YOU SHI GUAN XI CU CAO JI: BU QUE DING XING
JUE CE DE LI LUN YU FANG FA

优势关系粗糙集：不确定性决策的理论与方法

邓维斌　苟光磊　田帅辉◇著

科学出版社

优势关系粗糙集：不确定性决策的理论与方法

邓维斌　苟光磊　田帅辉　著

科学出版社
北京

内 容 简 介

优势关系粗糙集以优势关系代替经典粗糙集的不可分辨关系，更好地满足了描述实际问题中某些属性具有偏序关系和连续属性的需要。优势关系粗糙集既可以有效处理等价关系，又可以处理具有偏序关系的决策信息系统，现已成为处理不确定信息的重要理论模型，受到越来越多学者的关注。本书集结了作者近年来在该领域的研究成果，针对优势关系粗糙集对不精确、不一致、不完备等不确定性信息处理的核心问题，从变精度模型、不一致信息系统的一致性转化、数据驱动的自主式决策、置信优势关系模型及近似集的动态更新、属性约简等方面对优势关系粗糙集的不确定性信息处理理论展开研究，并给出优势关系粗糙集在电信客户价值评价和水质评价中的应用案例。

本书适合计算机科学、管理科学与工程、控制科学与工程、应用数学等专业的科技人员，高等院校高年级本科生、研究生及相关领域的工程技术研究人员阅读参考。

图书在版编目(CIP)数据

优势关系粗糙集：不确定性决策的理论与方法 / 邓维斌 等著. —北京：科学出版社，2017.7

ISBN 978-7-03-053825-3

Ⅰ. ①优… Ⅱ. ①邓… Ⅲ. ①集论-研究 Ⅳ. ①O144

中国版本图书馆 CIP 数据核字（2017）第 146561 号

责任编辑：张 展 孟 锐 / 责任校对：王 翔

责任印制：罗 科 / 封面设计：墨创文化

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

成都锦瑞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

*

2017 年 7 月第 一 版 开本：B5（720×1000）

2017 年 7 月第一次印刷 印张：12.75

字数：256 千字

定价：79.00 元

（如有印装质量问题，我社负责调换）

前　言

在现实生活和实际应用中，不确定性数据普遍存在于金融、军事、经济、商业、工业控制、电信等诸多领域。数据的不确定性往往会使数据挖掘的结果不可靠，甚至出现完全错误的结论。因此，对不确定性数据的研究逐渐受到广泛重视，并已成为智能信息处理的重要研究内容。目前，对不确定性数据进行处理的主要理论包括概率论、模糊集、Vague 集、粗糙集、可拓集、灰度理论、信息论等。其中，波兰科学家帕夫拉克于 1982 年提出了粗糙集(rough set, RS)的基本概念和相关理论，它把那些确定的对象归属到下近似集，将那些无法确认的个体归属到边界区域，边界区域被定义为上近似集与下近似集的差集。上近似集和下近似集均可通过等价关系给出确定的数学公式描述，所以粗糙集是一种用确定的数学方法处理不确定信息的有效工具。

为了处理具有连续属性和优势关系的信息系统，Greco 和 Slowinski 于 20 世纪 90 年代末对经典粗糙集进行扩展，提出了优势关系粗糙集理论。优势关系粗糙集以优势关系代替经典粗糙集的不可分辨关系，更好地满足了描述实际问题中某些属性具有偏序关系和连续属性的需要。由于优势关系粗糙集可以有效处理等价关系和具有偏序关系的决策信息系统，现已成为经典粗糙集一个很重要的扩展理论模型，受到越来越多学者的关注。目前，对优势关系粗糙集进行研究的学者队伍不断壮大，理论和应用研究成果日渐增多。为使更多的学者能够了解优势关系粗糙集理论和方法并参与研究，共同促进该研究领域的发展，现将作者近年来在该领域的研究成果结集成书，希望能对优势关系粗糙集的研究发展做出一定的贡献。

本书由重庆邮电大学邓维斌、田帅辉和重庆理工大学苟光磊共同撰稿完成。其中，邓维斌撰写第 1 章、第 3~5 章和全书统稿；苟光磊撰写第 6~9 章；田帅辉撰写第 2 章、第 10 章及附录部分。全书内容是作者多年来在优势关系粗糙集处理不确定性信息方面的研究成果，在此特别感谢王国胤教授的倾心指导，感谢万晓榆、张清华、胡峰、于洪等教授对作者在研究过程中提供的大量帮助。本书的写作引用了一些宝贵的资料，在此向作者表示深切谢意。

感谢重庆邮电大学出版基金、国家自然科学基金(61309014)、教育部人文社科规划基金(15XJA630003)、重庆市科委基础与前沿技术研究基金

(cstc2017jcyjAX0144)、重庆市教委科学技术研究基金(KJ1500416、KJ1600933)、重庆邮电大学博士启动基金(A2015-20)、重庆理工大学青年星火计划项目(2015XH15)的资助，感谢科学出版社的大力支持。

由于作者水平有限，书中难免存在不妥之处，敬请广大读者批评指正。

目　录

第1章 绪 论

近年来，随着数据获取、数据通信和数据存储技术的迅猛发展和云计算、社交网络、移动互联网等技术的广泛应用，各行各业都建立了数据库和数据仓库来管理人事、财务、市场等数据，如学校、超市、医院、电信企业、电子商务公司和银行等机构都积累了大量的数据。现实世界的多样性、复杂性和运动性，导致人们对事物和信息的表达往往是不精确、不确定的，甚至是模糊的。确定性是指客观事物联系和发展过程中有规律的、必然的、清晰的、精确的属性。不确定性是指客观事物联系和发展过程中无序的、偶然的、模糊的、近似的属性。确定与不确定揭示了客观事物联系和发展过程中的规律性与无序性、必然性与偶然性、清晰与模糊、精确与近似之间的关系。数据的不确定性往往会使决策的结果不可靠，甚至出现完全错误的结论。因此，对不确定性数据的决策方法研究受到广泛的重视，并已成为智能信息处理的重要研究内容。

1.1 不确定性信息概述

19 世纪以来，以牛顿理论为代表的确定性科学，创造了精确描绘世界的方法，将整个宇宙看作是一种确定性的动力学系统，按照确定和有序的规律运动，从牛顿到拉普拉斯，再到爱因斯坦，描绘的都是完全确定的科学世界图景。然而，随着人类社会的不断发展，确定论思想和方法在越来越多的研究领域遇到了无法克服的困难，量子力学的出现进一步揭示了不确定性是自然界的本质属性，不少科学家和哲学家都认为：从根本上说，不确定性将是人类所具有的认识能力的客观状态[1,2]。

在现实生活和实际应用中，客观世界本身所具有的复杂性、不稳定性和人们对其认识的不完备性，以及在数据采集、录入、编辑、处理、分析及表述过程中存在的各种误差(包括系统误差或随机误差)，概念的定性与定量转换等原因，导致随机、模糊、未确知等不确定性数据(uncertain data)普遍存在于金融、军事、经济、商业、工业控制、电信等诸多领域。因此，对不确定性数据的研究受到广泛的重视[3,4]，并已成为智能信息处理的重要研究内容。在相关研究中，可能使用不确定(uncertainty)、不精确(imprecision)、不一致(inconsistency)、不完备

(incompleteness)等词语表达数据的不确定性。下面将对不确定性信息的来源、表现形式进行介绍。

1.1.1 不确定性信息的来源

不确定性信息的来源较多，可能是原始数据本身不准确或是进行了数据粒度之间的变换，也可能是在数据预处理或数据集成过程中产生的。

(1)数据本身具有的不确定性。如 GIS 数据的来源多种多样，有的是从调查和统计数据中得来，有的是从地图和遥感数据中得来。其中，遥感数据本身就具有不确定性的特征，如“混合单元”问题。遥感数据虽经过了各种校正处理(如光谱校正和几何校正)，但仍然保存了“残余”误差，这使得经过解译后得到的各种专题图，也存在一定的属性和几何误差[5]。又如，在进行民意调查时，如要调查网民对某个新闻事件的态度，其答案本身也会具有不确定性。还有，在传感器网络应用与 RFID(radio frequency identification，射频识别)应用等场合，周围环境也会对原始数据的准确度造成影响。

(2)由缺失值造成的不确定性。当由于设备故障、历史原因、调查对象主观因素等原因无法获取某字段的信息而产生缺失值时，会采用两种操作方式，其一是根据一定的方法对数据进行补齐，其二是将具有缺失信息的记录删除，这两种操作方法都会在一定程度上对原始数据的分布特征造成影响。

(3)不同数据粒度变换造成的不确定性。在数据处理过程中，有时需要进行数据粒度的变换，包括从细粒度数据变到粗粒度数据和从粗粒度数据变到细粒度数据两种过程，不同粒度变换的过程会使数据产生不确定性。如假设某经济普查数据库是以县为基本单位记录全国的经济运行情况，而某应用需要以乡、镇为单位对数据进行粒度细化时，往往会以乡、镇所在的县进行数据映射，而同一县中不同乡、镇的经济发展肯定是有差别的，这就导致不确定性数据的产生。在数据粒度粗化过程中也会产生不确定性，如在对百分制成绩进行处理时，将百分制成绩离散化为以优、良、中、差等表示的等级制，就增大了原始成绩的不确定性。

(4)由于隐私等原因造成的不确定性。如在某些图片或视频展示中，出于保护当事人隐私的需要，常对图片或视频进行模糊化处理，增加了数据不确定性。另外，当涉及某些商业秘密时，也会对数据进行模糊化处理。如在银行、保险等行业中，由于市场调研和用户定位等需求，很多公司会委托专业咨询公司或数据分析公司，但是这些公司在提供调研所需的基础数据时，出于对客户隐私的保护，大量基础数据都会被进行人为干扰处理[6]。

(5)数据集成过程中产生的不确定性。在进行数据收集时，往往会从不同数据源中提取数据并进行集成，而在数据集成过程中会导致不确定性的产生。如 Web

页面中含有很多信息，由于页面更新等原因，许多页面的内容并不保持一致，在不同时间所收集到的信息也就出现了不确定性。

1.1.2　不确定性信息的分类

根据不同分类标准，可将不确定性信息进行不同的分类，王印清等将不确定性信息分为以下五类[7]。

(1) 随机信息：客观条件不充分或偶然因素的干扰，使几种确定性结果的出现呈现出偶然性，在某次试验中不能确定哪一个结果会发生。这种试验称为随机试验，由随机试验获得的信息称为随机信息。

(2) 模糊信息：由于事物的复杂性，其元素特征界限不明显，使多个事物的边界不清晰，对其概念不能给出确定性的描述，也不能给出确定的评定标准，这种信息称为模糊信息。

(3) 未确知信息：进行某种决策时，我们所研究和处理的某些因素和信息可能既无随机性又无模糊性，但决策者由于条件的限制而对它们认识不清。也就是说，所掌握的信息不足以确定事物的真实状态和数量关系。这种纯主观上、认识上的不确定性信息称为未确知信息。

(4) 灰色信息：由于事物的复杂性，或信道上各种噪声的干扰以及接收系统能力的限制，人类只能获得事物的部分信息或信息量的大致范围，而不能获得全部信息或确切信息。这种部分已知、部分未知的信息称为灰色信息。

(5) 泛灰信息：泛灰信息是灰色信息的扩张，它除了包括从正面描述的灰色信息之外，还包括了从反面描述的灰色信息。

1.1.3　不确定性信息的表现形式

不确定性信息的主要表现形式分为实例存在不确定性、属性不确定性和语义映射不确定性[6]。

(1) 实例存在不确定性：指一个实例在数据库中是否存在一个不确定值，通常表示为一个概率值。这种不确定性又分为实例之间的相互依赖关系和实例之间的相互独立性两种情况。

(2) 属性取值不确定性：指一个属性的取值存在不确定性，通常用概率密度公式或统计值(如标准差、方差)等来描述属性的不确定性信息。这种不确定性又可以根据属性取值的不同分为数值不确定性和非数值不确定性。

(3) 语义映射不确定性：指数据源和中介源之间的映射关系具有的不确定性。

1.2 不确定性决策理论概述

对于现实生活中的不确定性问题，需要用有效的工具来表达和处理。目前，对不确定性进行处理的主要理论包括概率论、模糊集、Vague 集、粗糙集、可拓集、灰度理论、信息论等。本节对几种常见的不确定性决策理论进行简单介绍。

1. 概率论

随机性又称偶然性，是指事件发生的条件不充分，使条件与结果之间没有决定性的因果关系，在事件是否出现上表现出的不确定性，可以用随机数学进行研究。随机性真正为人类所认识，要归功于苏联数学家柯尔莫哥洛夫。他在测度论基础上，于 1933 年在其“概率论基础”一文中，第一次提出并建立概率论的公理化方法。他的公理化方法成为现代概率论的基础，使概率论成为严谨的数学分支，对概率论的迅速发展起到积极的作用，也使人们可以用数学的方法研究随机性，将“随机性”用“概率”予以量化表示。借助于随机变量的分布函数，人们可以研究随机现象的全部统计特征。

定义 1.1(概率) 设 E 是随机试验，Ω 是它的样本空间。如果对于 E 的每个事件 A，无法有一实数 $p(A)$ 与之对应，且集合函数 $p(.)$ 满足下列条件：

(1) 对于每个事件 A，有 $p(A)\geqslant 0$；

(2) $p(\Omega)=1$；

(3) 若 $A_1, A_2, \cdots, A_n$，是两两不相容事件，即对于 $\forall i \neq j$，$A_i A_j = \Phi(I, j=1,2,\cdots,n)$，则有

$$p(A_1 \cup A_2 \cup \cdots \cup A_k) = p(A_1) + p(A_2) + \cdots + p(A_k) \tag{1.1}$$

则称 $p(A)$ 为事件 A 的概率。

在概率论中有条件概率、全概率、贝叶斯概率等几个重要的公式，分别介绍如下。

定义 1.2(条件概率) 设有 A，B 两个事件，且 $p(A)>0$，称

$$p(B \mid A) = \frac{p(AB)}{p(A)} \tag{1.2}$$

为事件 A 发生的条件下事件 B 发生的条件概率，同时可得 $p(A \mid B) = \dfrac{p(AB)}{p(B)}$。

定理 1.1(全概率公式) 设试验 E 的样本空间为 S，A 为 E 的事件，$B_1, B_2, \ldots, B_n$ 为 S 的一个划分，且 $p(B_i)>0\ (i=1,2,\ldots,n)$，则有

$$p(A)=p(A|B_1)p(B_1)+p(A|B_2)p(B_2)+\cdots+p(A|B_n)p(B_n) \tag{1.3}$$

全概率公式的主要用处在于：它可以将一个复杂事件的概率计算分解为若干个简单事件的概率计算问题，最后应用概率的可加性进行计算。

定理 1.2（贝叶斯公式）　设试验 E 的样本空间为 S，A 为 E 的事件，$B_1,B_2,\ldots,B_n$ 为 S 的一个划分，且 $p(A)>0$，$p(B_i)>0(i=1,2,\ldots,n)$，则有

$$P(B_i|A)=\frac{P(A|B_i)p(B_i)}{\sum_{j=1}^{n}P(A|B_j)p(B_j)},\quad (i=1,2,\cdots,n) \tag{1.4}$$

以贝叶斯公式为基础的贝叶斯理论，一直是处理不确定性决策的重要工具。贝叶斯网用图形模式表示随机变量间的依赖关系，提供一种框架结构来表示因果信息。贝叶斯网可以表达各个节点间的条件独立关系。人们可以直观地从贝叶斯网中得出属性间的条件独立以及依赖关系。另外，贝叶斯网还给出了事件的联合概率分布，根据网络结构以及条件概率表可以得到每个基本事件的概率。贝叶斯理论利用先验知识和样本数据来获得对未知样本的估计，而概率是先验信息和样本数据信息在贝叶斯理论中的表现形式。这样，贝叶斯理论使得不确定知识表示和推理在逻辑上非常清晰并且易于理解[7]。

此外，对于基于概率的不确定性知识表示研究方面，Shortliff 等提出了具有可信度的不确定推理。而后，Dempster 和 Shafer 进一步提出证据理论，引入信任函数和似然函数以描述命题的不确定性。证据理论满足比概率论弱的公理系统，又称为广义概率论。当先验知识很难得到时，证据理论可以区分不确定和不知道的差异，比概率论更具适应性。而当先验知识已知时，证据理论就变成了概率论，所以证据理论是概率论的推广。

2. 信息论

信息论是一门用数理统计方法来研究信息的度量、传递和变换规律的科学。它主要是研究通信和控制系统中普遍存在的信息传递的共同规律，以及研究解决信息的获取、度量、变换、储存和传递等问题的基础理论。信息论将信息的传递作为一种统计现象来考虑，给出了估算通信信道容量的方法。

克劳德·香农(Claude Shannon)被称为是“信息论之父”。人们通常将香农在 1948 年 10 月发表于《贝尔系统技术学报》上的论文“*A Mathematical Theory of Communication*”[8]作为现代信息论研究的开端。香农指出，任何信息都存在冗余，冗余大小与信息中每个符号(数字、字母或单词)出现的概率或不确定性有关。他借鉴了热力学的概念，将信息中排除了冗余后的平均信息称为“信息熵”，并给出了计算信息熵的数据表达式。

以下给出几个关于信息论的基本定义[8, 9]。

要描述一个由离散随机变量构成的离散信源，就是随机变量 X 的取值集合 $X=\{x_1,x_2,\ldots,x_n\}$ 及每个取值的概率测度 $p(x_i)$，则取值与概率的对应关系为

$$[X, p(x_i)] = \begin{pmatrix} x_1, x_2, \ldots, x_n \\ p(x_1), p(x_2), \ldots, p(x_n) \end{pmatrix} \tag{1.5}$$

定义 1.3(自信息量) 设离散的随机变量 X 取值为 x_i 的概率为 $p(x_i)$，则将 x_i 的自信息量 $I(x_i)$ 定义为

$$I(x_i) = -\log p(x_i) \tag{1.6}$$

定义 1.4(信息熵) 将信源的平均信息量定义为信息熵，表示为

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log(p(x_i)) \tag{1.7}$$

定义 1.5(条件熵) 设随机变量 X 的取值集合 $X=\{x_1,x_2,\ldots,x_n\}$，随机变量 Y 的取值集合 $Y=\{y_1,y_2,\ldots,y_m\}$，则条件熵 $H(Y|X)$ 表示在已知随机变量 X 的条件下，随机变量 Y 的不确定性，表示为

$$H(Y \mid X) = -\sum_{i=1}^{n} p(x_i) \sum_{j=1}^{m} p(y_j \mid x_i)\log(p(y_j \mid x_i)) \tag{1.8}$$

定义 1.6(互信息) 设两个随机变量 (X,Y) 的联合分布为 $p(x,y)$，边际分布分别为 $p(x)$、$p(y)$，互信息 $I(X;\ Y)$ 表示为

$$\begin{aligned} I(X;\ Y) = E[I(x_i;\ y_j)] &= \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j) I(x_i;\ y_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j)\log\frac{p(x_i / y_j)}{p(x_i)} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j)\log\frac{p(x_i y_j)}{p(x_i)p(y_j)} \end{aligned} \tag{1.9}$$

互信息是信息论里一种有用的信息度量，它可以看成是一个随机变量中包含的关于另一个随机变量的信息量，或者说是一个随机变量由于已知另一个随机变量而减少的不确定性。

信息论已成为不确定性信息处理的重要理论，现已在数据通信、数据挖掘、图像处理模式识别、信息安全、生物医学工程等众多领域得到非常广泛的应用。

3. 模糊集

19 世纪末，德国著名数学家 Contor(康托尔)创立了经典集合论。在经典集合论中，一个对象要么属于一个集合，要么不属于这个集合，不能模棱两可。即一个对象相对于一个集合来讲，只能在{0,1}中取值。若该对象不属于这个集合取 0，属于这个集合则取 1。因此，一个集合包含的对象是确定的，即集合的外延必须是分明的，这是经典集合论的基础。由此可看出，康托尔集合论只能处理“非此即彼”的现象，不能处理“亦此亦彼”的不确定性问题。

随着人类社会的进步和科学技术的发展，人们逐渐发现有些客观事物之间难以用分明的界限加以区分，也就是说模糊现象普遍存在。所谓模糊现象，是指客

观事物之间难以用分明的界限加以区分的状态，它产生于人们对客观事物的识别和分类之时，并反映在概念之中。外延分明的概念，称为分明概念，它反映分明现象；外延不分明的概念，称为模糊概念，它反映模糊现象。模糊现象是普遍存在的。在人类一般语言以及科学技术语言中，都大量存在着模糊概念，如高与矮、胖与瘦、美与丑、清洁与污染，甚至人与猿、脊椎动物与无脊椎动物、生物与非生物等这样一些对立的概念之间，都没有绝对分明的界限。为了处理概念之间的不分明性，Zadeh(扎德)于 1965 年提出了模糊集理论[10]，它将特征函数的取值从{0,1}扩充到[0,1]。

一般说来，分明概念是扬弃了概念的模糊性而抽象出来的，是把思维绝对化而达到的概念的精确和严格。传统数学以康托尔集合论为基础。集合是描述人脑思维对整体性客观事物的识别和分类的数学方法。康托尔集合要求其分类必须遵从排中律，它只能描述外延分明的“分明概念”，只能表现“非此即彼”，而不能描述和反映外延不分明的“模糊概念”。

模糊概念的外延是不明确的，其边界是不清晰的，因而相应的集合也是“模糊”的。就是说一个对象是否属于这个集合，不能简单地用“是”或“否”来回答。比如，对于“年轻人”这个概念，若要判断 20 岁的张三或 80 岁的李四是否是“年轻人”，答案自然是明确的！但要判断 28~35 岁的人是否属于“年轻人”的集合，就不那么好确定了。对于一个实际年龄不超过 30 岁而又没有几根头发的人，就更难确定是否属于“年轻人”的集合了。

有些现象是精确的，但是适当的模糊化可能使问题得到简化，灵活性大为提高。

例如，在地里摘玉米，若要找一个最大的很难，而且近乎迂腐。我们必须把玉米地里所有的玉米都测量一下，再加以比较才能确定。它的工作量跟玉米地面积成正比。土地面积越大，工作越困难。然而，只要稍稍改变一下问题的提法：不要求找最大的玉米，而是找比较大的，即按通常的说法，到地里摘个大玉米，这时，问题从精确变成了模糊，但同时也从不必要的复杂变成意外的简单，挑几个就可以满足要求。因此，适当的模糊反而灵活。

在许多场合，是与不是、属于与非属于之间的区别不是突变的，不是“一刀切”的，而是有一个边缘地带、量变的过渡过程。于是很自然地会提出疑问：为什么要把自己局限于只考虑“属于”和“不属于”两种极端情况？如果分别用 1 与 0 表示“属于”和“不属于”，称为元素属于集合的隶属度，上述问题就表示成：为什么非要规定隶属度只取 0、1 两个值呢？扎德正是创造性地允许隶属度可取 0 与 1 之间的其他值，从而用隶属函数来表示模糊集合。

定义 1.7(模糊集)　如果一个集合的特征函数 $\mu_A(x)$ 不是{0,1}取值，而是在闭区间[0,1]中取值，则 $\mu_A(x)$ 是表示一个对象 x 隶属于集合 A 的程度的函数，称为隶属度函数。