Robust Rough Sets and Applications

稳健粗糙集及应用

安爽 胡清华 于达仁 编著

清华大学出版社

Robust Rough Sets and Applications

稳健粗糙集及应用

安爽 胡清华 于达仁 编著

清华大学出版社
北京

内 容 简 介

本书系统总结了作者近几年在稳健粗糙集建模及算法设计方面的研究成果。该书针对实际应用中不可避免的噪声问题分别论述了未考虑数据概率分布和充分利用数据概率分布的稳健粗糙集建模方法。其中，基于变精度、软距离和稳健统计量的粗糙集模型是在未考虑数据概率分布信息的前提下研究的稳健模型，概率模糊粗糙集是一种适用于服从不同概率分布的数据集的稳健模型。本书从应用出发，将提出的稳健粗糙集模型用于设计稳健分类与预测模型，提出了模糊粗糙决策树模型、稳健模糊粗糙分类模型、原型选择及稳健分类模型和模糊粗糙回归预测模型。最后，本书将这些预测模型应用于太阳耀斑预报与风电预报，进一步验证稳健粗糙集模型及算法在实践中的稳健性和实用性。

图书在版编目(CIP)数据

稳健粗糙集及应用/安爽，胡清华，于达仁编著. --北京：清华大学出版社，2015
ISBN 978-7-302-42266-2

Ⅰ. ①稳… Ⅱ. ①安… ②胡… ③于… Ⅲ. ①集论－研究 Ⅳ. ①O144

中国版本图书馆 CIP 数据核字(2015)第 283675 号

责任编辑：袁勤勇　薛　阳
封面设计：傅瑞学
责任校对：时翠兰
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市中晟雅豪印务有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：10　　**字　　数**：247 千字
版　　次：2015 年 12 月第 1 版　　**印　　次**：2015 年 12 月第1 次印刷
印　　数：1～1000
定　　价：29.00 元

产品编号：067191-01

前　言

海量的数据中隐藏了丰富的、有价值的知识。然而，数据中不一致、不精确、不完备等不确定性给知识发现带来了巨大挑战。粗糙集理论是波兰学者 Z.Pawlak 于 20 世纪 80 年代提出的一种描述数据的不确定性的数学工具，能够有效地刻画不精确数据中的不一致性。1990 年，Dubois 和 Prade 针对 Pawlak 粗糙集无法处理实值和模糊数据的缺陷，提出了模糊粗糙集模型，扩展了粗糙集理论的应用领域，提升了该理论解决实际问题的能力。粗糙集理论在近十余年里得到了迅速发展，成为不确定性建模和机器学习领域十分活跃的分支。

然而，无论是 Pawlak 粗糙集，还是模糊粗糙集对数据噪声都十分敏感。在实际应用中，采集和存储的数据往往由于某种因素的影响存在不同程度的噪声。噪声的存在使得粗糙集的边界增大，降低了粗糙集理论处理不确定性的能力，严重制约了该理论在实际应用中的效果。粗糙集的稳健性问题成为该理论的研究热点之一。各国学者纷纷采取不同的措施改进粗糙集理论的稳健性能，拓展了经典粗糙集理论中的基本定义，提出了一些稳健的粗糙集模型。

本书根据应用中数据噪声的特点将稳健粗糙集模型划分为两大类：一类是不考虑数据概率分布信息的稳健模型，另一类是考虑数据概率分布信息的稳健模型。其中，基于可变精度的稳健粗糙集模型、基于软距离的稳健粗糙集模型和基于稳健统计量的模糊粗糙集模型是未考虑数据分布信息的稳健模型；概率模糊粗糙集模型是考虑噪声分布信息的稳健模型。本书不仅阐述了这些稳健粗糙集模型的基本性质，还设计了基于稳健粗糙集的分类方法。此外，本书以实际应用验证了稳健粗糙集模型的有效性。

理论始终是为实践服务的。本书的主要特色是从应用出发，将实际应用中遇到的问题抽象成数学模型，进而研究问题的解决方案，有效地将理论模型、学习算法与实际应用结合起来。

本书共分为 9 章。第 1 章综述稳健粗糙集理论的研究现状；第 2 章介绍数据噪声的类型、噪声检测方法以及一些抗噪声模型；第 3 章介绍 Pawlak 粗糙集模型及其拓展模型 —— 优势关系粗糙集、邻域粗糙集和模糊粗糙集；第 4~7 章分别介绍了 4 类稳健粗糙集模型，即基于可变精度的稳健粗糙集模型、基于软距离的稳健粗糙集模型、基于稳健统计量的模糊粗糙集模型、概率模糊粗糙集模型；第 8 章介绍 3 种基于稳健粗糙集的分类模型；第 9 章介绍稳健粗糙集的两种应用。

本书工作能够顺利完成离不开很多专家和朋友的帮助。黄鑫、谢宗霞、于霄、朱鹏飞和车勋建在研究过程中给予了大力帮助，加速了本书的出版进程；马诗咏和张保军为本书的顺利出版提供了莫大的支持。由于作者水平有限，书中难免存在不足，甚至错误之处，恳请读者批评指正。

本书相关研究受到国家自然科学基金（61202259，61222210）、河北省自然科学基金（F2013501052）和国家博士后基金（2013M530874）资助。

编　者

2015 年 11 月

目　录

第 1 章　绪　　论

1.1　稳健粗糙集理论的重要性

粗糙集理论是 Z.Pawlak 于 1982 年提出的 [129]，它提供了一种处理不确定性问题的数学工具。该理论已被成功地应用于机器学习、决策分析、过程控制、模式识别和数据挖掘等领域 [125,137,152,217]。然而，该理论只能处理符号数据的缺陷使它在一些实际应用中受到了限制，这也促进了粗糙集理论的进一步发展。20 世纪 90 年代邻域粗糙集 [113] 和模糊粗糙集 [37] 被提出用来处理数值数据的不确定性。其中，模糊粗糙集理论模拟了人类的推理方式，结合了粗糙逼近和模糊粒化两种互补的不确定性推理方法 [61]。在过去的十几年中，该理论引起了粒计算和不确定性推理等领域的广泛关注 [79,120,166,189,201]，成为机器学习领域十分活跃的一个分支。

研究表明，无论是 Pawlak 粗糙集，还是模糊粗糙集对数据噪声都十分敏感。然而，在实际应用中，采集的数据往往由于某些客观或主观因素的影响被叠加上不同程度的噪声。例如，在实际应用中人们为了节约空间和传输费用常常将这些数据在储存和传输的过程中截断为有限的有效数位。这不可避免地给真实的数值叠加了一定程度的噪声；数据在采集过程中由于测量仪器失灵、或测试人员对仪器不了解、或因思想不集中和粗心大意等导致的错误，使得采集的数值明显偏离真实值，进而给数据带来噪声。噪声干扰的存在使得粗糙集的上、下近似不同程度地偏离其真实值。尤其是孤立点（野点），仅仅一个噪声样本就使得模糊粗糙集的上、下近似产生巨大误差，使该理论在实际应用中处理不确定性的能力下降，甚至失效。

因此，从 20 世纪 90 年代起，粗糙集理论的研究者就针对该理论的稳健性问题展开了广泛的研究。至今，粗糙集理论的稳健性仍然是研究热点之一。

1.2　粗糙集理论的产生与发展

1970 年，波兰学者 Z.Pawlak 和一些逻辑学家在研究信息系统的逻辑特性的基础上提出了粗糙集理论的思想。1982 年，Z.Pawlak 发表了经典论文 Rough sets [129]，该论文的发表标志着粗糙集理论正式诞生。然而，最初的关于粗糙集的研究成果大多是以波兰文发表的，因此当时该理论并未引起国际上的重视。直到 20 世纪 90 年代初，粗糙集理论在知识发现中的成功应用引起了各国学者的高度关注。

1991 年，Z.Pawlak 出版了第一本关于粗糙集的专著*Rough Sets: Theoretical Aspects of Reasoning About Data*。1992 年，R.Slowinski 主编的关于粗糙集应用的论文集*Intelligence decision support: handbook of applications and advances of rough set theory*出版，有力地推动了各国学者对粗糙集理论及其应用的深入而广泛的研究。同年，在波兰 Kiekrz 召开了第 1 届国际粗糙集讨论会，这次会议着重讨论了集合近似定义的基本思想及其应用。1995

年，Z.Pawlak 等人在 ACM Communication 上发表了 Rough sets，极大地扩大了该理论的国际影响。此后，分别在加拿大、日本、美国、波兰、中国等国每年都召开以粗糙集理论为主题的国际研讨会。

中国学者也积极投身于粗糙集理论的研究。2001 年 5 月在重庆召开了第 1 届中国 Rough 集与软计算学术研讨会，邀请了创始人 Z. Pawlak 教授做大会报告。随后每年的研讨会在规模和质量上均呈良好的增长趋势。2002 年 10 月在苏州召开了第 2 届中国 Rough 集与软计算学术研讨会。2003 年 10 月在重庆召开了第 3 届中国 Rough 集与软计算学术研讨会，并同时举办第 9 届粗糙集、模糊集、数据挖掘和软计算的国际会议。2003 年，中国人工智能学会粗糙集与软计算专业委员会成立，粗糙集的研究队伍也随之壮大，研究成果在深度和广度上有了更大的发展。直到 2015 年，每年都举办一次以粗糙集理论为主题的学术研讨会。

近年来，粗糙集理论越来越受到重视，已经成为机器学习和模式识别的重要方法之一，其有效性已在许多科学与工程领域的成功应用中得到证实。

1.3 粗糙集理论的推广

粗糙集理论的提出引起了各国学者的广泛关注，从模型拓展、算法设计、稳健性以及工程应用等方面展开研究工作。

Pawlak 粗糙集理论模拟了人类思维中的粒化和近似两大特点，但是该模型只是人类思维的简单模型。回想一下人类的语言概念系统就会发现，人类思维中的基本概念并不一定是由等价关系生成的互斥的对象子集，而是由属性相似、距离相近或者功能一致等复杂关系形成的交叠的和分层的对象集合。其次，人类语言中的概念往往是模糊的，而不是清晰的。在人类思维和现实世界里存在十分复杂的粒化结构和近似形式，若要将基于等价关系的粗糙集理论应用于复杂问题分析就必须拓展粗糙集理论中的某些基本概念。

经典的粗糙集模型是定义在等价关系的基础上的，这使得该模型无法应用于连续数据集或模糊数据集的不确定处理问题。1988 年，Lin 提出邻域系统的概念用于数据库的近似检索。随后，Liu 和 Huang 提出了邻域粗糙集模型 [113]。该模型将样本集粒化为邻域等价类，再用邻域等价形成的精确集逼近粗糙集的边界。邻域粗糙集模型的提出将集合逼近问题从欧氏空间拓展到邻域近似空间，从而解决了用粗糙集理论处理实值属性数据集的不确定问题。2008 年，文献 [61] 从混合数据粒度计算的角度对邻域粗糙集模型进行了重新定义和解释，并揭示这一模型的研究意义。随后，Yao 对邻域近似算子的性质进行了深入的研究 [195]，Wu 和 Zhang 提出了 K 步邻域的概念 [185]，并分析了邻域近似空间的数学性质，2006 年，Yao 将邻域逼近的概念用于信息检索 [197]。

模糊集理论的创始人 Zadeh 教授认为，在人类推理的过程中绝大多数时候使用的都是模糊信息粒子，而非清晰的粒子。模糊信息粒化是人类具有在不精确、部分已知、部分确定和部分真实的环境下做出合理决策的这一不同寻常的能力的基础，也可以看作是人类的心智和感官在有限的处理细节和存储信息能力时所进行的必要简化。模糊信息粒化在人类推理和模糊逻辑中都占有十分重要的地位。

无论是 Pawlak 粗糙集中的等价信息粒子，还是邻域粗糙集中的邻域信息粒子，都是论域空间中的清晰子集。这些清晰粒子不能反映人类推理中的模糊边界。1990 年，Dubois

和 Prade 将粗糙集与模糊集相结合提出模糊粗糙集和粗糙模糊集的概念 [37]，实现了用模糊集逼近模糊集的推理方式。该理论是通过模拟人类的推理方式而形成的一种处理不确定性推理的数学工具，这一工具将模糊粒化和粗糙逼近两种互补的不确定性推理方法结合起来。模糊粗糙集模型的提出又将 Pawlak 粗糙集模型拓展到模糊近似空间。该模型一经提出又引起了一次研究浪潮。1998 年，Morsi 和 Yakout 引入模糊 T 等价关系、三角范数 T 及其诱导的剩余蕴含算子 θ 拓展了模糊粗糙集的定义 [123]。2004 年，Mi 和 Zhang 等人利用蕴含算子 θ 和其诱导的 σ 算子给出了广义的模糊粗糙集定义 [118]。2005 年，Yeung 和 Chen 等人将这些模糊上、下近似算子进行了归纳 [201]，并给出了公理化描述。2008 年，Li 等人再一次泛化了模糊粗糙集模型 [111]。在新模型中，无须计算对象之间的关系，只需要生成论域的模糊覆盖，这一泛化开拓了模糊粗糙集理论的新应用领域。

此外，Pawlak 粗糙集还被拓展为多种粗糙集模型。为了处理数据中的遗失值，1997 年到 2000 年提出了相似关系粗糙集 [93,94,155]，样本的遗失值被认为可以取任何可能的值，因此该样本被分到所有的相似类中。为了分析有序决策的不一致性，1999 年，Greco 等人提出了优势关系粗糙集 [52]；2000 年，Hu 等人提出了基于模糊偏序关系的粗糙集模型 [63]。研究者还针对不同的问题提出了一些其他的粗糙集模型，例如，混合关系粗糙集模型 [54,156]、基于覆盖的粗糙集模型 [111,212,215]、软模糊粗糙集 [160] 等。

1.4　稳健粗糙集及其研究现状

无论是 Pawlak 粗糙集、邻域粗糙集，还是模糊粗糙集，都存在一个共同的问题，即对数据噪声十分敏感。在实际应用中，采集的数据常常受到噪声干扰的污染，这使得粗糙集理论在实际应用中无法发挥其处理不确定问题的优势。于是，粗糙集理论的稳健性问题又成为国际研究热点问题之一。

为了解决 Pawlak 粗糙集模型对数据噪声敏感的问题，Ziarko 提出了可变精度粗糙集模型的定义 [216]。该定义改变了 Pawlak 粗糙集上下近似的计算方式，它认为一个等价类中只要大部分样本被包含在待逼近的粗糙集中，这个等价类就可以被划入粗糙集的正域；反之，如果这个等价类中大部分样本不被包含在待逼近的粗糙集中，这个等价类就应该被划入粗糙集的负域。随后，可变精度粗糙集模型被进一步发展为 Bayes 粗糙集模型和概率粗糙集模型 [153,154,181,196,218]，这两种模型都是从概率论的视角出发处理不确定性问题。文献 [194] 提出了决策理论粗糙集模型。该模型仍然是代数粗糙集的概率拓展模型。它改变了以决策正域为评价机制的思想，而以多数决策原则产生的总的决策错误率为标准来定义属性的评价标准。2008 年，Yao 等人将该模型用来设计属性约简算法 [198]。2007 年，Cornelis 等人针对 Pawlak 粗糙集对噪声的敏感性提出了模糊量化粗糙集的定义 [25]。该模型是通过模糊隶属度函数将等价类的包含度转化到区间 [0,1] 上来减小噪声和不一致样本带来的影响。2014 年，Zhou 提出了多类决策理论粗糙集 [211]。多类决策理论首先将确定的样本直接分类，然后通过对不确定的样本做进一步的判断来降低分类误差。这一思想与三支决策理论比较相似 [199]。同年，Qian 等人将多粒度思想引入决策理论粗糙集提出了多粒度决策理论粗糙集 [139]，不仅为多粒度粗糙集理论制定了统一的理论框架，同时还提出了一种稳健

的粗糙集模型。2015 年，Liang 和 liu 提出了一种在不确定信息下的基于决策理论粗糙集的风险决策制定方法 [112]。

与 Pawlak 粗糙集相似，邻域粗糙集在分类过程中同样存在对数据噪声敏感的缺陷。为了解决这个问题，一致度的概念被引入邻域粗糙集，提出了邻域一致度的概念 [61]。它是采用多数决策原则划分粗糙集的上下近似，在分类应用中可以有效地抑制数据噪声的干扰，克服了邻域粗糙集的不稳健的缺陷。

模糊粗糙集理论自从被提出便引起了各国学者的广泛关注，该理论的稳健性问题仍然是研究热点之一。2003 年，Fernandez 和 Salido 将变精度的思想引入了模糊粗糙集模型，提出了 β-精度模糊粗糙集模型的定义来减小噪声给上下近似计算带来的影响 [41]；2004 年，Mieszkowicz-Rolka 等人将 β-精度模糊粗糙集模型泛化为变精度模糊粗糙集模型 [120]。在分类问题中，上述两个基于变精度的稳健模型都是通过忽略掉一些边界样本进行抗噪声。2007 年，Cornelis 等人针对模糊粗糙集对噪声的敏感性将模糊量化模糊粗糙集模型推广到连续数据的情况 [25]。2009 年，Zhao 等人提出了模糊变精度粗糙集模型来提升经典的模糊粗糙集模型的稳健性 [208]。2013 年，Verbiest 等人针对模糊粗糙集的稳健性问题提出了最小粒度的概念 [168]，并将其用于模糊粗糙原型选择。此外，Yao 等人于 2014 年提出了一种新颖的基于模糊粒度的变精度 (θ,σ)-模糊粗糙集模型 [200]，并详细讨论了该模型的性质。

1.5 本书组织结构

本书首先介绍 Pawlak 粗糙集、模糊粗糙集、邻域粗糙集、优势关系粗糙集等模型的基本概念，分析这几种粗糙集模型的稳健性。然后，本书按照如图 1.1 所示的结构组织章节。这里将稳健粗糙集根据稳健原理划分为两大部分：一部分是未考虑数据分布情况的稳健模型，即基于可变精度的稳健粗糙集、基于软距离的稳健粗糙集和基于稳健统计量的模糊粗糙集；另一部分是考虑数据分布情况的稳健模型，即概率模糊粗糙集。最后，本书介绍了基于稳健粗糙集的分类模型及应用情况。

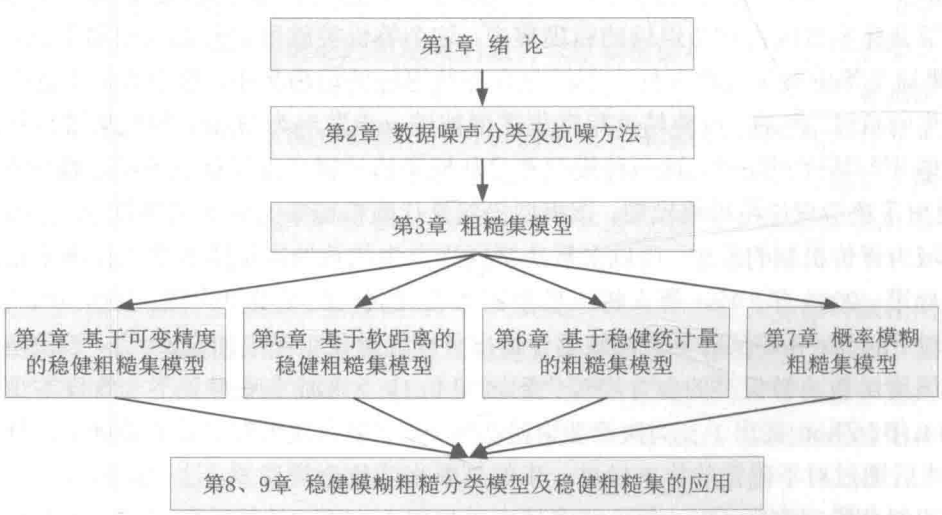

图 1.1 本书的组织结构

本书的具体安排如下。

第 1 章介绍粗糙集理论的起源与发展情况，稳健粗糙集模型的研究现状。第 2 章介绍数据噪声的分类情况和一些现有的抗噪声方法和模型，这里主要针对类噪声进行研究。第 3 章介绍 Pawlak 粗糙集、模糊粗糙集、邻域粗糙集和优势关系粗糙集模型的定义及性质。第 4 章介绍现有的基于可变精度的稳健粗糙集模型。第 5 章介绍基于软距离思想的稳健粗糙集模型。第 6 章介绍基于稳健统计量的粗糙集模型。第 7 章介绍数据分布意识下的稳健模型，即概率模糊粗糙集模型。第 8 章介绍基于稳健粗糙集的分类算法。第 9 章介绍稳健粗糙集在太阳耀斑预报和风速预报中的应用情况。

第 2 章　数据噪声分类及抗噪方法

在社会生产实践中，采集的数据往往受到噪声的污染。数据噪声的存在不仅导致数据质量下降和信息丢失，还会影响机器学习算法的性能和挖掘结果的可信性。为了降低和抑制数据噪声带来的干扰，研究者针对不同数据噪声的特性提出了一系列噪声检测方法和抗噪声方法。

2.1　数据噪声分类

数据中的噪声是指那些不遵守正常行为的模式 [18]，下面分别从数据噪声的特点、分布情况和属性类型三个方面对数据噪声进行分类。

2.1.1　根据数据噪声的特点分类

图 2.1 展示了一个二维数据集中的噪声特点。大部分点分布在 N_1 和 N_2 两个区域，因此它们可以称为正常的点分布区域。点 o_1、o_2 和集合 O_3 中的点远离 N_1 和 N_2，则它们被称为离群点或者数据噪声。

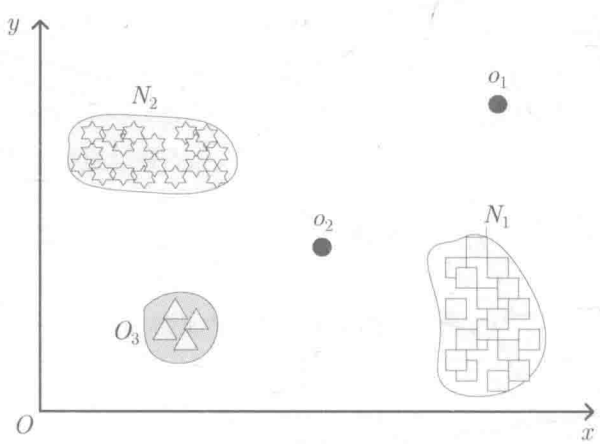

图 2.1　数据噪声

数据的分布不同，噪声的形式也是不一样的。图 2.2 展示了一幅时间序列图像，时刻 t_1 和 t_2 代表的时间是相同的，但是函数值却相差很大。虽然在时刻 t_2 的函数取值在值域范围之内，但是该点却被定义为噪声或者异常。这种噪声被称为条件噪声 [18]。还有这种情况，数据噪声不是单个的点，而是一个集合。这种噪声被称为整体噪声 (或整体异常)[18]。

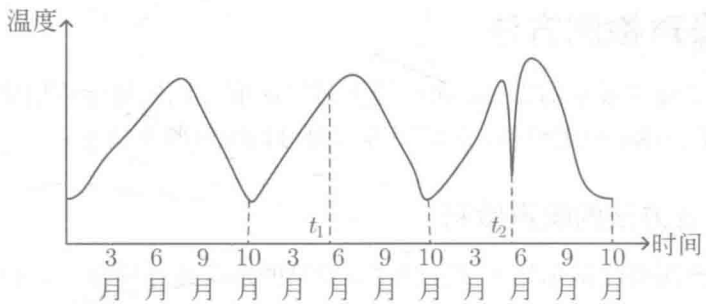

图 2.2　条件噪声

2.1.2　根据数据噪声的分布分类

数据噪声根据它们的分布特性可以分为高斯分布噪声、指数分布噪声、二项分布噪声、$\mathcal{X}$ 方分布噪声等。

2.1.3　根据属性类型分类

数据集中的每个样本都由一组属性描述，根据噪声所处的位置可以将噪声分为输入噪声和输出噪声。对于分类问题而言，输入噪声是指属性噪声，即属性值受到数据噪声的干扰。例如，四舍五入、测量仪器失灵等原因引入的噪声。输出噪声是指样本的类别受到噪声的污染，即类别号被错误标识的样本，简称类噪声。例如，一个样本应该标记为 1，而由于工作人员的失误，这个样本被标记为 2。这里将离群值也视作类噪声。如图 2.3 所示，样本 x 和 y 是被错误标记的样本，称它们为类噪声。同时，样本 x 和 y 也可以分别看作两类样本的离群值，即野点。Zhu 等人于 2004 年详细讨论了属性噪声和类噪声对学习模型的影响 [213]。此外，Frenay 和 Verleysen 于 2014 年再次归纳总结了分类问题中类噪声的存在形式、清洗方法和稳健模型 [43]。

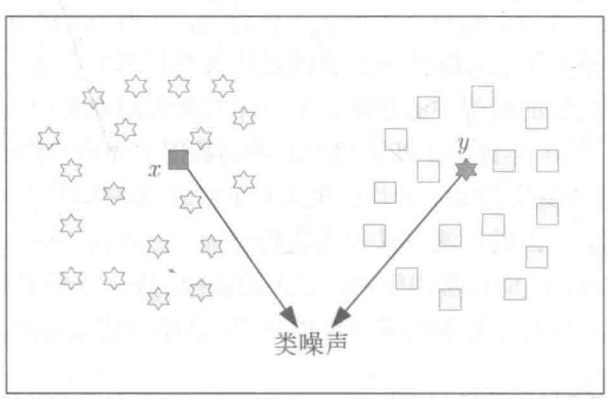

图 2.3　类噪声

2.2 数据噪声检测方法

为了减小数据噪声带来的干扰，在设计学习算法前可以先将噪声检测出来。这里分别从统计方法、聚类方法、分类模型、k 近邻等方面讨论噪声监测技术。

2.2.1 基于统计方法的噪声检测

数据噪声的检测研究最早始于统计领域。最早的基于统计的噪声点检测技术假定被检数据符合某个概率分布模型，不符合该分布模型的数据点被视为噪声点。这类检验方法进一步分为参数方法和非参数方法。参数技术就是假设数据服从某一分布，用数据估计分布模型参数，进而估计样本的概率密度。非参数技术就是不假设数据集的分布，直接估计样本的概率密度。常用的估计概率密度的参数方法有最大似然估计、3σ 准则 [149]、箱线图 [75]、Grubb 测试 [158]、回归模型 [1] 等，非参数技术有直方图 [32] 和 Parzen 窗估计 [128] 等。

基于统计的噪声点检测方法一般分两步进行：在训练阶段，拟合数据的统计模型或构建数据的概要。这个过程可以是有监督的、半监督的或无监督的。有监督的方法估计正常与噪声点的概率密度，半监督的方法仅估计正常或噪声点的概率密度，这取决于类标的可用性。无监督的方法则尝试构建能拟合数据集中绝大多数数据点的统计模型或概要。在检测阶段，检测给定数据点是否符合统计模型或数据概要，根据数据点与模型或概要的偏差，决定数据点的离群度。

此外，Breunig 等人分别于 1999 年 [14] 和 2000 年 [15] 提出用每个样本的相对密度作为判断该样本为噪声的程度的打分标准，密度越大分数越低。当数据集类密度不均匀时，用全局密度作为噪声打分的标准是不合理的，基于局部密度的噪声打分原理具有更大的优势。2004 年，Hautamaki 等人对局部野点因子进行了简化 [58]，用一个样本点作为别的样本的 k 近邻的次数作为噪声程度打分标准。

2.2.2 基于聚类方法的噪声检测

聚类是将相似样本划分到同一个簇里。基于聚类的噪声检测方法遵循的前提假设是正常的样本一定属于某一个簇，噪声样本不属于任何簇。然而，这种方法在识别噪声方面并得不到很好的效果。原因是该方法的潜在目的是给每个样本寻找簇。进而提出了第二种前提假设，即正常样本离它的聚类中心很近，噪声样本离它的聚类中心比较远。基于这个前提假设，Bejerano 在 2001 年提出了基于聚类的序列数据的噪声检测技术 [10]。随后，Smith 等人于 2002 年研究了自组织映射、k 均值聚类和 EM 聚类算法 [157]。然而，当数据中的噪声数据形成一个簇时，上面的检测方法便失去检测能力。从而，基于聚类的噪声检测方式的第三个前提假设被提出，即正常的样本属于大的浓密的簇，小的且稀疏的簇被视为噪声。基于这个假设提出了一些基于聚类的噪声检测算法，详细方法见文献 [40,59,81,127]。

2.2.3 基于分类模型的噪声检测

文献 [159] 提出基于分类模型的噪声检测方法可以归结为两种：一种是多类分类问题 (图 2.4(a))，另一种是一类学习问题 (图 2.4(b))。文献 [92] 中提出一种基于多类分类问题的

噪声检测方法。该方法假设训练样本属于多个正常的类别，并用训练集训练的分类器将测试样本划分到这多个正常类别中，那些没有被划分到任一正常类别中的样本被称为噪声。基于一类学习问题的噪声检测方法是用一类分类问题学习出正常样本分布的边界，将边界外的样本看成是噪声样本。用于噪声检测的分类模型有神经网络模型 [5,77]、贝叶斯网模型 [182,183]、规则学习 [34]、SVM[163] 以及 Fisher 判别式 [143] 等。

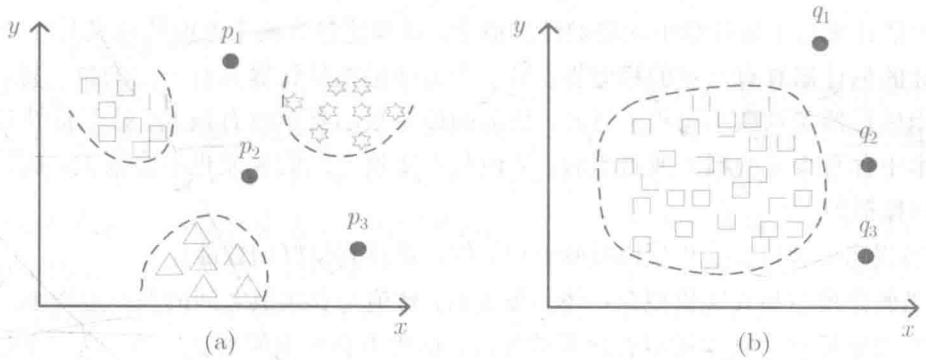

图 2.4　基于分类模型的噪声检测方法

2.2.4　基于 k-近邻的噪声检测

虽然 k-NN 对类噪声十分敏感，但是一些研究者却将这一方法与其他方法融合提出除噪方法。其中，这些方法中的部分方法都是基于统计的噪声检测方法 [36]。Wilson 和 Martinez 总结归纳了基于 k-NN 的训练数据的清洗方法 [179]，并提出一些新的案例剪枝方法来去除噪声。随后，文献 [56,142]提出了新颖的噪声检测方法。该方法将一个样本与它的第 k 个近邻之间的距离作为噪声打分的除噪方法，距离越大分数越高，则这个样本为噪声的可能性就越大。为了提升算法的稳健性，文献 [40] 将该方法扩展为一个样本到最近 k 个样本的距离和作为噪声打分。文献 [89,90]用某个样本的 d 邻域内样本的数量作为打分的标准，数量越少该样本为噪声的可能性就越小。

2.2.5　其他检测方法

信息论认为噪声样本会引起数据样本的复杂度，那些使得数据复杂度增加的最大的特征组成的最小子集被视作噪声。信息度量措施，如信息熵、互信息等都可以用来度量符号数据集的复杂性 [4]。谱分析是将数据集映射到一个低维空间上，在这个空间上正常的样本与噪声样本可以很清晰地分开 [82]。对于一些特殊分布类型的数据，例如环形数据、半月形数据等，Chen 提出了基于核空间深度的野点检测算法 [21]。该方法通过利用中位数的稳健性和引进核函数，可以检测如环形分布等数据的野点。此外，决策树剪枝是为了避免对训练数据中的噪声数据过拟合提出的抑制噪声的方法 [141]。Zhu 等人设计了一系列算法来检测和去除属性噪声和类噪声 [212,213]。

2.3 稳健模型

在应用中，数据预处理过程不能保证所有数据噪声都能够被准确剔除，在设计学习算法时应该考虑算法的稳健性。下面介绍一些稳健的统计模型和数据挖掘模型。

2.3.1 抗差估计

抗差估计来自于统计学中的稳健性的概念。该理论包含两个方面的意义[73]：一个方面是指设计的估计器具有一定的稳健性，另一个方面是指估计器具有一定的抗干扰性。稳健性是指当估计器受到微小噪声干扰时，估值的偏差很小甚至没有偏差。抗干扰性是指当采集的样本中含有少量的粗差或野点时，估值与真实值之间的偏差也不会很大。抗差估计的三个目标概括如下：

① 当假定模型与真实模型比较吻合时，估计器具有良好的性能；

② 当假定模型与真实模型存在较小偏差时，估值与真实值之间的偏差也较小；

③ 当假定模型与真实模型有严重偏差时，估值不会严重偏离真实值。这三个目标可以理解为所有稳健学习模型的目标。

下面给出抗差估计的几何解释 (见图 2.5)。

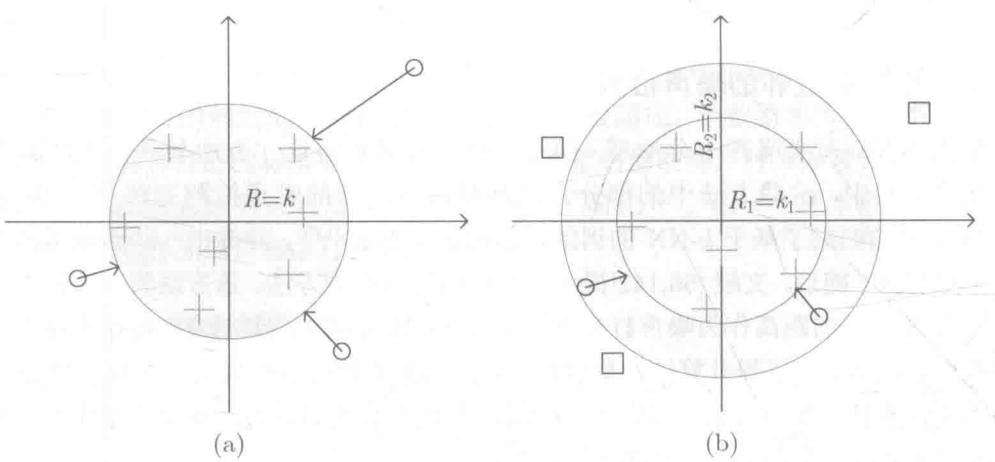

图 2.5 Huber 估计和 IGG 方案

Huber 估计[73] 的抗噪方式属于两段法，在区间 $\pm k\sigma$ 内保持原观测值不变，对 $\pm k\sigma$ 之外的观测值降权处理。图 2.5(a) 是以 R 为半径的球体，+ 表示球内的观测点，o 表示球外的观测点。Huber 估计的几何含义是保持球内观测点的位置不动，把球外的点沿着该点的径向拉到球面上。球越小，抗差能力越强。

IGG 方案[210] 的抗噪方式属于三段法，在区间 $\pm k_1\sigma$ 内保持原观测值不变；在区间 $k_1\sigma < |l| \leqslant k_2\sigma$ 内观测值降权处理；观测值在 $\pm k_2\sigma$ 之外则被淘汰。图 2.5(b) 由半径为 R_1 和 R_2 的两个同心球构成。+ 表示小球内的观测点，o 表示小球外大球内的观测点，□表示大球外的观测点。大球内的观测点与 Huber 处理方法一样，大球外的观测点被淘汰。