粒计算的不确定性分析与知识获取方法

孙 林 徐久成 著

科学出版社

软计算的不确定性分析与知识获取方法

国家自然科学基金项目资助
中国博士后科学基金面上项目资助
河南省科技攻关项目资助
河南省高等学校青年骨干教师培养计划资助
新乡市科技攻关计划项目资助
河南师范大学学术专著出版基金资助
河南师范大学博士科研启动费支持课题资助

粒计算的不确定性分析与知识获取方法

孙　林　徐久成　著

科 学 出 版 社
北　京

内 容 简 介

粒计算是当前人工智能研究领域中模拟人类思维和解决复杂问题的新理论与新方法，它涵盖了所有有关粒度的理论、方法和技术，是研究大规模复杂问题求解、大数据分析与挖掘、不确定性智能信息处理的有力工具。经过十多年的发展，在与多学科交叉研究过程中，粒计算正逐步形成其特有的研究体系和内容。本书介绍了粒计算的不确定性分析与知识获取方法的最新进展，内容涉及粒计算的基本理论、粒计算的不确定性分析、粒计算的扩展模型、基于粒计算的知识获取方法。

本书可供计算机、模式识别与智能系统、自动化等相关专业的研究人员、教师、研究生、高年级本科生阅读，也可供相关领域工程技术人员参考。

图书在版编目 (CIP) 数据

粒计算的不确定性分析与知识获取方法 / 孙林，徐久成著.
—北京：科学出版社，2018.1
ISBN 978-7-03-054337-0

Ⅰ. ①粒… Ⅱ. ①孙… ②徐… Ⅲ. ①人工智能－计算方法 Ⅳ. ①TP18

中国版本图书馆 CIP 数据核字(2017)第 215816 号

责任编辑：王 哲 纪四稳 / 责任校对：郭瑞芝
责任印制：张 伟 / 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

北京中石油彩色印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2018 年 1 月第 一 版 开本：720×1 000 B5
2018 年 1 月第一次印刷 印张：18
字数：360 000

定价：98.00 元

(如有印装质量问题，我社负责调换)

前　言

作为知识表示和数据挖掘的重要工具，粒计算是模拟人类思考问题和解决大规模复杂问题自然模式的一种新的理论、技术与方法，是在问题求解过程中使用信息粒，从不同角度、不同层次对现实问题进行描述、推理与求解。尽管“粒计算”术语于 1997 年正式提出，但与之相关的研究自 1992 年起就已经出现在一些重要的国际学术会议上。到目前为止，研究者已经对粒计算理论及其应用做了大量有意义的探索，使得粒计算的研究范围非常广泛，所有与粒度相关的理论、方法、技术和工具都可归为粒计算的研究范畴。粒计算改变了传统的计算观念，在海量数据挖掘的研究中享有独特的优势，在智能系统的设计和实现中起着重要作用。目前，粒计算在智能信息处理领域中得到了众多专家学者的广泛关注。随着粒计算研究工作的不断深入，人们从不同的角度研究得到不同的粒计算理论模型。当前，这些理论与方法依靠各自的特点和优势已经广泛应用于对不确定、不精确、不完整信息的处理，以及对大规模海量数据的挖掘和对复杂问题的求解等。粒计算理论模型中的知识粒具有不确定性，它直接影响问题求解的效率和精度。不确定性人工智能的新时代已经到来，研究不确定性信息的表示、处理和模拟，让机器模拟人类认识客观世界和人类自身的认知过程，使机器具有不确定性智能，已经成为人工智能领域的一个重要课题。于是，探索有效的知识不确定性分析技术及粒计算的知识获取理论与方法是当前粒计算研究的一项重要任务。

近年来，粒计算研究已经取得了令人鼓舞的成果，国内外学者在粒计算理论、不确定性分析及知识获取等方面进行了较为深入的研究，开展了国际、国内学术研讨会和暑期研讨会等多种形式的粒计算学术交流。结合粒计算的相关专题，国内外华人学者相继出版了一系列著作：《问题求解理论及应用——商空间粒度计算理论及应用(第 2 版)》(清华大学出版社，2007 年)、《粒计算：过去、现在与展望》(科学出版社，2008 年)、《商空间与粒计算——结构化问题求解理论与方法》(科学出版社，2010 年)、《不确定性与粒计算》(科学出版社，2011 年)、《决策粗糙集理论及其研究进展》(科学出版社，2011 年)、《云模型与粒计算》(科学出版社，2012 年)、《多粒度知识获取与不确定性度量》(科学出版社，2013 年)、《三支决策与粒计算》(科学出版社，2013 年)、《粒计算及其不确定信息度量的理论与方法》(科学出版社，2013 年)、《基于包含度的粒计算方法与应用》(科学出版社，2015 年)、《大数据挖掘的原理与方法——基于粒计算与粗糙集的视角》(科学出版社，2016 年)等。这些学术成果不仅能为大规模复杂系统的数据挖掘及不确定性分析提供新的技术与

方法，而且对推动决策管理、智能制造、故障诊断、工业机器人、无人驾驶、地震预报、采矿工程、空间数据分析、医疗诊断与预防等领域中大数据技术的发展具有重要的科学意义和应用价值，也进一步促进了粒计算理论、方法、模型及其应用在各个领域中的发展。

本书总结了粒计算研究组近年来在粒计算理论及其应用方面取得的最新研究成果和相关研究进展，旨在将粒计算与不确定性分析相结合，运用粒计算的理论与方法来度量和处理不确定信息，分析信息粒化与不确定性的关系，研究基于粒计算的不确定性度量的理论与方法。全书共 4 章，分为三部分内容，章节总体按照从理论到应用的思路进行编排。第一部分为粒计算的相关研究及基本理论概述，主要在第 1 章粒计算的基本理论中介绍。第二部分为粒计算的不确定性理论及模型研究，包括第 2 章粒计算的不确定性分析和第 3 章粒计算的扩展模型。第三部分为粒计算理论与方法的应用研究，主要是第 4 章基于粒计算的知识获取方法。各章之间内容相对独立，自成体系，但都紧密围绕粒计算的不确定性分析与知识获取方法的主题展开。本书的内容引用了粒计算领域相关专家学者及作者前期的一些研究成果，同时也包含作者部分最新的研究成果。因此本书既是对前期研究工作的总结，也是对未来研究的展望，可为读者进一步研究粒计算的理论与方法提供一些参考和帮助。

撰写本书时，引用和参考了国际、国内与粒计算有关的诸多研究成果，在这里对所涉及的专家、学者、研究人员表示深深的谢意。本书中所列的参考文献未能覆盖全部相关的研究成果，在此也对那些可能被遗漏文献的作者一并表示衷心的感谢。此外，孟慧丽、张灵均、施玉杰、李亚鸽等老师，以及硕士研究生张霄雨、孟新超、刘弱南、王云、王楠、徐战威等对书中的部分章节进行了整理和校对，在此同样表示感谢。同时，感谢科学出版社在本书的出版过程中给予的大力帮助和支持，也感谢国家自然科学基金项目(61772176、61402153、61370169、60873104、61040037)、中国博士后科学基金面上项目(2016M602247)、河南省科技攻关项目(142102210056、162102210261)、新乡市科技攻关计划项目(CXGG17002)、河南师范大学学术专著出版基金、河南师范大学博士科研启动费支持课题(QD15132)的资助。

由于作者自身知识水平有限，书中难免存在不妥之处，恳请读者朋友批评指正。

作　者

2017 年 3 月

目　　录

第 1 章　粒计算的基本理论

1.1　引　　言

根据维基百科的定义，大数据是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合。大数据=海量数据+复杂类型的数据[1-4]。伴随着移动互联网、社交网络、电子商务、物联网、云计算等技术和应用的飞速发展，数据量呈几何倍数增长，加速了大数据时代的到来，致使数据从简单的处理对象开始转变为一种基础性资源[5-7]。这些数据来源广泛，其中最主要的有科学研究、社交、搜索、电商、微博、微信、传感器、智慧地球、车联网、GPS、医学影像、安全监控、金融(银行、股市、保险)、移动通信(通话、短信)等，如果这些大数据能够互通共享、融合分析，将会给人类带来巨大的经济与社会效益[8-14]。截至 2011 年年底，中国互联网行业持有的数据总量为 1.9EB[7]。截至 2015 年 12 月，我国网民规模达 6.88 亿人。其中，手机网民规模达 6.20 亿人，移动互联网流量呈现爆发式增长。工业和信息化部统计显示，2015 年仅移动互联网接入的流量超过 3.8EB，同比增长 103%。根据国际数据公司(IDC)的估计，未来全球数据总量年增长率将维持在 50%左右，到 2020 年，全球数据总量将达到 40ZB，其中，我国数据量将达到 8.6ZB[7,15-17]。大数据在现代信息社会中的数据资源主体地位已成为学术界与企业界的共识。由于对经济活动与社会发展具有可预见的重要推动作用，大数据已经进入世界主要经济体的战略研究计划中。正如美国政府启动的“Big Data Research and Development Initiative”计划指出的“将大力推进大数据的收集、访问、组织和开发利用等相关技术的发展，提高从海量复杂的数据中提炼信息和获取知识的能力与水平”。从大数据中进行数据挖掘与知识发现是大数据应用的战略问题之一。

从大数据的外在来看，大数据经常呈现出大规模性、多模态性与增长性等特征，使得传统的数据分析理论、方法与技术面临可计算性、有效性与时效性等严峻挑战。粒计算是专门研究基于粒结构的思维模式、问题求解方法、信息处理模式的理论、技术和工具的学科，是当前智能信息处理领域中一种新的计算范式。通过分析大数据的表现形态、大数据挖掘面临的挑战与粒计算核心理念的内在关系可知，大数据自身具有天然的多层次/多粒度特性，数据挖掘任务也经常呈现多层次/多粒度特性，而大数据挖掘算法本身也要求可计算性、有效性、高效近似求解特性。这表明大数据的分析需求和粒计算框架有很强的契合性。

粒计算(granular computing，GrC)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。它覆盖了所有有关粒度的理论、方法和技术，是复杂问题求解、海量数据挖掘、模糊信息处理的有效工具[18]。粒计算作为一种方法论，旨在求解问题的过程中，用合适的“粒”作为处理的对象，从而在保证求得满意解的前提下，提高解决问题的效率，其合适的粒度常常是由提出的问题及问题的环境来决定的，这一点对设计基于粒计算的数据处理框架有非常重要的意义。截至目前，国内外研究人员已经对粒计算理论和模型进行了深入的研究，同时也将这些理论和模型与其他计算智能、机器学习的技术相结合，取得了丰硕的研究成果[19-32]。

根据粒计算的本质和宗旨，粒计算模拟了人类的思维规律，从粒计算的角度来思考大数据处理，从数据源头和处理过程考虑数据的合适粒度，将会使大数据处理理论和方法具有重要的科学研究意义和价值[33]。粒计算在解决大规模复杂问题及问题求解过程中使用了信息粒和信息粒化，作为一种信息处理范式得到了快速发展和研究，已有学者认为粒计算技术是处理大数据的最有效方法之一[34,35]。目前，粒计算已成为智能信息处理领域中一个非常活跃的研究方向[36-76]。对于粒计算，大数据的处理会对其产生两个重要作用：其一，如张钹院士指出，在大数据背景下，需要重视概率论和统计学在粒计算中的应用；其二，在深度学习、贝叶斯网络等受到广泛关注的今天，数据的表示仅仅使用基于集合的形式已经不再足够，应该在粒化基础上进一步探索信息粒之间的结构关系。针对大数据的特性，王国胤等提出统一的大数据问题粒计算解决框架[33]。针对大数据的规模性、多模态性与增长性给传统的数据挖掘方法带来的挑战，梁吉业等从粒计算的视角分析了应对这些挑战的新思路和新策略[77]。

针对大数据中普遍存在的不确定性，目前的数据集成技术和粒化方法并不能完全消除这些不确定性，因此，不确定性处理和概率数据挖掘技术已经成为当前大数据技术要解决的一个重大问题[33,77,78]。从粒计算的观点来看，人们对问题的分析及获取知识的表示都具有粒度性，这既与认知主体的主观局限有关，也受观测工具的客观因素影响。因此，粒计算模型中的知识粒具有不确定性，它直接影响粒计算方法解决复杂问题的效率和精度[79]。迄今为止，各种粒计算模型中知识的不确定性分析引起了众多学者的关注和研究，不确定性分析的研究对粒计算的发展也具有很大的推动作用[80]。

为了更好地刻画这一客观世界，一种新的结构化思维方式——粒化思维方式应运而生，粒计算正是系统研究粒化思维方式及其方法的一门新兴学科。在过去的十几年中，人们看到了粒计算研究的快速发展以及研究者对其急速增长的研究兴趣，许多粒计算的模型和方法以及扩展的模型和方法被提出[81]。这些结果增强了人们对粒计算的理解。具体地，粒计算理论是模糊集理论、粗糙集理论、商空间理论、云模型理论等具体粒计算模型的超集[18,19,82]。粗略地讲，凡是在分析问题和求解问题中，应用了分组、分类、分治和聚类手段的一切理论与方法均属于粒计算的范畴。粒计算模型大体分为两大类：一类以处理不确定性问题为主要目标，如以模糊集理

论为基础的词计算模型；另一类则以多粒计算为目标，如商空间理论、粗糙集理论和云模型等。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问题处理，后者多粒计算的目的是降低处理复杂问题的复杂性。随着粒计算研究工作的发展，粒计算模型的种类也层出不穷[36-41]。这里在文献[36]～[39]、[41]和[81]的粒计算理论基本研究内容的基础上，简单介绍粒计算模型中具有代表性的 4 个主要模型，即模糊集理论模型、粗糙集理论模型、商空间理论模型和云模型，以加深读者对粒计算相关理论和方法的理解。

1.2 模糊集理论

简单来说，Fuzzy 集(模糊集)就是用来表达模糊性概念的集合。模糊集理论是在经典集合论基础上发展起来的用以刻画模糊现象的一种理论方法。由于经典集合只能表现具有明确外延的概念，不能表现模糊概念，而在现实生活中，模糊现象是大量存在的。为了定量地刻画模糊概念和模糊现象，美国计算机与控制论专家 Zadeh 教授于 1965 年提出了模糊集这一重要概念，且目前在许多领域中都获得了卓有成效的应用。本节简单介绍模糊集的理论基础，包括模糊集的定义与表示、模糊集的运算和模糊性的度量。

1. 模糊集的定义与表示法

定义 1.1[37] 设 X 为一有限非空论域，x 为 X 中的元素，对于$\forall x\in X$，给定映射 $x\to\mu_A(x)\in[0, 1]$，则序偶组成的集合：

$$A=\{x \mid \mu_A(x)\},\ \forall x\in X$$

为 X 上的模糊集合。其中，$\mu_A(x)$为 x 对 A 的隶属度，μ 为隶属函数。

X 上的模糊集合的全体记做 $F(X)$，$\mu_A(x)$可简记为 $A(x)$，隶属度 $\mu_A(x)$表示 x 隶属于模糊集 A 的程度。

若 $\mu_A(x)=1$，则认为 x 完全属于 A；

若 $\mu_A(x)=0$，则认为 x 完全不属于 A；

若 $0<\mu_A(x)<1$，则认为 x 在 $\mu_A(x)$程度上属于 A。

设论域 $X=\{x_1, x_2,\cdots, x_n\}$，$X$ 上的任一模糊集合 A，其隶属函数为 $A(x_i)$ $(i=1, 2,\cdots, n)$，则模糊集合有如下三种表示方法。

(1) Zadeh 表示法：

$$A=\frac{A(x_1)}{x_1}+\frac{A(x_2)}{x_2}+\cdots+\frac{A(x_n)}{x_n}$$

其中，$\dfrac{A(x_i)}{x_i}$不表示“分数”，仅表示元素 x_i 隶属于 A 的程度为 $A(x_i)$；符号“+”也不表示“加号”，而是一种联系符号。

(2) 序偶表示法：

$$A = \{(x_1, A(x_1)), (x_2, A(x_2)), \cdots, (x_n, A(x_n))\}$$

(3) 向量表示法：

$$A = (A(x_1), A(x_2), \cdots, A(x_n))$$

向量表示法要求论域 X 中元素的先后顺序是确定的，把每个元素对应的隶属度按给定的顺序组成一个向量，该向量称为模糊向量[83]。在这里要注意，向量中如有 $A(x_i) = 0$，则该项不能省略。

2. 模糊集的基本运算与性质

由于模糊集中没有点和集之间的绝对属于关系，所以其运算的定义只能以隶属函数间的关系来确定。

定义 1.2　设 X 为论域，A 和 B 是 U 上的两个模糊集合，若 $\forall x \in X$, $A(x) \leqslant B(x)$，则称 B 包含 A，或称 A 包含于 B，记作 $A \subseteq B$；若 $\forall x \in U$，$A(x) = B(x)$，则称 A 与 B 相等，记作 $A = B$；若 $A \subseteq B$，但 $A \neq B$，则称 B 真包含 A，或 A 真包含于 B，记为 $A \subset B$。

用 $\varnothing$ 表示其隶属函数恒为 0 的模糊集，U 表示其隶属函数恒为 1 的模糊集，则 Fuzzy(X) 具有下列性质。

定理 1.1　设 $A, B, C \in F(X)$，则

(1) 有界性：$\varnothing \subseteq A \subseteq X$；

(2) 自反性：$A \subseteq A$；

(3) 反对称性：$A \subseteq B$，$B \subseteq A \Rightarrow A = B$；

(4) 传递性：$A \subseteq B$，$B \subseteq C \Rightarrow A \subseteq C$。

由定理 1.1 易知，$\subseteq$ 是 $F(X)$ 上的一种偏序关系，从而 $(F(X), \subseteq)$ 是一个偏序集。

定义 1.3　设 $A, B \in F(X)$，A 与 B 的交、并运算分别定义为 $\cap$、$\cup$，A 的补集记为 A^c，$\forall x \in U$ 有

$$(A \cap B)(x) = \min\{A(x), B(x)\} = A(x) \wedge B(x)$$

$$(A \cup B)(x) = \max\{A(x), B(x)\} = A(x) \vee B(x)$$

$$(A^c)(x) = 1 - A(x)$$

显然，$A \cap B, A \cup B, A^c \in F(X)$。

定理 1.2　设 $A, B, C \in F(X)$，则

(1) 幂等律：$A \cup A = A$，$A \cap A = A$；

(2) 交换律：$A \cup B = B \cup A$，$A \cap B = B \cap A$；

(3) 结合律：$A \cup (B \cup C) = (A \cup B) \cup C$，$A \cap (B \cap C) = (A \cap B) \cap C$；

(4) 分配律：$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$，$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$；

(5)吸收律：$A\cup(A\cap B)=A$，$A\cap(A\cup B)=A$；

(6)复原律：$(A^c)^c=A$；

(7)对偶律：$(A\cup B)^c=A^c\cap B^c$，$(A\cap B)^c=A^c\cup B^c$；

(8)两级律：$A\cup\varnothing=A$，$A\cap\varnothing=\varnothing$，$A\cup U=U$，$A\cap U=A$。

证明过程详见文献[84]，此处证明从略。

3. 模糊集的其他运算

定义 1.4　映射 T：$[0,1]^2\to[0,1]$，如果对$\forall a,b,c\in[0,1]$，满足条件：

(1)交换律：$T(a,b)=T(b,a)$；

(2)结合律：$T(T(a,b),c)=T(a,T(b,c))$；

(3)单调性：若 $a_1\leqslant a_2, b_1\leqslant b_2, T(a_1,b_1)\leqslant T(a_2,b_2)$；

(4)边界条件：$T(1,a)=a$。

则称为 t-三角模，也称为 T 范数。

定义 1.5　映射 S：$[0,1]^2\to[0,1]$，如果对$\forall a,b,c\in[0,1]$，满足条件：

(1)交换律：$S(a,b)=S(b,a)$；

(2)结合律：$S(S(a,b),c)=S(a,S(b,c))$；

(3)单调性：若 $a_1\leqslant a_2, b_1\leqslant b_2, S(a_1,b_1)\leqslant S(a_2,b_2)$；

(4)边界条件：$S(a,0)=a$。

则称为 s-三角模，也称为 S 范数(T 余范)。

T 范数和 S 范数统称为三角算子。

显然，三角模 T 和反三角模 S 都是[0, 1]上的二元运算，通常将它们统称为三角模算子，记为 $\varDelta$。三角模算子的交换性表明其取值不依赖输入变量的顺序，单调性说明在输入变量增加时其值不应该减小；而利用结合性则可以把三角模算子的定义从二元函数扩展为 n 元函数：

$$\varDelta(x_1,x_2,\cdots,x_n)=\varDelta(x_1,\varDelta(x_2,\varDelta(\cdots,\Delta(x_n-1,x_n))))$$

即

$$T(x_1,x_2,\cdots,x_n)=T(x_1,T(x_2,T(\cdots,T(x_n-1,x_n))))$$

$$S(x_1,x_2,\cdots,x_n)=S(x_1,S(x_2,S(\cdots,S(x_n-1,x_n))))$$

定理 1.3　三角模算子 T 和 S 是对偶算子。

性质 1.1　设 T 是 T 范数，则$\forall a,b\in[0,1]$，有

(1)$0\leqslant T(a,b)\leqslant a\wedge b$；

(2)$T(a,0)=0$。

性质 1.2　设 S 是 S 范数，则$\forall a,b\in[0,1]$，有

(1)$a\vee b\leqslant S(a,b)\leqslant 1$；

(2) $S(a, 1)=1$。

定义 1.6　设 $A,B\in F(X)$，对$\forall u\in U$，规定：

$$(A\cup B)(u)=A(u)\vee^* B(u)$$

$$(A\cap B)(u)=A(u)\wedge^* B(u)$$

其中，$\vee^*$、$\wedge^*$是[0, 1]中的二元运算，简称模糊算子。令 $a=A(u), b=B(u)$，常用算子如下。

(1) Zadeh 算子$\vee$、$\wedge$：

$$\begin{cases} a\vee b=\max(a,b) \\ a\wedge b=\min(a,b) \end{cases}$$

(2) 最大乘积算子$\vee$、$\bullet$：

$$\begin{cases} a\vee b=\max(a,b) \\ a\bullet b=ab \end{cases}$$

(3) 代数算子$\dot{+}$、$\bullet$：

$$\begin{cases} a\dot{+}b=a+b-ab \\ a\bullet b=ab \end{cases}$$

(4) 有界算子$\oplus$、$\odot$：

$$\begin{cases} a\oplus b=\min(a+b,1) \\ a\odot b=\max(0,a+b-1) \end{cases}$$

另外还有强烈算子、Einstein 算子、Hamacher 算子等，这里不做一一介绍，读者如有兴趣，可参阅文献[84]以及其他模糊集研究相关文献。

4. 模糊性的度量

模糊集是实现定量刻画模糊性对象的概念，而不同的模糊性对象的模糊程度是不一样的，于是不同的模糊集的模糊程度是有区别的。模糊集的 Fuzzy 性度量可以从几个方面来度量，在这里仅介绍几种常用的 Fuzzy 性度量方法，具体可参看文献[84]。

1) 模糊集的 Fuzzy 度

一个 Fuzzy 集到底模糊到什么程度，又如何来度量这种程度，这都属于 Fuzzy 集的 Fuzzy 性范畴。

定义 1.7　若映射 d：$F(U)\to[0, 1]$满足条件：

(1) 当且仅当 $A\in F(X)$ 时，$d(A)=0$；

(2) 当且仅当 $A(x)\equiv\dfrac{1}{2}$ 时，$d(A)=1$；