不确定性多目标优化的数据挖掘

理论及应用

张志旺
高广霞
邹海林 著

清华大学出版社

不确定性多目标优化的数据挖掘

理论及应用

张志旺
高广霞
邹海林　著

清华大学出版社
北京

内 容 简 介

本书是在作者多年从事数据挖掘行业实践和相关科学研究的基础上编写而成，书中包括数据挖掘理论研究及实际应用的现状分析、研究内容的组织框架、研究方法与技术路线的描述、数据挖掘理论及应用的综述、不确定性理论、多目标优化的分类器方法、模糊多目标优化的分类器模型和算法、基于粗糙集和统计贡献度的特征选择算法、基于粗糙集预处理和粗近似的多目标优化的分类器模型和算法以及基于模糊化、核方法和惩罚因子的多目标优化的分类器模型和算法等内容。本书含有不确定性多目标优化的数据挖掘在信用评分、Web 客户忠诚度分析、蛋白质交互的热点区域预测以及重大疾病的医疗诊断和预测等几个经典领域中的实际应用的描述。最后，通过对研究内容和实际应用效果的总结，展望了进一步研究和应用的方向。

本书可供从事数据挖掘、机器学习与知识工程领域的科学工作者、相关专业的本科生和研究生，以及从事数据分析和处理的工程技术人员参考。

图书在版编目（CIP）数据

不确定性多目标优化的数据挖掘理论及应用/张志旺，高广霞，邹海林著. —北京：清华大学出版社，2015

ISBN 978-7-302-42166-5

Ⅰ. ①不… Ⅱ. ①张… ②高… ③邹… Ⅲ. ①数据采集 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 269035 号

责任编辑：白立军
封面设计：傅瑞学
责任校对：焦丽丽
责任印制：沈　露

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
课件下载：http://www.tup.com.cn，010-62795954
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×230mm　　**印　　张**：13.5　　**字　　数**：252 千字
版　　次：2015 年 11 月第 1 版　　**印　　次**：2015 年 11 月第1次印刷
印　　数：1～500
定　　价：39.00 元

产品编号：065337-01

前　言

数据挖掘是从海量数据中发现并提取隐藏的、前所未知的、有价值的信息或知识，然后利用这些信息或知识辅助做出重要决策的过程。因此，分析数据库、数据集市和数据仓库中收集的历史数据能够帮助商业决策者更好地了解他们的客户、提升产品或服务的质量、评价企业在市场中的地位、提高和改进决策的准确性和增加其竞争力。随着理论研究的不断深入和大量的实际应用的推广，数据挖掘的方法逐渐发展成为一门新的学科和技术，它主要包括最近邻算法、贝叶斯分类器、决策树、神经网络、数学规划、模糊集、粗糙集和支持向量机等方法。

然而，某些现有的确定性数据挖掘方法和模型在解决实际问题时又存在各种各样的局限性，特别是当数据中存在不确定、不精确、不一致、不完整等数据或信息时，这些方法和模型的性能表现较差，有时甚至无法有效地得到所求问题的解。因此，本书在深入研究模糊集、粗糙集和核方法等不确定理论、多目标规划和决策以及分类问题等的基础上，尤其是在客观地分析了这些方法之间存在的互补性的前提下，提出了建立一系列不确定性多目标优化的模型和算法，并将它们主要用于解决数据挖掘中的分类问题，以提高分类的准确性、分类模型和算法的求解效率，以及它们在新数据上的泛化能力。

因此，本书主要探讨了不确定情形的多目标优化分类器的各种模型和算法，包括模糊多目标规划分类器、基于粗糙集预处理的多目标规划分类器和粗近似多目标规划分类器的各种新的模型及算法流程。此外，在全面分析传统的特征选择方法存在的不足的情形下，提出基于统计属性贡献度的特征选择模型和算法；然后，本书分析了粗糙集在属性和信息约简方面的优势，同时，又由于一般的粗糙集方法本身存在缺点，所以在描述的基础上给出了一种扩展的、基于粗糙集理论的属性约简模型和算法，综合考虑了粗糙集的代数和逻辑方法。针对数据中广泛存在的噪声、异常、类别不平衡和非线性可分的情形，提出一类基于核方法、模糊化和惩罚因子的多目标优化的分类模型和算法，并将这类新的分类

模型和算法用于客户信用评分、蛋白质交互的热点区域的预测以及重大疾病的医疗诊断和预测。

总之，通过信用评分、Web客户关系分析、蛋白质交互的热点区域的预测和重大疾病的医疗诊断及预测等的实际应用，将其分类结果与传统方法对比分析的结论表明，不确定性多目标规划分类方法能够显著地提高分类的性能，模型运行的效率，对不同类型数据的适应性、稳定性和灵活性以及它们在新数据集上的推广能力。对于商业应用而言，上述各种模型和算法能够较好地改进和提高商业决策的质量和效益。

由于作者水平有限，书中难免会有错误和不妥之处，恳请广大读者批评指正。

本著作的出版得到国家自然科学基金(No. 61170161)和山东省自然科学基金(No. ZR2012FL13)的资助。感谢清华大学出版社的大力支持。

作　者

2015年6月

目 录

第1章　引言

随着现代信息技术的不断发展，由各类业务系统所产生的数据呈爆炸性增长，人们在面临各种各样复杂的决策问题时又常常缺乏足够的信息和知识来辅助其做出及时准确的决策。数据挖掘技术是20世纪80年代后期兴起的一门交叉学科，已逐渐成为获取信息和知识的重要工具之一，并成为商业智能的一个重要组成部分。尤其在信息泛滥的今天，从历史数据中获取有用的信息和知识显得尤为重要。数据挖掘作为一门新兴的学科，其涉及的理论背景和知识还不是特别深入，研究可以从许多方面入手，这使得数据挖掘成为一个热门的研究课题。本书的宗旨是研究和探索当数据中存在不确定性信息时多目标规划在数据挖掘中的应用，开辟不确定理论、多目标规划和数据挖掘这一交叉研究课题，拓宽数据挖掘理论研究的背景，丰富数据挖掘模型和算法的内容以及提供在实际应用时可供选择的、适当的方法和工具。

本章将从整体上对本书的研究内容做一个概括性的介绍，包括描述本书研究的背景、选题意义，介绍全书的组织结构、研究内容、研究方法和技术路线。

1.1　研究背景与意义

数据挖掘技术是当前机器学习、人工智能、模式识别、计算机科学、智能计算、应用数学、统计学习理论、信息检索以及智能机器人研究中重要的课题。如何从已有的数据中分析、提炼和挖掘隐含的、前所未知的、新颖的并对制定决策有潜在价值的信息和知识，已经越来越成为各行各业迫切需要解决的问题。随着有关此类问题研究的不断深入和进步，关于数据挖掘的模型和算法以及软件工具也层出不穷，不断地推动着数据挖掘向前发展和趋于完善。

就其发展趋势而言，国际上学术界关于数据挖掘这一新兴学科的研究逐渐表现为如

下几个方面(张志旺,2009)。

(1) 数据挖掘技术所处理的数据从结构化、半结构化到非结构化演变。

(2) 数据挖掘技术所使用的理论和方法也从确定性逐渐转向近似性和不确定性,不确定理论主要包括粗糙集、模糊集、证据推理和未确知理论等。

(3) 数据挖掘方法所使用的数学模型的结构特征从线性走向非线性。

(4) 从静态模型转变为动态建模,如针对流数据的分析和处理。

(5) 从单一模型到混合模型和交互式模型发展。

(6) 相关的算法也从线性计算模式向分布式、并行计算和云计算方向扩展。

(7) 从数据的规模来看,随着数据的不断积累,特别是网络和移动网络的应用使得数据从少量的低维数据向高维海量数据发展,真正进入了大数据挖掘和知识发现的时代。

(8) 从数据的来源和产生而言,从单一来源到多源异构的复杂化演变。

(9) 从数据的形态来看,从数值型为主向几何、图形、图像、视频、音频、数字信号、文本、时间和空间等不同模态的混合类型方向发展。

(10) 出现了其他新的发展,并向不同的学科渗透,如生物数据挖掘、化学数据挖掘和商业数据挖掘等。

总之,一个明显的趋势是数据挖掘与其他学科更加交叉,新的方法与模型层出不穷和更加丰富多彩,应用的领域也更加宽广和深入。

实际上,在自然科学、社会科学和工程技术领域中,都不同程度地涉及对不确定因素和对不完备信息的处理。在实际业务系统中采集的数据常常包含噪声、不够精确甚至不完整的信息;如果采用纯数学的假设来消除和回避这些不确定的信息,实际的效果又不理想,反之如果正视它,即对这些信息进行处理往往有助于实际问题的解决。此外,人们的主观知识、经验判断甚至猜测也往往具有很大的不确定性。因此,通过研究基于不确定信息的数据挖掘方法能够有效地解决许多实际问题:如不确定或不精确的信息或知识的表达和应用;经验知识(或专家知识)表示和应用;使用具有不确定、不完整、不一致性的数据进行建模分析和推理;在不损失关键信息的情形下进行属性的约简;提高和增进数据挖掘方法在实际数据集上的稳定性和在新观测数据上获得较好的推广能力等。

目前在国内外学术界,基于模糊的、粗糙的、随机的以及核函数的支持向量机的研究

和应用正处于如火如荼的地步，同时，基于不确定理论和多目标决策的数据挖掘方法刚刚起步，并且将会成为未来的研究热点之一，它具有较强的生命力和广阔的发展空间及应用前景。

此外，传统的确定性的数据挖掘方法和模型在解决实际问题时面临着种种困难和挑战，例如，经常涉及有模型在新数据集上的准确率和推广能力急剧下滑的问题、重复计算的问题、面对噪声和缺失数据时模型无法得到解决的问题、模型的求解效率低下占用了太多的系统资源的问题以及模型的伸缩性较差等。

在数据挖掘和知识发现中，对于分类方法来说，衡量某类模型和算法的优劣主要考察它们的性能和泛化能力两个方面。对于性能而言，模型的运行速度和所解决问题的规模又是两个关键的因素。而泛化能力主要关注模型在新的数据集上所表现出来的准确率和推广能力。但是，数据挖掘模型和算法的最终表现又取决于数据的现状、对数据的了解程度和所使用的数据分析手段。而通常数据中又往往存在各种各样的不确定、不完整、不一致和不精确的数据或信息，所以在不确定的情形下建立新的数据挖掘模型和算法具有重要的理论和现实意义。

因此，本文主要探讨和研究在不确定情形下多目标规划的新的数据挖掘分类模型和算法，并通过实际应用的结果来检验这些方法的有效性和性能特征，同时与传统的方法进行必要的对比分析并得出一定的分析结论。

1.2 本书的主要内容与组织结构

本书研究的对象是不确定情形下基于多目标规划的新的数据挖掘分类模型和算法以及它们在实际领域的应用。研究内容上，首先从分析多目标决策和数学规划的基本理论和已有的相关的数据挖掘分类方法入手，研究并归纳该分类方法的数学原理、方法论、模型和算法。然后从不确定性的角度对多目标规划的新的数据挖掘分类模型和算法进行比较系统的研究。最后给出不确定性的多目标规划分类方法在实际应用中的性能评价和结果分析。研究的重点集中在不确定情形下多目标规划的分类模型和算法上，目标是探索具有较高准确率的分类模型和较高性能的分类算法，进一步完善和丰富数据挖掘的方法

和工具。

在本书的组织结构安排上，首先对研究背景、选题意义、研究内容和方法予以说明；然后描述本研究中涉及的基本理论和方法；接着，通过分别将模糊集和粗糙集与多目标规划方法相结合，探讨了模糊多目标规划分类器、基于粗糙集预处理的多目标规划分类器、粗近似多目标规划分类器和基于模糊化、核方法和惩罚因子的多目标优化分类器的模型与算法；最后，在实际的应用领域中对上述各类模型和算法进行检验及评价。本书的研究内容的组织框架结构图如图 1.1 所示。

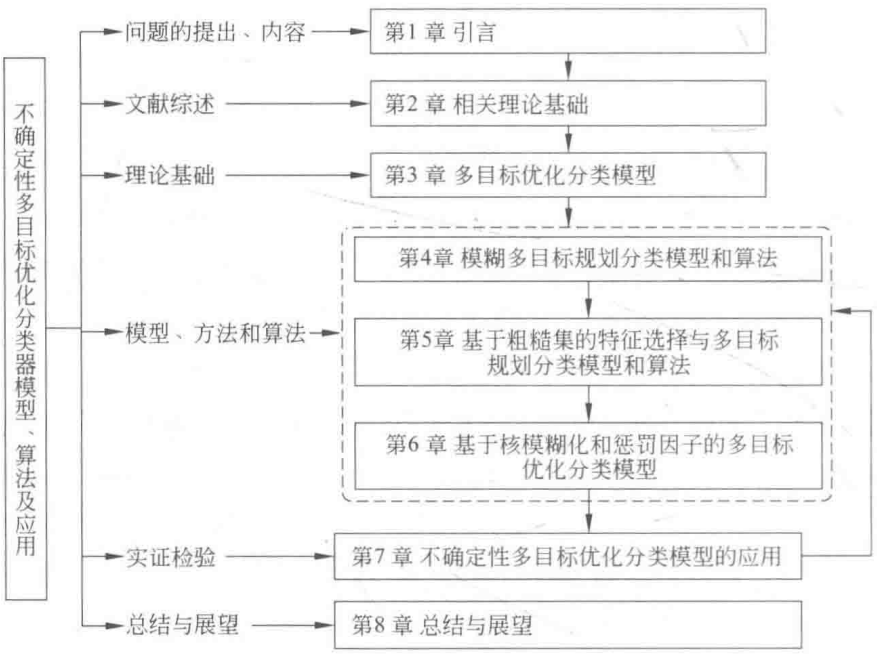

图 1.1　研究内容的组织框架结构图

第 1 章首先介绍有关本书研究的背景、研究的发展趋势，并分析当前数据挖掘研究中存在的一些问题，在此基础上给出该研究的动机和选题意义，然后提供研究的主要内容、结构安排和研究所采用的方法和技术路线。

第 2 章描述了国内外相关理论和方法研究的发展过程和现状，主要包括数据挖掘、最优化理论、分类、基于支持向量机和数学规划的最优化分类方法、基于模糊集和粗糙集的

不确定理论及相关数学规划的分类方法，并对这些文献做了简要的分析和评价。

第 3 章分析并总结了分类方法的数学表示、分类性能评价方法、数据挖掘中的支持向量机和多目标线性规划及多目标二次规划分类的基本原理、方法或模型，它们将构成本书研究内容的理论基础。

第 4 章研究并建立各类模糊多目标规划分类器方法或模型，主要包括模糊多目标线性规划和模糊多目标二次规划的分类模型，此外，根据模糊策略在最优化问题中出现的位置，分别讨论了对称的模糊多目标规划分类方法、非对称的模糊多目标规划分类方法以及多阶段的模糊多目标规划分类方法，并将这些方法同传统的多阶段的模糊线性规划分类方法进行对比分析，并给出分析结论。

第 5 章首先研究了基于统计属性贡献度和基于粗糙集的一般信息约简方法，然后，在分析粗糙集的一般信息约简方法存在的局限性以后给出基于粗糙集代数和逻辑方法的约简模型和算法；随后，根据粗糙集和多目标规划分类方法之间存在的各种互补性，提出建立基于粗糙集预处理的多目标规划的分类模型和算法。当决策域中本身存在不确定、不完整和不精确的信息时，根据粗糙集利用近似方法求解分类问题的一般方法，并提出建立粗近似多目标线性规划和粗近似多目标二次规划的分类模型和算法流程。

在实际应用中，由于数据中普遍存在噪声、异常、类别不平衡和非线性可分等问题，传统的分类器方法、支持向量和多目标优化分类器的性能会迅速退化，难以满足实际应用的需要。因此，在第 6 章研究了模糊化、核方法和类别不平衡的学习技术，并在此基础上提出并建立一系列基于模糊化、核方法和惩罚因子的多目标优化分类器模型和算法。第 4 章～第 6 章将构成本书研究不确定性多目标规划的分类方法的主要内容和主体部分。

在第 7 章，通过信用评分、Web 客户忠诚度分析、蛋白质交互的热点区域预测和重大疾病的医疗诊断及预测的实际应用案例检证了上述各类分类模型和算法的性能，并给出分类结果和性能评价的有关结论。

第 8 章就上述各类方法的理论分析和实际应用的表现，得出最终的结论，并根据现有方法和模型存在的可以改进的地方提出今后进一步研究的方向和建议。

1.3 研究方法与技术路线

本书采用的研究方法主要是理论结合实际应用的方法，在深入研究不确定理论、多目标决策理论和数据挖掘的分类方法等的基础上，探讨和发现在不确定情形下将它们相互结合的新的数据挖掘分类模型和算法，然后将其应用到实际的数据集中进行模型和算法的性能检验、结果评估和获得改进思路，并与在确定信息的情形下的各种算法进行比较分析，最后得出该研究的结论以及它们具有的优势与存在的不足。

本书中各个研究内容的技术路线如图 1.2 所示。

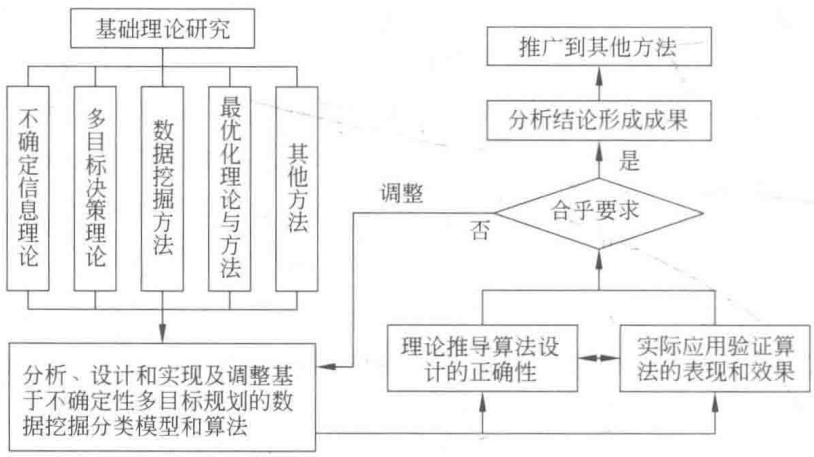

图 1.2 本研究的技术路线

本研究采用的技术路线主要可以分为 3 个阶段和层次：首先，从不确定理论、最优化理论以及它们在数据挖掘中的应用入手，并主要针对最优化理论中多目标规划的数据挖掘方法作为研究的理论基础；然后，探讨各类基于不确定理论和多目标规划的分类模型和算法；最后，通过把这些方法应用到实际的数据挖掘模型中去，并分析其实际效果，得出一定的结论，根据得到的分析结论和与传统方法的对比分析的结果对模型和算法做出进一步的调整和改进，直到得到满意的模型和算法。

第 2 章　相关理论基础

不确定性多目标优化的分类方法以不确定理论、多目标决策理论和方法以及数据挖掘分类方法为理论基础。因此，本章将从数据挖掘、最优化理论、分类方法、最优化分类方法和不确定理论及其数学规划等方面入手，并对它们的基本概念、方法论、实际应用以及优势和不足进行总结、分析和评价，对这些方面的研究的发展过程做一个整体的描述和梳理，以便为以后的研究内容提供基础和帮助。

2.1　数 据 挖 掘

数据挖掘(Data Mining)技术是 20 世纪 80 年代后出现的一门新兴的交叉学科，它是在数据库、统计学、人工智能、模式识别、机器学习、信息检索、算法理论、图像与信号处理、多媒体技术、空间数据处理技术等学科的基础上发展起来的。随着信息化时代的来临，数据将充斥人们生活中的每一个角落，并且各行各业存储的数据量将呈爆炸式增长，出现了“数据丰富，但知识贫乏”的现象，因此，数据挖掘逐渐成为学术界一个研究的热点之一，同时，数据挖掘已经被许多组织用来从海量数据中提取自己需要的信息或知识，并用这些有价值的信息做出关键的业务决策。通过分析存储在数据仓库或数据集市中的历史数据可以更好地洞察客户、提升产品或客户服务的质量、了解企业在行业中所处的位置，以及改进决策的准确性和科学性并有效地提升企业的竞争力。

从数据挖掘过程构成的角度来看，一般地，数据挖掘主要包括 5 个步骤。

(1) 数据挖掘任务的描述，即根据用户的专业知识和应用领域的数据特征来描述和定义数据挖掘的目标。

(2) 数据的选择和集中，即选取并集成所有与数据挖掘任务有关的数据集。

(3) 数据预处理，包括数据的完整性和一致性检验、噪声去除、修正错误数据、填充缺