首都经济贸易大学出版基金资助
国家自然科学基金天元基金项目

决策分析与决策树算法优化

高　静◎著

JUECE FENXI YU
JUECESHU SUANFA YOUHUA

首都经济贸易大学出版社
Capital University of Economics and Business Press

首都经济贸易大学出版基金资助
国家自然科学基金天元基金项目
（项目批准号:11426158）

决策分析与决策树算法优化

JUECE FENXI YU JUECESHU SUANFA YOUHUA

高　静◎著

首都经济贸易大学出版社
Capital University of Economics and Business Press
·北　京·

图书在版编目(CIP)数据

决策分析与决策树算法优化/高静著. —北京:首都经济贸易大学出版社,2017.4

ISBN 978-7-5638-2630-8

Ⅰ.①决… Ⅱ.①高… Ⅲ.①决策学②决策树—最优化算法 Ⅳ.①C934

中国版本图书馆 CIP 数据核字(2017)第 053527 号

决策分析与决策树算法优化

高 静 著

责任编辑	王 猛
封面设计	砚祥志远·激光照排 TEL: 010-65976003
出版发行	首都经济贸易大学出版社
地 址	北京市朝阳区红庙(邮编 100026)
电 话	(010)65976483 65065761 65071505(传真)
网 址	http://www.sjmcb.com
E-mail	publish@cueb.edu.cn
经 销	全国新华书店
照 排	北京砚祥志远激光照排技术有限公司
印 刷	人民日报印刷厂
开 本	710 毫米×1000 毫米 1/16
字 数	176 千字
印 张	10
版 次	2017 年 4 月第 1 版 2017 年 4 月第 1 次印刷
书 号	ISBN 978-7-5638-2630-8/C·131
定 价	29.00 元

前　言

本书针对决策技术中的决策树算法进行了深入分析和研究，与其他技术结合，提出了大量融合算法；创新性地借鉴了认知物理学的研究思想，借鉴认知物理的信息扩散理论讨论了参数波动变化时规则的取舍；借鉴这样的理论思想对传统的 ID3 算法进行了改进，在认知物理原有的信息熵的概念上提出了信息补偿，并在这种新的信息启发下，提出了基于信息补偿量的决策树生成算法 CID3 算法，有效地解决了 ID3 取值偏向多值属性的问题。

本书较完备地分析和整理了决策树与粗糙集的理论及其方法。由于 ID3 算法不能较好地处理带有不一致信息的数据集，这里选择了基于信息熵的属性约简进行数据预处理。由于经典的基于信息熵的属性约简算法的时间复杂度不太理想，而结合差别矩阵的方法通俗易懂，所以本书提出了一个新的基于信息熵的属性约简的差别矩阵算法。该算法的时间复杂度较以前算法的时间复杂度要小，用新算法预处理数据集，可以预先去除一些不重要的属性，从而可以生成简单易懂的决策树，提高决策树的泛化能力和预测能力。

本书在对数据进行预处理后，有效结合了决策树和粗糙集理论各自的优点，提出了基于粗糙边界的决策树优化算法。在该算法中，引入抑制因子，对即将扩张的结点，在常用的终止条件的基础上加入一个新的终止条件，这样不用通过剪枝就能生成一棵较合理的决策树，从而避免了树的过分细化而生成过于庞大的决策树，便于用户的理解，提高了决策树的泛化能力和对未来数据的分类、预测能力。

本书针对多种基于粗糙集理论的决策树生成算法，证明了基于粗糙边界和基于正区域以及基于依赖度的决策树的构造算法是等价的。在此基础上，提出了伴随正区域的概念，进一步改进了基于正区域的决策树算法，提出新的具有较好预测效果的决策树算法，它较好地避免了 ID3 算法生成决策树后再剪枝的额外开销，从而提高了决策树生成算法的效率。

同时，由于最优决策树的确定已被证明是 NP－完全问题，为避免领域知识

参与所造成的巨大代价，提高特征构建的自动化程度，更好地适用于海量数据的挖掘，本书提出了“近似信息增益”的新的评价标准，在 C4.5 算法的基础上，提出了 AR－C4.5 算法，利用新生成的属性和数据本身固有的属性，共同构造决策树。

本书使用具体领域的数据，用实例验证了上述的各种优化算法，并将决策树与粗糙集理论以及其他理论的研究进一步有机结合，针对需要解决的具体问题探索出更优化的算法，挖掘出更深层次的规律，这不仅是对分类技术理论应用范围的进一步扩展，也是对整个知识发现研究方法的有益探索。

本书是作者多年数据挖掘及其相关技术教学以及研究之后进行的关于决策分析与决策树创建及其优化的全面总结，对该领域的研究提供了一种可借鉴的方法论。

本书可以作为计算机及其相关专业的本、专科生的数据挖掘的辅助教材，也可以作为计算机科技人员的参考书。

全书由首都经济贸易大学信息学院高静编著。首都经济贸易大学杨一平教授、马慧教授、徐天晟教授及各位领导和同事也对本书的编写给予了热情的帮助和指导。本书的出版得到了首都经济贸易大学出版基金的资助。首都经济贸易大学出版社杨玲副社长及各位编辑也为本书的出版付出了辛勤的劳动和汗水。本书的编写过程也离不开父母、爱人和女儿的支持。

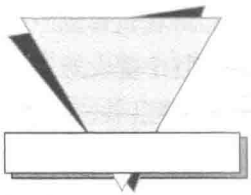
在此，谨向所有给予我们支持和帮助的各位同仁和家人表示衷心的感谢。

作　者

2016 年 9 月于北京

目 录

图表目录



引言

随着计算机、网络和通信等技术的高速发展，信息处理迅速产业化，在技术上表现为整个社会对大规模数据操作的产业化。这使得人们所积累的数据越来越多，并且数据与信息系统中的不确定性更加显著。海量杂乱的数据背后隐藏着许多重要的信息，人们希望能够对其进行深入分析，以便更好地利用这些数据所隐藏的信息。目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势[1]。正是由于缺乏挖掘数据背后隐藏的信息的手段，导致了“数据爆炸但知识贫乏”的现象。

传统的数据分析做法费时费力，效率较低，且只能获得这些数据的表层信息，不能获得数据属性的内在关系和隐含信息，即不能有效地获取人们感兴趣的知识[2,3]。所以，一种能自动分析数据，并提取出隐藏的为人所理解的知识的数据挖掘(Data Mining，DM)[4,5]算法应运而生。它的出现为实现自动和智能地把海量数据转化为有用的知识提供了有力的手段[6]。人们把原始数据看作形成知识的源泉，就像从矿石中采矿一样。原始数据可以是结构化的，如关系数据库中的数据；也可以是半结构化的，如文本、图形、图像数据；甚至是非结构化的异构数据，如分布在网络上的 Web 数据。

数据挖掘技术从一开始就是面向应用的，它要对这些数据进行微观、中观乃至宏观的统计、分析、综合和推理，以指导实际问题的求解，发现事件间的相互关联，甚至利用已有的数据对未来的活动进行预测。如数据挖掘在零售业中的应用，能够识别顾客的购买行为，发现顾客的购买模式和趋势，从而改进服务质量，取得更好的顾客保持力和满意程度，提高货品销量，减少商业成本。数据挖掘在电信业中的应用有助于理解商业行为，确定电信模式，捕捉盗用行为，更好地利用资源和提高服务质量。此外，数据挖掘在金融系统和生物医学等方面的研究与应用也获得了较大的成功，并促进了这些行业的发展。需要指出的是，数据挖掘所发现的知识都是相对的，具有特定的前提和约束条件，同时，为了易于理解，最好能用自然语言来表达发现的结果。

近年来，分类技术也已被有效地应用于科学实验、医疗诊断、气象预报、信贷审核、商业预测、案件侦破等领域，取得了良好的效果。依据分类所采用的不同模型，主要可分为：基于决策树模型的数据分类、基于神经网络模型的数据分类、基于统计模型的数据分类、基于遗传算法模型的数据分类、基于粗糙集模型的数据分类[7]。

上述分类中，基于决策树的分类模型以其特有的优点广为人们采用。首先，决策树方法结构简单，便于人们理解；其次，决策树模型效率高，对训练集数据量较大的情况较为适合；第三，决策树方法通常不需要受训数据外的知识；第四，决策树方法具有较高的分类精确度。一些学者曾利用决策树算法解决中文分词中的未登录词识别问题，取得了良好的识别效果[8]。

虽然，目前常用的 ID3 算法及其衍生算法 C4.5 被人们认为是标准决策树学习算法中最优秀的算法，并且大量的改进算法也是基于它们的核心策略的，但是 ID3 算法以及 C4.5 算法仍有很多不足之处。例如，ID3 算法中用到的属性选择标准——信息增益，具有倾向于取值较多的属性的缺陷；ID3 算法生成的决策树由于分支过细，导致其泛化能力较差，从而导致决策树的预测精度下降，为了提高决策树的泛化能力，提高预测精度，往往要对生成的决策树进行剪枝。但是，这会使生成决策树的代价过大，特别是对于大型数据是非常不利的，对于一些实时性要求较高的预测领域更加不利。

与决策树算法结合的方法有很多，粗糙集方法是其中之一。粗糙集理论是 20 世纪 80 年代初由波兰数学家 Pawlak Z. 教授提出的，用于研究不完整数据和不精确知识的表达、学习归纳的数学分析理论[9]。其特点是算法简单，无需提供数据之外的任何先验信息，可直接从给定问题的描述集合出发，通过不可分辨关系和等价类确定给定问题的近似域，从而找出该问题的规律。属性约简是粗糙集理论中一个重要的研究课题[10]。

由于大型数据库中常常包含许多对发现规则来讲是冗余的、不必要的属性，研究人员发现，如果能将冗余属性剔除，将大大提高系统潜在知识的清晰度，降低知识发现的时间复杂性，提高算法效率。

随着数据挖掘的兴起，加之粗糙集理论的特点，它越来越受到众多研究人员的重视。粗糙集是一个强大的数据分析工具，它能表达和处理不完备信息；能在保留关键信息的前提下对数据进行化简并求得知识的最小表达式；能识别并评估数据之间的依赖关系，揭示出概念的简单模式；能从经验数据中获取易于证实的规则知识，特别适用于智能控制。但是，由于粗糙集理论的分类通常是确定的，并且缺乏交互验证功能，所以结果往往不稳定，精度不高。决策树是一种决策集的树状结构，决策树方法具有速度快，容易转换成简单且便于理解的分类规则，容易转换成数据库查询语言等优点。然而，当数据集中的属性过多时，用决策树分类易出现结构性差，难以发现一些本来可以找到的有用的规

则信息等情况[11]。

由于粗糙集和决策树具有很强的优势互补性，因此，如果将两种方法有机结合，即采用粗糙集进行数据约简，去除冗余属性，然后利用决策树方法来产生分类所用到的规则，有可能形成新的有效分类方法。基于粗糙集和决策树结合的数据挖掘算法可以更好地服务于数据挖掘的领域，能提高对大型数据库中的不完整数据进行分析和学习的能力，因而具有广泛的应用前景和实用价值。

1）本书写作的目的与意义

（1）具有开创性。本书借鉴认知物理学的研究思想，改进了传统的 ID3 算法，提出了基于信息补偿量的决策树算法，有效地解决了 ID3 取值偏向多值属性的问题；结合决策树和粗糙集的理论，提出了基于粗糙边界的决策树优化算法，证明了基于粗糙边界和基于正区域以及基于依赖度的决策树的构造算法是等价的，并进一步改进了基于正区域的决策树算法，提出新的具有较好预测效果的决策树算法，避免了 ID3 算法生成决策树后再剪枝的额外开销，从而提高了决策树生成算法的效率。同时，由于最优决策树的确定已被证明是 NP－完全问题，为避免领域知识参与所造成的巨大代价，提高特征构建的自动化程度，更好地适用于海量数据的挖掘，本书提出了“近似信息增益”的新评价标准，在 C4.5 算法的基础上，提出了 AR－C4.5 算法，利用新生成的属性和数据本身固有的属性，共同构造决策树。

（2）具有重要的理论意义。本书在决策树算法和粗糙集理论以及预测模型日益不断发展的情况下，借鉴现有的决策树算法以及数据挖掘领域的其他成果，深入地研究决策树的算法，进行分析对比，从而提出高效的决策树构造算法，从多种领域的海量数据中提取特点、寻找规律，使用新的分类算法，验证其科学性。

（3）具有一定的应用价值。现代企业数据分析与决策运用的好坏与否直接决定了企业的成功与失败。市场竞争中，企业面对海量的信息，若无法做出有效决策，将使得其市场风险大大增加。本书研究了对各种敏感信息进行分类决策的技术，利用信息挖掘技术挖掘相关资源的最新动向的数据。它的实现将为企业的智能决策提供重要的依据，使企业能够制定有效的策略，在市场竞争上取得优势。

本书运用现有的知识发现领域的技术和成果，结合粗糙集理论的优点，提出了预测效果更好的决策树算法，为企业预测提供了一个新的较好的技术

手段。

2)本书的主要创新点

(1)创新性地将认知物理学的思想引入到决策树算法的构建中,改进了传统的决策树算法——ID3 算法,提出了基于信息补偿量的决策树算法,有效解决了 ID3 取值偏向多值属性的问题。

(2)针对粗糙集理论的基本知识,较完备地分析、整理和应用了粗糙集与决策树的优势互补的理论和方法,将两种方法有机结合,即采用粗糙集进行数据约简,去除冗余属性,然后利用决策树方法来产生分类所用到的规则。并在知识工程研究已取得的相关研究成果基础上,进行了重要改进和创新——构建了基于粗糙边界的决策树优化的新算法。

(3)针对三种基于粗糙集的决策树生成算法,分析了基于正区域、基于粗糙边界和基于依赖度的属性选择标准的关系,并证明了这三种属性选择标准彼此等价。然后以正区域的属性选择标准为代表,分析了基于正区域的决策树生成算法的优点和不足。针对这些不足,制定出了一种新的属性选择标准,即基于伴随正区域的属性选择标准。用新的属性选择标准生成决策树,能较好地避免 ID3 算法生成决策树后再剪枝的额外开销,从而提高了决策树生成算法的效率,具有较强的泛化能力。

(4)分析现有的决策树的研究状况,发现此研究主要集中在利用各种启发信息来度量属性的重要程度,或利用各种策略对决策树进行剪枝。为避免领域知识参与所造成的巨大代价,提高特征构建的自动化程度,更好地适用于海量数据,提出了基于关联规则的决策树优化算法:只选择那些近似精确规则,而不是挖掘得到的全体规则来构建新的属性。针对关联规则挖掘得到的新属性,利用信息增益思想提出了“近似信息增益”这一新的评价标准。在 C4.5 算法的基础上,提出了 AR - C4.5 算法,利用新生成的属性和数据本身固有的属性,共同构造决策树。

3)本书的内容安排

首先是引言部分,简单介绍了本书的研究背景、主要研究内容,简述研究的目的与意义和主要创新点。

第 1 章是绪论,简述了决策树算法的产生与发展,介绍了目前认知物理学的研究现状,以及传统的 ID3 算法的起源、思想和对它进行改进的算法 C4.5;介绍了和研究内容相关的认知物理学的研究情况,并借鉴了物理学中的原子模