和科技处的帮助和支持。

最后，欢迎广大读者参与大数据下粒计算方法的研究，对于书中的不足之处，恳请批评指正(联系方式：qjqjlqyf@163.com)。

作　者

2017 年 10 月

目　录

第1章 概 论

粒计算是一个新兴的学科，主要从多粒度角度进行问题求解和信息处理，其主要理论和方法有模糊集理论、粗糙集理论、商空间理论等。粗糙集理论作为粒计算的三大模型之一，是一种刻画不精确、不确定性的数学工具，其主要利用上下近似逼近概念。知识约简是粗糙集理论的重要研究内容之一，是数据挖掘中知识获取的关键步骤。通过知识约简，可以直接从给定的数据中提取出简洁、易懂且有利用价值的知识。

1.1 粒 计 算

1.1.1 概述

随着数据库技术的迅速发展以及数据库管理系统的广泛应用，积累的数据越来越多。激增的数据背后隐藏着许多重要的信息，人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。目前，大多数数据库应用系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的模式或规则，无法根据现有数据预测未来的发展趋势。缺乏挖掘潜在知识的手段将导致数据爆炸但知识贫乏。如何帮助人们有效地收集和选择感兴趣的数据，更关键的是如何帮助用户在日益增多的数据中自动发现新的概念并自动分析它们之间的关系，使之能够真正做到信息处理的自动化，已成为信息技术领域的热点问题。知识发现(Knowledge Discovery in Database，KDD)就是为满足这种需求而诞生并迅速发展起来的，可用于开发新的信息资源。

知识发现[1,2]是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。它从数据“矿山”中找到蕴藏的知识“金块”，将信息变为知识，为知识创新和知识经济的发展做出贡献。知识发现中大多数的数据挖掘技术是在人工智能、数据库、统计学、模糊集等领域中发展起来的，所处理的对象要么是海量数据，要么是涉及大量属性的复杂数据，其复杂程度不言而喻。实际上，能够对数据模式或规则起主导作用的关键属性是有限的，即构成最终知识的属性并不多。因此，对海量数据事先进行知识约简，从不同层次进行决策分析，能够有效地降低问题求解的复杂性，提高知识发现的效率。

粒计算(Granular Computing，GrC)[3-5]是一种粒化的思维方式及方法论，是一种新的信息处理模式，而这种模式是粒化及分层思想在机器问题求解中的具体实现，成为近二十年人工智能领域的一个新研究热点。自 Zadeh 1979 年发表论文 *Fuzzy sets and information granularity* 以来，研究人员对信息粒度化的思想产生了浓厚的兴趣。Zadeh 于 1996 年提出词计算理论，认为人类的认知可以概括为信息粒度化、信息组织和因果推

理等能力；Pawlak[6]于 1982 年提出粗糙集理论，可以利用它有效地表示不确定或不精确的知识，并进行推理[7]；Hobbs[8]于 1985 年提出粒度理论，指出在不同粒度上概念化世界的能力和在不同粒度世界转换的能力是人类智能的基础；Gordon 等[9]于 1992 年指出，人的感知得益于人可以在不同的粒度层次上分析问题，并在不同粒层间转换；Love[10]于 2000 年也注意到人可以在多个抽象层次上频繁地使用和获取知识；在 Lin[11]的研究基础上，姚一豫[4,12]结合邻域系统对粒计算进行了详细的研究，发表了一系列研究成果，并将它应用于知识挖掘等领域，建立了概念之间的 if-then 规则与粒度集合之间的包含关系，提出利用由所有划分构成的格求解一致分类问题，为数据挖掘提供了新的方法和视角。姚一豫给出了粒计算的 3 种观点：①从哲学角度看，粒计算是一种结构化的思想方法；②从应用角度看，粒计算是一个通用的结构化问题求解方法；③从计算角度看，粒计算是一个信息处理的典型方法。据此，姚一豫提出了粒计算三元论(多视角、多层次粒结构和粒计算三角形)。该研究框架阐述粒计算的哲学、方法论和计算模式 3 个侧面，用来指导人们进行结构化问题和机器问题的求解。

国内外学者积极参与粒计算的研究，开展了一系列如国际国内会议、暑期学术研讨会等多种形式的交流活动。在会议方面，有现在每年举办一届的粒计算国际会议(International Conference on Granular Computing)、中国粒计算会议(CGrC)等。在暑期学术研讨会方面，2010～2017 年分别由安徽大学、同济大学、北京邮电大学和西南交通大学承办了以商空间与粒计算、不确定性与粒计算、云模型与粒计算和三支决策与粒计算为主题的研讨会，并出版了相关学术著作《商空间与粒计算——结构化问题求解理论与方法》、《不确定性与粒计算》、《云模型与粒计算》和《三支决策与粒计算》。在专著方面，2007 年，张钹和张铃出版了《问题求解理论及应用——商空间粒度计算理论及应用(第 2 版)》；海内外华人学者苗夺谦等合作出版了首部粒计算专著《粒计算：过去、现在与展望》；2010 年，周献中等出版了《不完备信息系统知识获取的粗糙集理论与方法》；2011 年，杨习贝等出版了《不完备信息系统及粗糙集理论——模型与属性约简(英文版)》；2012 年，胡清华等出版了《应用粗糙计算》；2013 年，张清华等合作出版了《多粒度知识获取与不确定性度量》等；2016 年，李天瑞等出版了《大数据挖掘的原理与方法:基于粒计算与粗糙集的视角》。以上这些学术活动的开展及专著的出版促进了粒计算理论及其应用的迅速发展。

1.1.2　粒计算内涵

1. 粒计算基本概念

从粒计算的角度看计算的对象，可形成不同的计算模型。如果从多粒度计算的角度去看，这个计算模型大体分为粒、粒层和粒结构。

(1) 粒。粒是构成粒计算模型的最基本元素，是计算模型的原语。一个粒可以看作由内部属性描述的个体元素的集合，以及由它的外部属性所描述的整体。

(2) 粒层。粒层是对问题空间或计算对象的一种抽象化描述，按照某个实际需求的粒化准则得到的所有粒子的全体构成一个粒层。同一层的粒子内部往往具有某种相同的

性质或功能。粒化程度的不同导致同一问题空间会产生不同的粒层，各个粒层的粒子具有不同的粒度，即粒的不同大小。粒计算模型的主要目标是能够在不同粒层上进行问题的求解，且不同粒层上的解能够相互转化。

(3) 粒结构。一个粒化准则对应一个粒层，不同的粒化准则对应多个粒层，粒层之间的相互联系构成一个关系结构，称为粒结构。在一般的粒计算理论中，把同一粒层的粒子看成一个集合，通常并不考虑粒子之间的结构关系。在商空间理论中[13]，粒层中的粒子间具有结构关系，因此粒结构既指粒层间的结构关系，又指粒层中的结构。

2. 粒计算的基本问题

根据粒计算的基本概念，粒计算中的两个基本问题主要为粒化和基于粒化的计算，即如何构造这个模型以及如何根据这个模型进行计算。粒化是问题空间的一个划分过程，可以简单理解为在给定粒化准则(如等价关系)下得到一个粒层的过程，是粒计算的基础。通过粒化可以得到问题空间层次间与层次内部的结构。在同一或者不同的粒化准则下均可得到多个粒层，形成多层次的网络结构。粒计算通过访问粒结构求解问题，包括在层次结构中自上而下或者自下而上两个方向的交互以及在同一层次内部的移动，即不同粒层上粒子之间的转换与推理以及同一粒层上粒子之间相互交互，从而形成基于粒化的计算。

3. 粒计算的主要模型

粒计算模型大体分为两大类：一类以处理不确定性为主要目标，如以模糊处理为基础的计算模型和以粗糙集为基础的模型；另一类则以多粒度计算为目标，如商空间理论。这两类模型的侧重点有所不同，前者在粒化过程中，侧重于计算对象的不确定性处理，Zadeh 认为“在人类推理与概念形成中，其粒度几乎都是模糊的”，因此他认为以模糊概念为基础的词计算，是粒计算的主要组成部分。以 Pawlak 为首的波兰学者提出的粗糙集理论的基础也是“思维的计算，即关于含糊、不清晰概念的近似推理”。而多粒度计算的思想则来源于 Hobbs 的如下思想：“人类问题求解的基本特征之一，就是具有从不同的粒度上观察世界，并很容易地从一个抽象层次转换到其他层次的能力，即分层次地处理它们。”因此，多粒度计算的目的是降低处理复杂问题的复杂性。

1.2 粗　糙　集

1.2.1 概述

粗糙集理论[6,7]是一种刻画不精确、不确定和不完备的数学工具。Pawlak 于 1991 年撰写的第一本粗糙集理论专著 *Rough Sets—Theoretical Aspects of Reasoning about Data* 的问世和 Slowinski 于 1992 年主编的关于粗糙集应用与相关方法比较研究论文集的出版，极大地推动了粗糙集理论研究。自 1992 年以来，国际上每年都召开以粗糙集理论为主题的学术研讨会，如 RSCTC、RSKT、GrC、RSFDGrC 等，而国内 2001 年 5 月在重庆召

开了第一届中国 Rough 集与软计算学术研讨会，此后每年举办一次，以及与之联合召开的中国 Web 智能学术研讨会和中国粒计算学术研讨会。近二十年来，国内外学者发表了大量高水平学术论文和专著，进一步促进了粗糙集理论的发展。

目前，粗糙集理论中知识约简能够帮助解决“数据丰富、知识缺乏”这一难题，而且获得了广泛的应用和巨大的成功，但也面临许多问题和挑战。经典知识约简算法主要处理离散型数据，然而现实世界中经常会遇到包含缺失值、连续值的数据，这使实际应用效果不是很理想。为此，许多学者扩展了经典粗糙集模型，提出了模糊粗糙集模型[14]、决策粗糙集模型[15]、变精度粗糙集模型[16]、相容粗糙集模型[17]、相似粗糙集模型[18]、覆盖粗糙集模型[19,20]、区间值粗糙集模型[21,22]、层次粗糙集模型[23]等。面对海量的数据，并行约简可能是一个重要的途径。于是，许多学者提出了一些并行知识约简算法[24-32]，但是仍然没有解决海量数据的知识约简问题。而当前业内并无有效的并行计算解决方案，无论是编程模型、开发语言还是开发工具，距离开发实用的数据挖掘平台还有很大的差距[33-37]。

1.2.2　基本概念

粗糙集理论将知识理解为分类能力，其主要思想是从数据中挖掘规则，利用知识库中的知识来刻画不精确或不确定目标概念。本章在回顾粗糙集理论基本概念的基础上，详细分析和比较已有的基于正区域的知识约简算法、基于差别矩阵的知识约简算法和基于信息熵的知识约简算法，并阐述影响传统知识约简算法效率的关键因素。

下面简要介绍本书主要用到的一些Rough集的基本概念，具体请参考文献[6]和文献[7]。

定义 1.1　四元组 $S=(U,\mathrm{At},\{V_a\mid a\in\mathrm{At}\},\{I_a\mid a\in\mathrm{At}\})$是一个信息系统，其中 $U=\{x_1,x_2,\cdots,x_n\}$表示对象的非空有限集合，称为论域；At 为全体属性集；V_a 是属性 $a\in\mathrm{At}$ 的值域；$I_a: U\to V_a$ 是一个信息函数，它为每个对象赋予一个信息值。每一个属性子集 $A\subseteq\mathrm{At}$ 决定了一个二元不可区分关系 IND(A)：

$$\mathrm{IND}(A)=\{(x,y)\in U\times U\mid\forall a\in A, I_a(x)=I_a(y)\} \tag{1.1}$$

关系 IND(A)构成了 U 的一个划分，用 U/IND(A)表示，简记为 U/A 或 π_A。U/A 中的任何元素 $[x]_A=\{y\mid\forall a\in A,\ I_a(x)=I_a(y)\}$称为等价类。

在分类问题中，可将 At 分成条件属性集 C 和决策属性集 D 两部分，即 $\mathrm{At}=C\cup D$ 且 $C\cap D=\varnothing$，其中，$C=\{c_1,c_2,\cdots,c_m\}$表示条件属性的非空有限集，D 表示决策属性的非空有限集，一个信息系统 S 就变成了一个决策表。需要说明的是，这里所讨论的决策表均为完备的，并且假设 $D=\{d\}$，表示决策表仅有单一决策属性。若决策表中存在多个决策属性，则可将所有决策属性的属性组合值映射为不同的单一的新决策值，从而将多个决策属性转化为单个决策属性。U/C 中任一元素称为论域关于条件属性集 C 的条件类，U/D 中任一元素称为决策类。显然，划分 U/C 具有最细的信息粒度而划分 $U/\varnothing$ 具有最粗的信息粒度。

定义 1.2　在决策表 $S=(U,\mathrm{At}=C\cup D,\{V_a\mid a\in\mathrm{At}\},\{I_a\mid a\in\mathrm{At}\})$中，对于每个子集 $X\subseteq U$ 和不可区分关系 $A\subseteq C\cup D$, X 的下近似集与上近似集分别可以由 A 的基本集定义

如下：

$$\underline{A}X = \cup\{x \in U \mid [x]_A \subseteq X\} \tag{1.2}$$

$$\overline{A}X = \cup\{x \in U \mid [x]_A \cap X \neq \varnothing\} \tag{1.3}$$

下近似集由肯定属于 X 的对象集构成，表示根据现有的知识可以判断出肯定属于 X 的对象所组成的最大集合，上近似集由可能属于 X 的对象集构成，表示根据现有知识判断出可能属于 X 的对象所组成的最小集合。通过两个精确的上、下近似集，可从两个侧面对概念 X 进行逼近，从而可以近似地描述概念 X。

下近似 $\underline{A}X$ 也称为 X 关于 A 的正域，记为 $\mathrm{POS}_A(X)$。上近似集与下近似集的差别部分称为 X 关于 A 的边界域，即 $\mathrm{BND}_A(X) = \overline{A}X - \underline{A}X$，表示不能完全确定是否属于 X 的对象所组成的集合，刻画了关于概念 X 分类的不确定对象。$\mathrm{NEG}_A(X) = U - \overline{A}X$ 称为 X 关于 A 的负域，表示根据现有的知识判断出肯定不属于 X 的对象所组成的集合。若 $\underline{A}X = \overline{A}X$，即 $\mathrm{BND}_A(X) = \varnothing$，则概念 X 是精确的，说明根据现有知识能够确定 X 中的所有分类对象，否则概念 X 是粗糙的，即 $\mathrm{BND}_A(X) \neq \varnothing$，说明 X 中存在不确定的分类对象。

定义 1.3　在决策表 $S = (U,\ \mathrm{At} = C \cup D,\ \{V_a \mid a \in \mathrm{At}\},\ \{I_a \mid a \in \mathrm{At}\})$中，$\forall A \subseteq C$，$X \subseteq U$，决策属性 D 关于 A 的正区域 $\mathrm{POS}_A(D)$ 定义为

$$\mathrm{POS}_A(D) = \bigcup_{X \in U/D} \underline{A}X \tag{1.4}$$

在决策表 S 中，决策属性 D 导出的 U 上划分记为 $\pi_D = \{D_1,\ D_2,\ \cdots,\ D_k\}$。

定义 1.4　在决策表 $S=(U, \mathrm{At} = C \cup D, \{V_a \mid a \in \mathrm{At}\}, \{I_a \mid a \in \mathrm{At}\})$中，称 $\mathrm{POS}_C(D)$ 中的对象为相容对象，即 $\mathrm{POS}_C(D) = \cup_{i=1}^{k} \underline{C}D_i$；称 $U - \mathrm{POS}_C(D)$ 中的对象为矛盾对象，记为 $\underline{C}D_{k+1}$。若 $\mathrm{POS}_C(D) = U$, 则称决策表 S 是一致决策表，$\underline{C}D_{k+1} = \varnothing$；否则是不一致决策表。

将决策表 S 中所有矛盾对象归为一类，划分$\{\underline{C}D_1,\ \underline{C}D_2,\ \ldots,\ \underline{C}D_k,\ \underline{C}D_{k+1}\}$则既可以将属于不同决策类的相容对象分开，又可以将相容对象与矛盾对象分开，这样的不一致决策表就可以看成“相容”决策表了。因此，相容决策表不过是不一致决策表的“特例”。

定理 1.1　在决策表 S 中，$A \subseteq C$，则 $\mathrm{POS}_A(D) = \mathrm{POS}_C(D)$ 的充分必要条件是 $\underline{A}D_i = \underline{C}D_i$，$\forall i \in \{1,2,\cdots,k,k+1\}$。

证明　由定义 1.3 和定义 1.4 直接证得。

定义 1.5　在决策表 $S = (U, \mathrm{At} = C \cup D, \{V_a \mid a \in \mathrm{At}\}, \{I_a \mid a \in \mathrm{At}\})$中，记 $U/C = \{[x_1']_C,\ [x_2']_C,\ \cdots,\ [x_s']_C\}$，$U' = \{x_1',\ x_2',\ \cdots,\ x_s'\}$，$U'_{\mathrm{POS}} = \{x'_{i_1},\ x'_{i_2},\ \cdots,\ x'_{i_t}\}$，其中，$U'_{\mathrm{POS}}$ 中的对象为相容对象，U'_{BND} 为 $U' - U'_{\mathrm{POS}}$，则 $S' = (U' = U'_{\mathrm{POS}} \cup U'_{\mathrm{BND}},\ \mathrm{At} = C \cup D, \{V_a \mid a \in \mathrm{At}\}, \{I_a \mid a \in \mathrm{At}\})$为简化决策表。

不一致决策表中的对象可分为相容对象和不相容对象，故相容决策表可看作不一致决策表的特例，而简化决策表则是从不一致决策表中删除相容等价类和不相容等价类中冗余的对象。对象、等价类和决策表可理解为分析问题的 3 个视角层次，分别为从低到高、从具体到抽象。

定义 1.6　在决策表 $S=(U,\mathrm{At}=C\cup D,\{V_a|a\in\mathrm{At}\},\{I_a|a\in\mathrm{At}\})$中，$a\in C$，若 $\mathrm{POS}_{C-\{a\}}(D)\neq\mathrm{POS}_C(D)$，则称属性 a 在 C 中是不可缺少的，否则称属性 a 在 C 中是不必要的。C 中所有不可缺少的属性集合称为 C 的 D-核(简称核)，记为 $\mathrm{CORE}_D(C)$。

1.3　知 识 约 简

知识约简是粗糙集理论的重要研究内容和热点之一，也是知识获取的关键步骤。所谓知识约简是指在不影响知识表达能力的条件下，通过消除冗余知识，从而获得知识库简洁表达的方法。研究表明，信息系统中有些属性是冗余的，若将这些属性删除，不仅不会改变信息系统的分类或决策能力，反而会提高系统潜在知识的清晰度。知识约简反映了一个决策表的本质属性，一般先自底向上逐步增加属性，然后对结果采用逐步删除冗余属性的搜索策略来获取关键属性。通过知识约简可以导出现实问题的更简单、对决策更有效的决策规则，从而帮助做出一些预测或辅助决策。现有知识约简方法主要包括基于正区域的知识约简算法、基于差别矩阵的知识约简算法和基于信息熵的知识约简算法等，一般采用以下搜索策略：①以空集为起点，自底向上逐步增加属性来计算约简；②以初始条件属性集为起点，采用逐步删除属性来获取约简；③先自底向上逐步增加属性，然后对结果逐步删除冗余属性。策略 1 和策略 2 一般得到的是约简的超集，通常采用策略 3 进行知识约简。计算所有约简是 NP-Hard 问题，因此现有的知识约简方法采用启发式方式获取一个约简。

为了提高知识约简效率，达到知识获取实用化的目的，许多学者提出了基于正区域的知识约简、基于差别矩阵的知识约简和基于信息熵的知识约简等方法。粗糙集创始人 Pawlak[6,7]从形式上定义了基于正区域的知识约简模型，该方法主要保持约简前后正区域对象不变以保证确定性规则的分类能力不变化。Skowron 等[38]提出了基于差别矩阵的知识约简算法，该算法简单、易理解。苗夺谦等[39,40]将信息论的思想引入粗糙集，运用信息熵来表达知识的粗糙度，并提出基于互信息的启发式知识约简算法，并深入探讨了粗糙集理论中基本概念和运算的信息表示，进一步讨论了粗糙集理论中知识的粗糙性和信息熵的关系。该启发式算法简单实用，能够推动粗糙集理论在实际问题中的应用。许多学者在这些算法的基础上改进并提出了一些高效知识约简算法。

1.3.1　基于正区域的知识约简算法

定义 1.7　给定决策表 $S=(U,\mathrm{At}=C\cup D,\{V_a|a\in\mathrm{At}\},\{I_a|a\in\mathrm{At}\})$，一个属性集 $A\subseteq C$ 是 C 的 D-约简，如果

(1) $\mathrm{POS}_A(D)=\mathrm{POS}_C(D)$；

(2) $\forall a\in A$，$\mathrm{POS}_{A-\{a\}}(D)\neq\mathrm{POS}_A(D)$。

由定义 1.6 和定义 1.7 可得

$$\mathrm{CORE}_D(C)=\cap\mathrm{Red}_D(C) \tag{1.5}$$

其中，$\mathrm{Red}_D(C)$ 表示所有相对约简。

定义 1.8　给定决策表 $S=(U,\mathrm{At}=C\cup D,\{V_a\mid a\in\mathrm{At}\},\{I_a\mid a\in\mathrm{At}\})$，则 D 对 C 的依赖度定义如下：

$$r_C(D)=\frac{|\mathrm{POS}_C(D)|}{|U|} \tag{1.6}$$

称 D 以依赖度 $r_A(D)$ 依赖于 A，并称

(1) D 完全依赖于 C，当 $r_C(D)=1$ 时；

(2) D 部分依赖于 C，当 $0<r_C(D)<1$ 时；

(3) D 独立于 C，当 $r_C(D)=0$ 时。

依赖度 $r_A(D)$ 表示利用条件属性集 A 可以被正确分类到 U/D 中对象的个数占论域对象总数的比例。这样，可以利用 $r_{A\cup a}(D)$ 来度量属性 a 的重要性。由于 $0\leqslant r_{A\cup a}(D)\leqslant 1$ 并且 $r_{A\cup a}(D)$ 具有单调性，可以使用它构建一种基于正区域的知识约简算法。

基于正区域的知识约简算法获得的约简结果通常不是唯一的，人们希望能找到具有最少属性的约简，即最佳约简。然而，找到一个最佳约简是一个 NP-Hard 问题，解决这一问题通常采用启发式搜索方法，求出最佳或次最佳约简。文献[41]利用快速排序算法，提出了基于正区域的渐增式计算的属性约简算法，其时间复杂度为 $O(|C|^2|U|\log|U|)$。文献[42]以基数排序思想设计了一个新的计算 U/P 算法，在已知的信息 U/P 上递归地求出 $U/(P\cup\{a\})$，得到时间复杂度为 $\max(O(|U||C|),O(|C|^2|U/C|))$ 的属性约简算法。文献[43]提出了基于 Hash 的正区域计算方法和知识约简算法，将基于正区域的知识约简算法的时间复杂度降为 $O(|C|^2|U/C|)$。文献[44]利用正向近似思想，提出了一种知识约简算法框架模型，能够将基于正区域的约简算法时间复杂度降为 $O(|U||C|+\sum_{i=1}^{|C|}(|C|-i+1))$。

1.3.2　基于差别矩阵的知识约简算法

Skowron 等[38]提出一个用来存储任意两个对象之间差别信息的差别矩阵，通过该差别矩阵可以计算一个约简或所有约简。在该差别矩阵中，定义 $\mathrm{pos}(x_i)$ 表示对象 x_i 是否属于正区域，利用该信息来帮助判断任意两个对象之间是否存在差别信息。针对不一致决策表所产生的差别矩阵问题和差别矩阵空间复杂度，目前提出了许多基于差别矩阵的改进算法。文献[45]给出了属性序下时间复杂度为 $O(|C|^2|U|^2)$ 的差别矩阵属性约简算法。文献[46]对差别矩阵进行简化，主要是不用生成差别矩阵，就给出了时间复杂度为 $O(|C|^2|U|\log|U|)$ 的属性约简算法。文献[47]给出了时间复杂度为 $\max(O(|C|^2(|U'_{\mathrm{pos}}||U/C|)),O(\sum_{i=1}^{|C|}|k_i||U|))$ 的基于简化差别矩阵的属性约简算法。文献[48]利用分治策略思想提出了属性序下属性约简算法，其平均时间复杂度为 $O(|U||C|(|C|+\log|U|))$，空间复杂度为 $O(|C|+|U|)$。文献[49]具体探讨了 Pawlak 粗糙集模型下各种知识约简算法中性质保持的含义，给出了一个广义的性质保持定义。文献[50]和文献[51]提出了条件信息熵的概念，给出了时间复杂度为 $O(|C|^2|U|^2)$ 的属性约简算法。文献[52]分析了现有条件信息熵的不足，给出了时间复杂度为 $O(|C|^2|U|\log|U|)$ 的基于新的条件熵的高效知识约简算法。文献[53]引入决策表中

基于条件信息熵的近似约简概念，提出时间复杂度为$O(|C|^2|U|^2)$的基于条件信息熵的近似约简算法。

下面首先探讨一些主要的差别矩阵定义[38,47,54-57]。

定义 1.9[38]　给定一个决策表S，对应的差别矩阵M^C中任一元素定义为

$$m_{ij}=\begin{cases}\{a\in C\mid I_a(x_i)\neq I_a(x_j)\}, & x_i,x_j\in \mathrm{POS}_C(D)\\ & \text{或}\mathrm{pos}(x_i)\neq \mathrm{pos}(x_j)\\ \varnothing, \quad \text{其他}\end{cases} \tag{1.7}$$

在定义 1.9 中，Skowron 等提出的差别矩阵考虑到了不一致决策表的情况。

定义 1.10[54]　给定一个决策表S，对应的差别矩阵M_1^C中任一元素定义为

$$m_{ij}=\begin{cases}\{a\in C\mid I_a(x_i)\neq I_a(x_j)\}, & I_D(x_i)\neq I_D(x_j)\\ \varnothing, \quad \text{其他}\end{cases} \tag{1.8}$$

文献[54]中得出了如下结论：当且仅当某个m_{ij}为单属性时，该属性属于$\mathrm{CORE}_D(C)$。该结论在大量的 Rough 集理论文献中被引用。文献[55]对文献[54]中的这个结论提出了质疑，举例说明了该方法的缺陷，提出了新的差别矩阵定义并给出了求核的方法。

定义 1.11[55]　给定一个决策表S，对应的差别矩阵M_2^C中任一元素定义为

$$m'_{ij}=\begin{cases}m_{ij}, & \min\{|D(x_i)|,|D(x_j)|\}=1\\ \varnothing, \quad \text{其他}\end{cases} \tag{1.9}$$

其中，$D(x_i)$表示与对象x_i相同对象的所有不同决策值的集合。

文献[55]给出并证明了如下结论：当且仅当某个m'_{ij}为单属性时，该属性属于$\mathrm{CORE}_D(C)$。但在构造差别矩阵时，对每个矩阵元素m'_{ij}均要求给出并判定$\min\{|D(x_i)|,|D(x_j)|\}$的值，因此计算代价高。

定义 1.12[56]　给定一个决策表S，对应的差别矩阵M_3^C中任一元素定义为

$$m''_{ij}=\begin{cases}\{a\in C\mid I_a(x_i)\neq I_a(x_j)\}, & I_D(x_i)\neq I_D(x_j)\}\text{且}x_i\in U_1,x_j\in U_1\\ \{a\in C\mid I_a(x_i)\neq I_a(x_j)\}, & x_i\in U_1,x_j\in U'_2\\ \varnothing, \quad \text{其他}\end{cases} \tag{1.10}$$

其中，$U_1=\bigcup_{i=1}^{k}\underline{C}D_i$；$U_2=U-U_1$，$U'_2$是对$U_2$删除多余的不相容对象后的集合(不同的不相容对象只保留一个)。

文献[56]给出并证明了如下结论：当且仅当某个m''_{ij}为单属性时，该属性属于$\mathrm{CORE}_D(C)$。但文献[56]仅对不相容对象进行了处理，对相容对象没有进行处理。另外，差别矩阵中还存在许多空集元素。

定义 1.13[57]　给定一个决策表S，记$U/C=\{[x'_1]_C,[x'_2]_C,\cdots,[x'_s]_C\}$，$U'=\{x'_1,x'_2,\cdots,x'_s\}$，$U'_{\mathrm{POS}}=\{x'_{i_1},x'_{i_2},\cdots,x'_{i_t}\}$，其中，$U'_{\mathrm{POS}}$中对象为相容对象，$U'_{\mathrm{BND}}$为$U'-U'_{\mathrm{POS}}$，则$S'=(U'=U'_{\mathrm{POS}}\cup U'_{\mathrm{BND}}$，$\mathrm{At}=C\cup D$，$\{V_a\mid a\in\mathrm{At}\}$，$\{I_a\mid a\in\mathrm{At}\})$为简化决策