粒计算及其不确定信息度量的理论与方法

徐久成　孙　林　张倩倩　著

科学出版社

国家自然科学基金项目资助
河南师范大学学术专著出版基金资助

粒计算及其不确定信息度量的理论与方法

徐久成　孙　林　张倩倩　著

科学出版社
北京

内 容 简 介

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，它涵盖了所有有关粒度的理论、方法和技术，是研究复杂问题求解、海量数据挖掘和不确定性信息处理等问题的有力工具。经过十多年的发展，在与多学科交叉研究的过程中，粒计算正逐步形成其特有的研究体系。本书介绍了粒计算及其不确定信息度量的理论与方法的最新进展，内容涉及不确定信息处理的基本理论，Rough 集理论及其不确定信息度量，边界不确定信息的处理——Rough 集、Fuzzy 集和 Vague 集理论，基于信息粒度与 Rough 集的决策细化的理论分析，基于粒计算的知识约简，基于粒计算的基因表达谱数据挖掘研究，基于粒计算的图像检索，时间序列下的粒度决策演化模型等。

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、高年级本科生和工程技术人员参考。

图书在版编目(CIP)数据

粒计算及其不确定信息度量的理论与方法/徐久成，孙林，张倩倩著. —北京：科学出版社，2013

ISBN 978-7-03-037542-1

Ⅰ. ①粒… Ⅱ. ①徐… ②孙… ③张… Ⅲ. ①人工智能-研究 Ⅳ. ①TP18

中国版本图书馆 CIP 数据核字(2013)第 107919 号

责任编辑：王 哲/责任校对：桂伟利

责任印制：张 倩/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

源海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

*

2013 年 9 月第 一 版 开本：B5(720×1000)

2013 年 9 月第一次印刷 印张：20 3/4

字数：418 300

定价：80.00 元

(如有印装质量问题，我社负责调换)

前　言

粒计算是人工智能领域中崛起的一个新方向，它从实际出发，用尽力而为的满意解替代精确解，其主要思想是在不同的粒度层次上进行问题求解，在很大程度上体现了问题求解过程中的智能。粒计算从提出到现在已有30多年，近年来受到广泛关注。到目前为止，研究人员已经对粒计算理论及其应用作了大量有意义的探索。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。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了不同的粒计算理论模型，主要有模糊集理论模型、粗糙集理论模型、商空间理论模型和云模型等。当前，基于这些模型的不确定性度量方法有很多，这些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。不确定性人工智能的新时代已经到来，认知的不确定性必然导致不确定性人工智能研究的不断深入。研究不确定性信息的表示、处理和模拟，寻找并且形式化地表示不确定性信息中的规律性，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，已经成为人工智能研究领域的一项重要任务。

经过十多年的发展，粒计算研究已经取得了令人鼓舞的成果，其中国内学者对此起到了很大的推动作用，开展了国际、国内学术会议和专题研讨会等多种形式的粒计算学术交流与合作，相继出版了一系列著作，如2007年张钹、张铃的《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)，2008年由13位海内外华人学者合著的《粒计算：过去、现在与展望》(科学出版社)，2010年6月他们又联合出版了《商空间与粒计算——结构化问题求解理论与方法》(科学出版社)，2011年8月结合不确定性与粒计算专题研讨会，出版了《不确定性与粒计算》(科学出版社)，2011年11月结合决策粗糙集理论专题研讨会，出版了《决策粗糙集理论及其研究进展》，2012年8月结合云模型与粒计算专题研讨会，出版了《云模型与粒计算》(科学出版社)等。另外，“中国粗糙集与软计算学术会议”、“国际粒计算高峰论坛”、“IEEE International Conference on Granular Computing”、“Advances in Granular Computing”等国内外有关学术活动的开展，也极大地促进了粒计算理论及其应用的发展。

本书总结了粒计算及其不确定信息度量理论与方法的主要成果，介绍了目前粒计算与不确定信息度量理论与方法的最新研究进展。本书旨在将粒计算与不确定信息度量相结合，运用粒计算的理论与方法度量和处理不确定信息，分析信息粒

化与不确定性的关系，探索粒表示、粒运算、粒推理和粒信息处理中的不确定度量方法，以获取不确定问题的粒化求解理论，进而研究基于粒计算的不确定信息度量的一系列理论与方法；从理论上探讨知识粒的公理化定义，研究知识粒与不确定信息度量方法之间相互融合的表示形式，建立基于粗糙集的粒计算度量和处理不确定信息的理论体系。全书内容共 8 章，分为 3 部分，章节总体按照从理论到应用的思路进行编排。第 1 部分为不确定信息处理的基本理论概要，主要在第 1 章不确定信息处理的基本理论中介绍。第 2 部分为粒计算与不确定信息度量理论与方法的研究，包括第 2 章介绍 Rough 集理论及其不确定信息度量，第 3 章介绍边界不确定信息的处理——Rough 集、Fuzzy 集和 Vague 集理论，第 4 章介绍基于信息粒度与 Rough 集的决策细化的理论分析，第 5 章介绍基于粒计算的知识约简。第 3 部分为粒计算与不确定信息度量的应用研究，包括第 6 章介绍基于粒计算的基因表达谱数据挖掘研究，第 7 章介绍基于粒计算的图像检索，第 8 章介绍时间序列下的粒度决策演化模型。各章之间内容相对独立，自成体系，但都紧密围绕粒计算及其不确定信息度量理论与方法的主题展开。本书的内容引用了作者前期的一些研究成果，同时也包含了作者部分最新研究成果，因此本书既是对前期研究成果的总结，也是对未来研究的展望，可为读者进一步研究粒计算提供参考。

编写本书时，参考了国内外有关研究成果，在此对所涉及的专家和研究人员表示衷心的感谢。书中所列的参考文献可能不够全面，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，硕士研究生徐天贺、胡玉文、李晓艳、高云鹏、任金玉、李双群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。同时感谢国家自然科学基金项目(60873104，61370169)、河南省科技攻关重点项目(112102210194)、河南师范大学学术专著出版基金的资助。

由于作者自身知识水平有限，书中难免存在不妥之处，恳请广大读者批评指正。

徐久成

2013 年 6 月

目　录

第 1 章　不确定信息处理的基本理论

由于客观世界的复杂多样性，不确定性信息在现实生活中大量存在，不确定性主要包括随机性、模糊性、含糊性和知识的不完备性。为了更好地刻画客观世界，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与方法的一门新兴学科。在不确定性的研究中涌现出许多理论和方法，本章重点介绍不确定信息处理的几种基本理论，包括 Rough 集理论、Fuzzy 集理论、Vague 集理论、商空间理论和粒计算理论。

1.1　Rough 集理论

由于计算机科学的发展，特别是网络技术的飞速发展，各个领域的数据和信息量急剧增加，面对如此巨大的信息量，人们往往希望从海量的信息中挖掘出潜在有用的信息。因此，近 20 年来，知识发现和数据挖掘研究日益受到人工智能学界的广泛重视。波兰华沙理工大学 Pawlak 教授于 20 世纪 80 年代初提出的 Rough 集（粗糙集）理论就是一种新的处理知识发现和数据挖掘的方法，该理论能定量分析处理不确定、不精确、不完整的信息与知识[1]。

Rough 集理论以不可分辨关系为基础，建立在分类机制的基础上，它将分类理解为特定空间上的等价关系，而等价关系则构成了对该空间的划分。其主要思想是在保持信息系统分类能力不变的前提下，通过知识约简，删除其中的冗余知识，进一步导出问题的决策或分类规则。

粗糙集理论与其他处理不确定和不精确问题理论的最显著区别是它不必提供处理问题所需数据集合之外的任何先验信息，可以从现有数据集中直接约简、导出决策规则，因而对问题的不确定性描述或处理比较客观，在机器学习与知识获取、数据挖掘、医疗数据分析、专家系统、决策分析、模式识别等方面得到了广泛应用，现已成为一个热门的研究领域[2]。由于该理论在处理不精确或不确定原始数据方面的欠缺性，所以与概率论、模糊数学和证据理论等其他处理不确定或不精确问题的理论有很强的互补性。

1.1.1　信息系统与决策信息系统

人之所以有智能行为是因为有知识，要让机器具有智能行为的能力，就必须让其具有相应的知识。知识表示就是研究用机器表示知识可行的、有效的、通用

的原则和方法。近年来对知识表示的研究引起了广泛的关注。目前，常用的知识表示方法有逻辑模式、框架、语义网络、产生式规则、状态空间等。一种基于信息表的知识表示形式，即信息系统，是粗糙集理论中对知识进行表示和处理的基本工具[3]，粗糙集理论研究的信息系统通常用一个数据表来表示。

定义 1.1 称 $S=<U,A,V,f>$ 是一个信息系统，其中 U 是非空的对象集，即 $U=\{x_1,x_2,\cdots,x_n\}$，称为论域，U 中的每个 $x_i(1\leqslant i\leqslant n)$ 称为一个对象；A 表示属性的非空有限集合；$V=\bigcup\{V_a \mid a\in A\}$，$V_a$ 为属性 a 的值域；$f:U\times A\to V$ 是一个信息函数，它为每个对象的每个属性赋予一个信息值，即 $\forall a\in A, x\in U$，有 $f(x,a)\in V_a$。

信息系统 $S=<U,A,V,f>$ 也称为知识表达系统，有时也简记为 $S=<U,A>$。

定义 1.2 设 U 是对象集，令

$$U^2=U\times U=\{(x_i,x_j)\mid x_i,x_j\in U\}$$

则 $R\subseteq U^2$ 称为 U 上的一个等价关系，若 R 满足以下条件。

(1)自反性：$(x_i,x_i)\in R(1\leqslant i\leqslant n)$。

(2)对称性：$(x_i,x_j)\in R\Rightarrow(x_j,x_i)\in R(\forall i,j\leqslant n)$。

(3)传递性：$(x_i,x_j)\in R,(x_j,x_k)\in R\Rightarrow(x_i,x_k)\in R(\forall i,j,k\leqslant n)$。

设 R 是 U 上的一个等价关系，记

$$[x_i]_R=\{x_j\in U\mid (x_j,x_i)\in R\}$$

则 $[x_i]_R$ 称为包含 x_i 的等价类。

对于每个属性子集 $B\subseteq A$，定义一个二元不可分辨关系(indiscernibility relation)IND(B)，即 $\mathrm{IND}(B)=\{(x,y)\in U\times U\mid \forall a\in B, a(x)=a(y)\}$。

显然，IND(B)是一个等价关系，它构成论域 U 上的一个划分 $U/\mathrm{IND}(B)$，划分 $U/\mathrm{IND}(B)$ 中等价类的个数用 $|U/\mathrm{IND}(B)|$ 表示。在不发生混淆的情况下，也可用 B 代替 IND(B)。

事实上，信息系统可直观地表示为一个二维表的形式，通常称该二维表为信息表，它是表达描述知识的数据表格。信息表的行对应要研究的对象，列对应对象的属性，对象的信息是通过指定对象的各属性值来表示的。容易看出，一个属性对应一个等价关系，一个表可以看做一族等价关系。所以本书以后提到的信息系统或信息表都是指同一个概念，可以混用。

例如，表 1-1 就是一个关于病人的信息系统(信息表)，这里 $S=<U, A>$，其中 $U=\{1, 2, 3, 4, 5, 6\}$ 是病人编号的集合，$A=\{$头疼，肌肉疼，体温$\}$ 是属性的集合。

表 1-1　一个信息系统

对象编号	头疼	肌肉疼	体温
1	是	是	正常
2	是	是	高
3	是	是	很高
4	否	是	正常
5	否	否	高
6	否	是	很高

信息表这种数据表格知识表达系统是对客观对象的描述和罗列，表达的是说明性的知识。当信息表包含的数据足以反映论域的时候，通过属性所对应的等价关系就可以体现论域中的过程知识，即概念之间的逻辑关系或规则知识。事实上，从信息表所表达的说明性知识中发现过程性知识（规则知识）就是知识发现（knowledge discovery）的研究内容。

一个信息系统对应一个关系数据表；反过来，一个关系数据表也对应着一个信息系统。因此，信息系统 $S=<U, A, V, f>$ 是关系数据表的一种抽象描述。

定义 1.3[4]　一个决策信息系统 S 可以表示为 $S=<U, A, V, f>$。其中，U 是对象的集合，也称为论域，$U=\{x_1, x_2, \cdots, x_n\}$，$A=C\cup D$ 是属性集合，属性子集 C 和 D 分别称为条件属性集和决策属性集且 $C\cap D=\varnothing$，$V=\bigcup\{V_a \mid a\in A\}$ 是属性值的集合，V_a 表示属性 a 的值域；$f: U\times A\to V$ 是一个信息函数，它指定 U 中每一个对象 x 的属性值。

决策信息系统又称决策表，它表示当满足某些条件时，决策（行为、操作、控制）应当如何进行，条件属性 C 和决策属性 D 的等价关系 IND(C) 和 IND(D) 的等价类分别称为条件类和决策类。决策信息系统是一类特殊而重要的信息系统，多数决策问题都可以用决策表来表示，其在决策应用中起着重要的作用。

例如，表 1-2 为关于病人诊断的决策信息系统（决策表），这里 $S=<U, A>$，其中 $U=\{1,2,3,4,5,6\}$ 是病人编号的集合，$A=\{$头疼，肌肉疼，体温，流感$\}$ 是属性的集合，$C=\{$头疼，肌肉疼，体温$\}$ 是条件属性集，$D=\{$流感$\}$ 是决策属性集。

表 1-2　流感诊断决策表

对象编号	头疼	肌肉疼	体温	流感
1	是	是	正常	否
2	是	是	高	轻度
3	是	是	很高	重

续表

对象编号	头疼	肌肉疼	体温	流感
4	否	是	正常	否
5	否	否	高	否
6	否	是	很高	较重

1.1.2 近似集及其性质

设 U 是一个非空有限论域，R 是 U 上的二元关系，则 R 称为不可分辨关系，$S=<U,R>$ 称为近似空间，$\forall (x,y)\in U\times U$，若 $(x,y)\in R$，则称元素 x 与 y 在 S 中是不可分辨的。U/R 是 U 上由 R 生成的等价类全体，它构成 U 上的一个划分。U/R 中的集合称为基本集或原子集，任意有限个基本集的并和空集均称为可定义集，否则称为不可定义集。可定义集也称为精确集，不可定义集也称为 Rough 集。

令 $[x]_R=\{y\in U\mid (x,y)\in R\}$，则称 $[x]_R$ 为由 R 决定的 x 的 R 等价类，关系 R 的等价类称为 S 中的基本集(基本概念)或原子。

定义 1.4[3] 给定近似空间 $S=<U,R>$，对于每个子集 $X\subseteq U$ 和不分辨关系 $B(B\subseteq R)$，X 的下近似集和上近似集分别可以由 B 的基本集定义为

$$\underline{B}(X)=\bigcup\{Y_i\mid (Y_i\in U/\mathrm{IND}(B)\wedge Y_i\subseteq X)\} \tag{1-1}$$

$$\overline{B}(X)=\bigcup\{Y_i\mid (Y_i\in U/\mathrm{IND}(B)\wedge Y_i\cap X\neq\varnothing)\} \tag{1-2}$$

式中，$U/\mathrm{IND}(B)$ 是不可分辨关系 B 对 U 的划分。

Rough 集的下近似集(lower approximation)和上近似集(upper approximation)也可通过集合来定义，即

$$\underline{B}(X)=\{x\mid x\in U\wedge [x]_B\subseteq X\} \tag{1-3}$$

$$\overline{B}(X)=\{x\mid x\in U\wedge [x]_B\cap X\neq\varnothing\} \tag{1-4}$$

即当且仅当 $[x]_B\subseteq X$，$x\in\underline{B}(X)$；当且仅当 $[x]_B\cap X\neq\varnothing$，$x\in\overline{B}(X)$。

集合 $\mathrm{POS}_B(X)=\underline{B}(X)$ 称为 X 的 B 正域(positive region)，它可以解释为由那些根据现有知识 B，U 中所有一定属于集合 X 的元素所组成的集合；$\overline{B}(X)$ 可以解释为由那些根据现有知识 B，U 中所有一定能和可能归入集合 X 的元素所组成的集合；集合 $\mathrm{BN}_B(X)=\overline{B}(X)-\underline{B}(X)$ 称为 X 的 B 边界域(boundary region)，它可以解释为由那些根据现有知识 B，U 中所有判断出可能属于 X，但不能完全肯定是否一定属于 X 的元素所组成的集合；$\mathrm{NEG}_B(X)=U-\overline{B}(X)$ 称为 X 的 B 负域(negative region)，它可以解释为由那些根据现有知识 B，判断出 U 中肯定不属于 X 的元素所组成的集合。

当且仅当 $\underline{B}(X)=\overline{B}(X)$，称 X 是 B 可定义的，即对于 B，X 为经典集；当且仅当 $\underline{B}(X)\neq\overline{B}(X)$，称 X 是 B 不可定义的，即对于 B，X 为 Rough 集。

可将 $\underline{B}(X)$ 看做 X 中的最大可定义集，将 $\overline{B}(X)$ 看做含有 X 的最小可定义集。从近似集的定义，可以得到下近似和上近似的下列性质[3,5-9]，证明过程详见文献[9]。

定理 1.1　$\underline{B}(X)$ 和 $\overline{B}(X)$ 有下列性质：

(1) $\underline{B}(X)\subseteq X\subseteq\overline{B}(X)$；

(2) $\underline{B}(\varnothing)=\overline{B}(\varnothing)=\varnothing$；$\underline{B}(U)=\overline{B}(U)=U$；

(3) $\overline{B}(X\cup Y)=\overline{B}(X)\cup\overline{B}(Y)$；

(4) $\underline{B}(X\cap Y)=\underline{B}(X)\cap\underline{B}(Y)$；

(5) $X\subseteq Y\Rightarrow\underline{B}(X)\subseteq\underline{B}(Y)$；

(6) $X\subseteq Y\Rightarrow\overline{B}(X)\subseteq\overline{B}(Y)$；

(7) $\underline{B}(X\cup Y)\supseteq\underline{B}(X)\cup\underline{B}(Y)$；

(8) $\overline{B}(X\cap Y)\subseteq\overline{B}(X)\cap\overline{B}(Y)$；

(9) $\underline{B}(-X)=-\overline{B}(X)$；

(10) $\overline{B}(-X)=-\underline{B}(X)$；

(11) $\underline{B}(\underline{B}(X))=\overline{B}(\underline{B}(X))=\underline{B}(X)$；

(12) $\overline{B}(\overline{B}(X))=\underline{B}(\overline{B}(X))=\overline{B}(X)$。

1.1.3　Rough 集理论中的近似度量方法

在粗糙集理论中，定义了粗糙集意义下的粗糙隶属函数(rough membership function)。通过使用不可分辨关系 B，定义元素 x 对集合 X 的粗糙隶属函数为

$$\mu_X^B(x)=\frac{\mathrm{card}(X\cap[x]_B)}{\mathrm{card}([x]_B)}\tag{1-5}$$

显然，$0\leqslant\mu_X^B(x)\leqslant1$。

定理 1.2[10]　粗糙隶属函数满足以下性质：

(1) $\mu_X^B(x)=1$，当且仅当 $x\in\underline{B}(X)$；

(2) $\mu_X^B(x)=0$，当且仅当 $x\in-\overline{B}(X)$；

(3) $0<\mu_X^B(x)<1$，当且仅当 $x\in\mathrm{BN}_B(X)$；

(4) 如果 $B=\{(x,x)\mid x\in U\}$，则 $\mu_X^B(x)$ 是 X 的特征函数；

(5) 如果 $(x,y)\in B$，则 $\mu_X^B(x)=\mu_X^B(y)$；

(6) $\mu_{U-X}^B(x)=1-\mu_X^B(x)$，$\forall x\in U$；

(7) $\mu_{X\cup Y}^B(x)\geqslant\max(\mu_X^B(x),\mu_Y^B(x))$，$\forall x\in U$；

(8) $\mu_{X\cap Y}^{B}(x)\geqslant\min(\mu_{X}^{B}(x),\mu_{Y}^{B}(x))$，$\forall x\in U$；

(9) 如果 $X=\{X_1, X_2, \cdots, X_n\}$是 U 的一族互不相交的子集，则 $\mu_{\cup X}^{B}(x)=\sum_{X_i\in X}\mu_{X}^{B}(x)$，$\forall x\in U$。

在 Rough 集理论中，Rough 集 X 的不可定义性(不确定性)是由 Rough 集 X 的边界不确定引起的。集合 X 的边界区域越大，其确定性程度就越低。为了更准确地表达集合的精确性，可以用集合 X 的近似精度和 Rough 度这两个概念来描述 Rough 集 X 的不确定性程度。

定义 1.5 假定集合 X 是论域 U 上的一个关于知识 B(属性子集)的 Rough 集，定义其 B 精度(在不发生混淆的情况下，也简称精度)为

$$d_B(X)=\frac{\underline{B}(X)}{\overline{B}(X)} \tag{1-6}$$

其中，$X\neq\varnothing$；如果 $X=\varnothing$，此时可定义 $d_B(X)=1$。

显然，对每个 B 和 $X\subseteq U$ 有 $0\leqslant d_B(X)\leqslant 1$。

当 $d_B(X)=1$ 时，X 的 B 边界区域为空集，集合 X 为 B 可定义的；当 $d_B(X)<1$ 时，集合 X 有非空的 B 边界区域，集合 X 为 B 不可定义的，即粗糙的。集合 X 关于知识 B 的精度也称为近似精度。

在以后章节的讨论中，将等价关系、属性、知识等概念混用，不予分辨。

定义 1.6[3] 假定集合 X 是论域 U 上的一个关于知识 B(属性子集)的 Rough 集，定义其 B 的 Rough 精度(在不发生混淆的情况下，也简称 Rough 度)为

$$\rho_B(X)=1-d_B(X) \tag{1-7}$$

Rough 集 X 的精度是一个区间[0,1]上的实数，它定义了 Rough 集 X 的可定义程度，即集合 X 的确定度。X 的 Rough 度与精度恰恰相反，它定义了集合 X 的知识的不完全程度。

除了用数值(B 精度)来表示 Rough 集的特征外，也可根据上近似和下近似的特征，对 Rough 集 X 的不确定性程度作如下定义：

(1) 如果 $\underline{B}(X)\neq\varnothing$ 且 $\overline{B}(X)\neq U$，则称 X 为 B 粗糙可定义的；

(2) 如果 $\underline{B}(X)=\varnothing$ 且 $\overline{B}(X)\neq U$，则称 X 为 B 内不可定义的；

(3) 如果 $\underline{B}(X)\neq\varnothing$ 且 $\overline{B}(X)=U$，则称 X 为 B 外不可定义的；

(4) 如果 $\underline{B}(X)=\varnothing$ 且 $\overline{B}(X)=U$，则称 X 为 B 全不可定义的。

可以对上述定义作如下的直观理解。

当 $\underline{B}(X)=\overline{B}(X)$时，集合 X 的边界域为空，即根据属性集 B 就可以确定元素是否属于集合 X，即 X 对应的概念是一个确定的概念。对于 Rough 集，边界域的存在导致部分元素不能被确定地判定。如果集合 X 为 B 粗糙可定义，则可以确