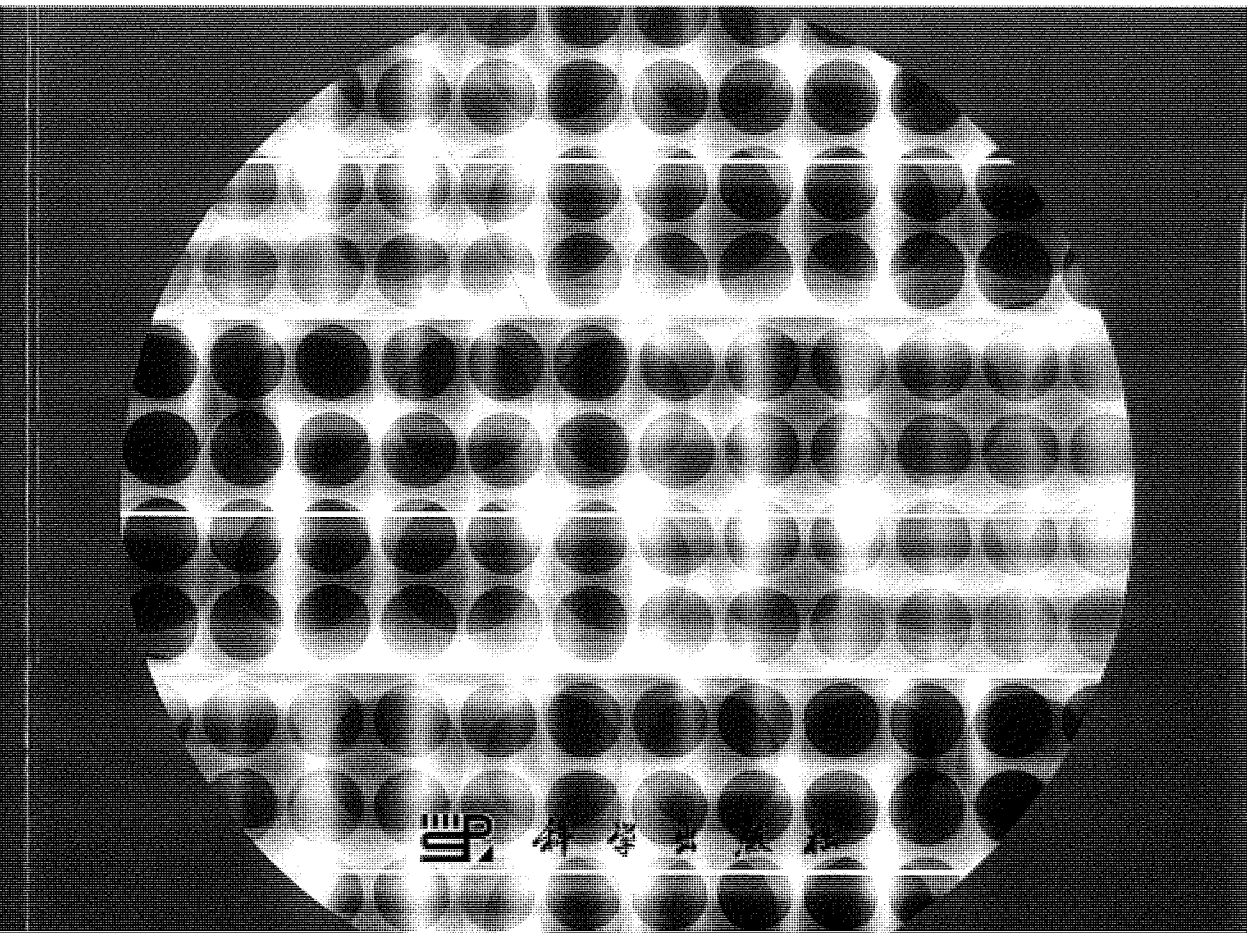
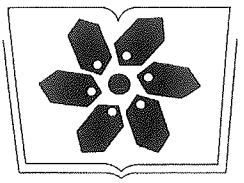
信息科学技术学术著作丛书

中国科学院科学出版基金资助出版

基于不确定性的决策树归纳

王熙照　翟俊海 著





中国科学院科学出版基金资助出版

信息科学技术学术著作丛书

基于不确定性的决策树归纳

王熙照　翟俊海　著

科学出版社

北京

内 容 简 介

本书主要介绍不确定性及不确定环境下的决策树归纳方法，包括模糊决策树归纳、最优割点的模糊化处理、决策树优化、主动学习与特征选择在模糊决策树中的应用、模糊决策树的集成学习等内容。本书结合作者近年来关于决策树归纳学习的研究成果，以决策树归纳学习的基本理论为基础，全面系统地讨论了决策树归纳学习中的主要问题。

本书可作为应用数学、智能科学与技术、自动化等专业高年级本科生和研究生的教材，也可供从事相关研究工作的科研人员参考。

图书在版编目(CIP)数据

基于不确定性的决策树归纳/王熙照，翟俊海著．—北京：科学出版社，2012
(信息科学技术学术著作丛书)
ISBN 978-7-03-034635-3

Ⅰ．基… Ⅱ．①王…②翟… Ⅲ．决策树-归纳 Ⅳ．C934

中国版本图书馆 CIP 数据核字(2012)第 117452 号

责任编辑：魏英杰 雷 旸/ 责任校对：林青梅
责任印制：张 倩 / 封面设计：陈 敬

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

北京虎彩印刷科技有限公司 印刷
科学出版社发行 各地新华书店经销
*
2012 年 6 月第 一 版 开本：787×1092 1/16
2012 年 6 月第一次印刷 印张：21 3/4
字数：424 000
POD定价： 108.00元
(如有印装质量问题，我社负责调换)

《信息科学技术学术著作丛书》序

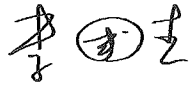
21世纪是信息科学技术发生深刻变革的时代，一场以网络科学、高性能计算和仿真、智能科学、计算思维为特征的信息科学革命正在兴起。信息科学技术正在逐步融入各个应用领域并与生物、纳米、认知等交织在一起，悄然改变着我们的生活方式。信息科学技术已经成为人类社会进步过程中发展最快、交叉渗透性最强、应用面最广的关键技术。

如何进一步推动我国信息科学技术的研究与发展；如何将信息技术发展的新理论、新方法与研究成果转化为社会发展的新动力；如何抓住信息技术深刻发展变革的机遇，提升我国自主创新和可持续发展的能力？这些问题的解答都离不开我国科技工作者和工程技术人员的求索和艰辛付出。为这些科技工作者和工程技术人员提供一个良好的出版环境和平台，将这些科技成就迅速转化为智力成果，将对我国信息科学技术的发展起到重要的推动作用。

《信息科学技术学术著作丛书》是科学出版社在广泛征求专家意见的基础上，经过长期考察、反复论证之后组织出版的。这套丛书旨在传播网络科学和未来网络技术，微电子、光电子和量子信息技术，超级计算机、软件和信息存储技术，数据知识化和基于知识处理的未来信息服务业，低成本信息化和用信息技术提升传统产业，智能与认知科学、生物信息学、社会信息学等前沿交叉科学，信息科学基础理论，信息安全等几个未来信息科学技术重点发展领域的优秀科研成果。丛书力争起点高、内容新、导向性强，具有一定的原创性；体现出科学出版社“高层次、高质量、高水平”的特色和“严肃、严密、严格”的优良作风。

希望这套丛书的出版，能为我国信息科学技术的发展、创新和突破带来一些启迪和帮助。同时，欢迎广大读者提出好的建议，以促进和完善丛书的出版工作。

中国科学院计算技术研究所所长



前　言

机器学习是计算机系统获取智能的本质途径，是人工智能的一个研究领域，主要研究如何让机器具有学习能力。机器学习是一个基于特定目的的知识获取过程，其内在行为是获取知识、发现规律，外在表现是改进性能、适应环境。

归纳学习（或称有监督学习）是机器学习领域的重要分支之一。它的主要任务是从数据中抽取规则，被认为是知识发现的一种重要手段。归纳学习是一种以归纳推理为基础的学习，是一种从多个示例中归纳出一般概念或一般性规律的学习方式，被公认为是专家系统发展的瓶颈。关于归纳学习研究，从近 20 年来出现的浩瀚文献中可以发现大量的归纳学习新方法和新技术，并且成功应用于故障诊断、模式识别、生物信息处理等实用领域。

决策树归纳是归纳学习中一种高效实用的学习方法。决策树方法最早产生于 20 世纪 60 年代中期，Quinlan 提出的 ID3 算法是决策树算法的典型代表，它以信息增益作为选择扩展属性的标准。由于决策树算法结构简单，计算量小，且适用于大规模数据集学习问题，故而已经成为归纳学习的一个重要分支，并在众多的实际领域得到应用。

以 ID3 算法为代表，早期出现的众多构造决策树的算法都是在假定属性取值及分类值是明确的前提下建立的，这些算法都不能处理与人的思维和感觉有关的不确定性。正如 Quinlan 指出的，“决策树的分类结果是分明的，它不能处理分类过程中潜在的不确定性。当属性的取值有微小变化时，有可能导致分类结果明显不合适的突变。生成的决策树一般不具有稳健性，数据信息的不精确或缺少可能完全阻止了样例的分类”。为克服这些缺陷，Quinlan 曾建议采用一种概率方法来构造决策树以处理不确定性。在这个模型下，属性值的不精确性被认为是一种噪声，分支的阈值被软化，最后的分类结果被指定为一种概率估计。这种关于分类问题的不确定性是统计上的，主要源于随机性误差。

有许多种关于不确定性的定义，但总体上可分为两大类，即统计上的和认识上的。统计上的不确定性所处理的是由随机性或系统误差等所产生的现象。与统计上的不确定性不同，认识上的不确定性所表示的现象来源于人的思维、推理、认识和感觉过程，这种认识上的不确定性可进一步细分为不可指定性和模糊性。一般来说，不可指定性代表一种一对多的关系，即从一个具有多个可选择项的问题中选择一个所具有的不可指定性；模糊性主要代表一种边界不分明现象，也就是与不能进行精确区分有关的不确定性。

本书首先介绍什么是不确定性，以及几种常见的不确定性：随机性、模糊性、不可指定性和粗糙性。通过讨论这几种不确定性之间的关系，为后面基于不确定性的决策树归纳学习提供基础。其次介绍不确定环境下决策树归纳过程中不确定性的表示、度量及应用。最后介绍不确定环境下的决策树生成算法、匹配策略、决策树优化算法、特征选取和样本选取此外，本书还介绍了不确定环境下的决策树集成和其他的归纳学习方法。

本书的特点是结合作者近年来关于决策树归纳学习的研究成果，以决策树归纳的基本理论（不确定性、模糊决策树的产生机制）为基础，全面讨论决策树归纳学习中的主要问题（不确定环境下决策树扩展属性启发式标准的设计、决策树优化、特征选择、决策树的集成学习等）。本书大部分内容取材于作者王熙照的博士论文《模糊示例学习研究》(1998)和1998年至今作者及其研究团队在此领域的相关研究成果。参加本书撰写和讨论的有翟俊海、冯慧敏、高相辉、陈爱霞、孟庆武、何玉林和董令彩，最后由王熙照和翟俊海定稿。本书的出版得到了2010年度中国科学院科学出版基金项目、国家自然科学基金项目(61170040)、河北省自然科学基金项目(F2008000635，F2010000323)和河北省应用基础重点研究项目(08963522D)的资助，在此一并表示感谢！另外，感谢科学出版社魏英杰老师的帮助。

由于作者水平所限，书中不足之处在所难免，敬请各位读者指正。

王熙照　翟俊海

2011年7月于河北大学

目　录