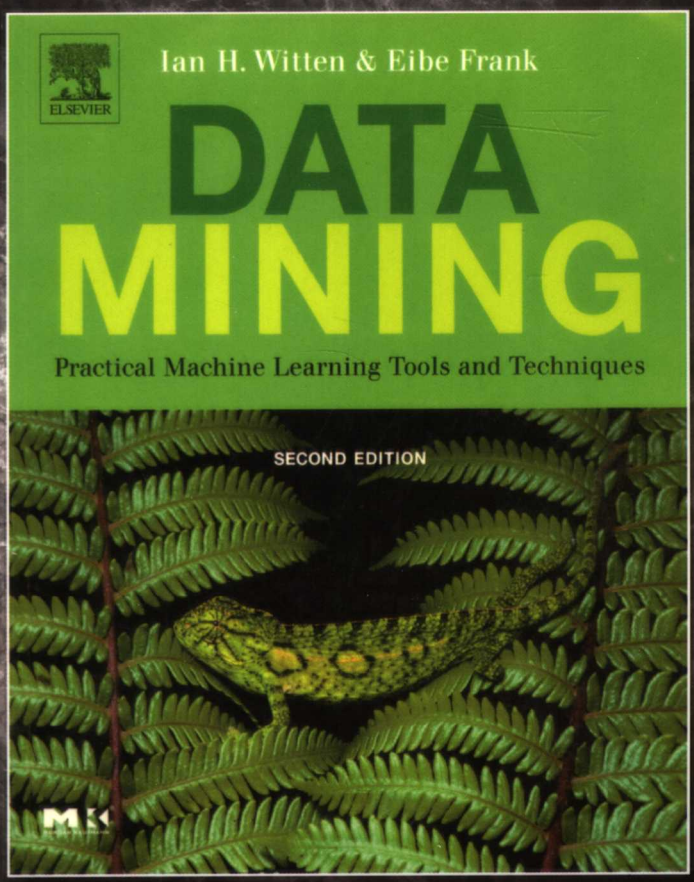
计 算 机 科 学 丛 书

原书第2版

数据挖掘
实用机器学习技术

(新西兰) Ian H. Witten Eibe Frank 著
董琳 邱泉 于晓峰 吴韶群 孙立骏 译



Data Mining
Practical Machine Learning Tools and Techniques
Second Edition

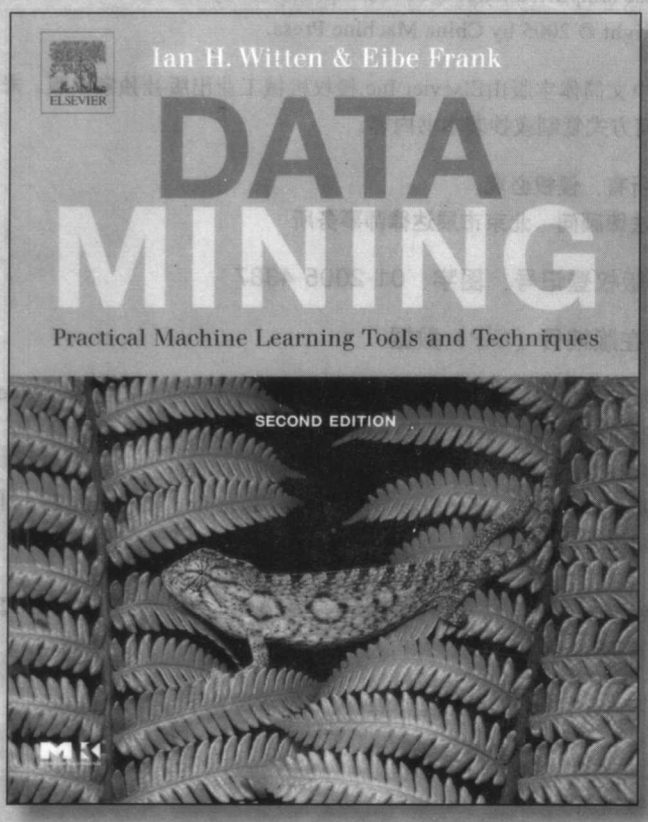
机械工业出版社
China Machine Press

计 算 机 科 学 丛 书

原书第2版

数据挖掘
实用机器学习技术

(新西兰) Ian H. Witten Eibe Frank 著 董琳 邱泉 于晓峰 吴韶群 孙立骏 译



Data Mining
Practical Machine Learning Tools and Techniques
Second Edition

机械工业出版社
China Machine Press

本书介绍数据挖掘的基本理论与实践方法。主要内容包括：各种模型（决策树、关联规则、线性模型、聚类、贝叶斯网以及神经网络）以及在实践中的运用，所存在缺陷的分析。安全地清理数据集、建立以及评估模型的预测质量的方法，并且提供了一个公开的数据挖掘工作平台Weka。Weka系统拥有进行数据挖掘任务的图形用户界面，有助于理解模型，是一个实用并且深受欢迎的工具。

本书逻辑严密、内容翔实、极富实践性，适合作为高等学校本科生或研究生的教材，也可供相关技术人员参考。

Ian H. Witten & Eibe Frank: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (ISBN: 0-12-088407-0).

本书版权登记号：图字：01-2005-4387

图书在版编目（CIP）数据

数据挖掘：实用机器学习技术（原书第2版）/（新西兰）威滕（Witten, I. H.），（新西兰）弗兰克（Frank, E.）著；董琳等译. －北京：机械工业出版社，2006.2

（计算机科学丛书）

书名原文：Data Mining: Practical Machine Learning Tools and Techniques, Second Edition

ISBN 7-111-18205-7

Ⅰ. 数…　Ⅱ. ① 威…　② 弗…　③ 董…　Ⅲ. ① 数据采集 ② 机器学习　Ⅳ. ① TP274　② TP181

中国版本图书馆CIP数据核字（2005）第157509号

机械工业出版社（北京市西城区百万庄大街22号　邮政编码　100037）

责任编辑：王　玉

北京京北制版厂印刷 · 新华书店北京发行所发行

2006年2月第1版第1次印刷

787mm × 1092mm 1/16 · 24印张

印数：0 001 -4 000册

定价：48.00元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

本社购书热线：（010）68326294

出版者的话

文艺复兴以降，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭橥了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短、从业人员较少的现状下，美国等发达国家在其计算机科学发展的几十年间积淀的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章图文信息有限公司较早意识到“出版要为教育服务”。自1998年开始，华章公司就将工作重点放在了遴选、移译国外优秀教材上。经过几年的不懈努力，我们与Prentice Hall，Addison-Wesley，McGraw-Hill，Morgan Kaufmann等世界著名出版公司建立了良好的合作关系，从它们现有的数百种教材中甄选出Tanenbaum，Stroustrup，Kernighan，Jim Gray等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及庋藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力襄助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专程为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍，为进一步推广与发展打下了坚实的基础。

随着学科建设的初步完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都步入一个新的阶段。为此，华章公司将加大引进教材的力度，在“华章教育”的总规划之下出版三个系列的计算机教材：除“计算机科学丛书”之外，对影印版的教材，则单独开辟出“经典原版书库”；同时，引进全美通行的教学辅导书“Schaum's Outlines”系列组成“全美经典学习指导系列”。为了保证这三套丛书的权威性，同时也为了更好地为学校和老师们服务，华章公司聘请了中国科学院、北京大学、清华大学、国防科技大学、复旦大学、上海交通大学、南京大学、浙江大学、中国科技大学、哈尔滨工业大学、西安交通大学、中国人民大学、北京航空航天大学、北京邮电大学、中山大学、解放军理工大学、郑州大学、湖北工学院、中国国家信息安全测评认证中心等国内重点大学和科研机构在计算机的各个领域的著名学者组成“专家指导委员会”，为我们提供选题意见和出版监督。

这三套丛书是响应教育部提出的使用外版教材的号召，为国内高校的计算机及相关专业

的教学度身定造的。其中许多教材均已为M. I. T., Stanford, U.C. Berkeley, C. M. U. 等世界名牌大学所采用。不仅涵盖了程序设计、数据结构、操作系统、计算机体系结构、数据库、编译原理、软件工程、图形学、通信与网络、离散数学等国内大学计算机专业普遍开设的核心课程，而且各具特色——有的出自语言设计者之手、有的历经三十年而不衰、有的已被全世界的几百所高校采用。在这些圆熟通博的名师大作的指引之下，读者必将在计算机科学的宫殿中由登堂而入室。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证，但我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。教材的出版只是我们的后续服务的起点。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

电子邮件：hzjsj@hzbook.com
联系电话：（010）68995264
联系地址：北京市西城区百万庄南街1号
邮政编码：100037

专家指导委员会

（按姓氏笔画顺序）

尤晋元	王　珊	冯博琴	史忠植	史美林
石教英	吕　建	孙玉芳	吴世忠	吴时霖
张立昂	李伟琴	李师贤	李建中	杨冬青
邵维忠	陆丽娜	陆鑫达	陈向群	周伯生
周克定	周傲英	孟小峰	岳丽华	范　明
郑国梁	施伯乐	钟玉琢	唐世渭	袁崇义
高传善	梅　宏	程　旭	程时端	谢希仁
裘宗燕	戴　葵			

译 者 序

无论是数字化管理的需要还是后工业化进程的要求，都使我们日益面对以前无法想像的海量数据。大型消费品公司和商场的市场部门、信用卡公司日复一日地面对如山的销售数据，绞尽脑汁地挖掘其中潜在的市场信号，工业企业的质管部门则需要正确读懂源源不断的质量情况波动报表。毫无疑问，理解乃至最终能够利用这些数据，是值得认真对待的问题。

本书由新西兰怀卡托大学计算机科学系的Ian H. Witten 和Eibe Frank 两位专家集多年的研究和教学成果精心撰写而成。本书（1999年初版）以及配套的Weka 软件一直受到全世界读者和用户的好评。2004年，Witten教授荣获了国际信息处理研究协会（IFIP）颁发的 Namur奖项，这是一个两年一度、用于奖励那些在信息和通信技术的社会应用方面做出杰出贡献及具有国际影响的荣誉奖项。2005年8月，在第11届ACM SIGKDD国际会议上，怀卡托大学的Weka小组荣获了数据挖掘和知识探索领域的最高服务奖，Weka系统得到了广泛的认可，被誉为数据挖掘和机器学习历史上的里程碑，是现今最完备的数据挖掘工具之一（已有11年的发展历史）。在本书第二版的翻译之际，我们欣喜地发现，Weka的每月下载次数已超过万次。本书被很多大学选作专业教材，并在许多学术研究文献中被频繁引用。这些都从侧面验证了本书的杰出成就。

本书的一大特点是能满足不同程度的读者的需求，既有基本理论介绍又有实践应用，读者可以根据需要进行选读而不失连贯性。本书作者Witten教授一再强调，数据挖掘及其不可或缺的技术基础——机器学习，是一个新兴的、充满希望的领域，其应用的前景是极其宽广的。本书既面向有一定专业技术基础、想对技术层面作全面深入了解的读者，也适合于技术基础有限的普通初学者。对于初学者来说，要在短期内尝试入门是难以想像的，然而这部分人群也正是本书要服务的对象。难能可贵的是本书所述的核心技术大都在Weka 系统中得以实现，这不仅可作为学习工具，读者还可以按照自己的需要使用Weka进行数据分析，或在此基础上自行开发创新。

对高等学校的学生来说，本书无疑是一本逻辑严密、内容翔实、极富实践性的教科书；原本对数据挖掘一无所知或有意了解一番的人们，无论是证券专业投资者还是超级市场数据分析员，甚至是面对一大堆彩票数据试图做些什么的业余爱好者，相信我们，读完这本循序渐进、例证充分的入门书籍，再辅之以新西兰怀卡托大学免费下载的、功能强大的Weka系统，成为一个专业水平的数据挖掘者绝非遥不可及的梦想!

我们在翻译的过程中，时时刻刻感受到Witten教授和Frank博士的良苦用心和巨大付出。他们体会到了广大机器学习潜在用户的需求所在，在成功地领导、设计并开发出Weka系统后，于1999年出版了此书的第一版。近年来随着数据挖掘技术的更新和发展，经过Weka研究小组的辛勤工作，Weka软件也日趋成熟完善。两位作者在5年之后重新审视了与Weka软件配套的这本著作，推出了第二版。

本书的翻译是集体完成的，参与翻译工作的有董琳、邱泉、吴韶群、于晓峰、孙立骏，都来自中国，或是对新西兰充满憧憬的新移民，或是负笈留学的莘莘学子。前面四位译者都

曾经或仍然学习、工作于怀大计算机科学系。在翻译过程中，我们无一不被Witten教授和Frank博士的睿智和巨大付出所打动，秉持“信、雅、达”的翻译原则，诚惶诚恐、尽了最大的努力，希望奉献给广大读者一部忠实反映原著风貌的科技书籍。

当然，要翻译好一本书并不是一件容易的事情，我们的水平还很欠缺，本书的翻译一定存在不少问题，还望各位读者批评指教。最后，在翻译过程中，Witten教授和Frank博士对我们的翻译工作给予了许多宝贵的建议和指导，Weka研究小组的成员也对翻译工作给予了很多的关心和支持，我们在此诚致谢意。真诚希望广大中文读者在读完本书后，会认可我们在这白云之乡为大家所尽的绵薄之力。

译　者

2005年9月30日

于新西兰

中文版前言

Nihau（“你好”应为“Nihao”，——译者注）！有机会为新版的《数据挖掘》中文版写上三言两语无疑是很令人高兴的事。我们中的一位 (Ian H. Witten)，从30年前算起曾在武汉、南京、北京等处逗留过，在寥寥几次短暂的中国之行里，亲眼目睹了这个国家从一个只能被描绘成Mamahuhu（“马马虎虎”——译者注）的基础上迅速发展成拥有成熟和先进技术的国家。我们俩人都为本书能对数据挖掘和机器学习领域的关键技术的进一步发展做出贡献而感到欣慰。

西方的作者们时常诧异于偶然发现他们的书籍出现在中国，有时是在中国的再版，有时是完全的中文版。若这次不是我们的学生，我们也将对此一无所知。许多年以前，有人带来一本名为《海量数据管理》的中文版书籍，我们俩中的一位是此书的作者之一。令人遗憾的是，这个翻译本没有被很好地审校，最为突出的一个严重错误是该书的封面上只出现了一位作者的名字——实际上另两位作者在书中的任何部分都没有提到！大家可以想像一下他们的感受。

遇到执著如一的学生是做学问的巨大快乐之一。大约18个月以前，我们的研究生给我们看了《数据挖掘》一书在中国的再版。我们吃惊不已——更让我们吃惊的是随着课程的进展，他们打算将此书翻译成中文。当时，我们的第二版即将完成，我们因此能有机会将中文的翻译工作托付给我们一直以来了解和可信赖的人。

也许可以给作者的最大的肯定莫过于为他们的作品做艰辛的翻译。我们深深地、真诚地感激我们的学生和朋友：董琳、邱泉、于晓峰、吴韶群和孙立骏。感谢他们为此所付出的大量时间和巨大努力。Xiexie（“谢谢”——译者注）！我们知道他们已经理解掌握了书中的内容，因为他们指出了英文版中的一些隐藏的错误，并且向我们提出了一些尖锐的、使我们难以回答的问题。

希望你们能够得益于他们的努力。

英文原文

Nihau! *It is a great pleasure to have the opportunity to write a few words for this new Chinese edition of Data Mining. During a few all-too-short visits to China, beginning thirty years ago with sojourns in Wuhan, Nanjing and Beijing, one of us (IHW) has seen the country blossom in technical maturity and sophistication from a base that all those years ago can only be described as* mamahuhu. *We are both delighted that our book will contribute to further extended growth in the key technologies of data mining and machine learning.*

Western authors are often bemused by occasional sightings of their books in China, sometimes reprints, sometimes complete Chinese editions. We would never know about them were it not for our students. Many years ago someone brought in a Chinese version of a book

co-authored by one of us entitled Managing Gigabytes. *Embarrassingly, this translation had not been properly checked, and a devastatingly prominent error was that only one author's name appeared on the cover of the book-indeed the other two were not mentioned anywhere inside either! You can imagine how they felt.*

Dedicated and committed students are one of the greatest pleasures of being an academic. About eighteen months ago members of our graduate course showed us a reprint of Data Mining *that had been produced in China. We were astonished-even more so because as the course progressed they conceived the idea of actually translating the book into Chinese. At that time we were just completing the second edition, which gave the opportunity of having a Chinese translation by reliable people who we knew well that was bang up to date.*

Perhaps the greatest compliment anyone can pay an author is to painstakingly translate their work for them. We are deeply and sincerely grateful to our students and good friends Lin Dong, Quan Qiu, Xiaofeng Yu, Shaoqun Wu and Lijun Sun for the tremendous amount of time and energy they have put into this project. Xiexie! *We know they have mastered the material because they point out obscure errors in the English edition and ask us penetrating questions that we are hard pressed to answer ourselves.*

We hope you will benefit from their efforts.

Ian H. Witten

Eibe Frank

2005年12月7日

序

技术的发展让我们能够捕获和存储大量的数据。在这些数据集中寻找模式、趋势和异常之处，并且以简单的数量模型归纳之，是当今信息时代的巨大挑战之一——将数据转化为信息，将信息转化为知识。

数据挖掘和机器学习已获得了令人瞠目结舌的进步。统计学、机器学习、信息理论以及计算技术的有机结合，创建了一门具备坚实数学基础和强大工具的完备科学。Witten 和Frank在本书中展示了这个进展的许多方面，辅之以关键算法的实现。因此，这是数据挖掘、数据分析、信息理论以及机器学习多方结合的一个里程碑。如果你在过去的十年里未能追随这个领域（的进步），本书是赶上这个激动人心的进展的绝好机会。如果你一直与该领域同步，那么由Witten 和Frank所撰写的本书及配套的开源工作平台Weka ，可以成为你的工具箱的有用补充。

本书展示了自动从数据中提取模型，然后验证这些模型的基本理论。非常出色地解释了各种模型（决策树、关联规则、线性模型、聚类、贝叶斯网以及神经网），以及如何在实践中运用它们。在这个基础上，讲解了各种方法的实施步骤以及所存在的缺陷。解释了如何安全地清理数据集、如何建立模型，以及如何评估模型的预测质量。本书的大部分是教学指导，但第二部分则全面地描述了系统是如何工作的，并且引导读者游历了作者们在网站上提供的一个公开的数据挖掘工作平台。这个Weka工作平台拥有一个引导读者进行数据挖掘任务的图形用户界面、并拥有出色的、有助于理解模型的可视工具。Weka系统本身是一个有用并深受欢迎的工具，同时又是对本书的一个绝佳补充。

本书以非常容易理解的方式展示了这门新的学科：既是用来训练新一代实际工作者和研究者的教科书，同时又能让像我这样的专业人员受益。Witten和Frank热衷于简单而优美的解决方案。他们对每个主题都采用这样的方法，用具体的实例来讲解所有的概念，促使读者首先考虑简单的技术，当简单的技术不足以解决问题时，便提升到更为复杂的高级技术。

如果你对数据库有兴趣，并且对机器学习方面所知甚少，本书无疑是赶上这个激动人心的技术进步的好机会。如果你有数据要分析和理解，本书和所附的Weka工具将是一个绝佳的起步。

Jim Gray，微软研究院

前　言

计算和通讯的结合建立了一个以信息为源的新领域。但绝大多数信息尚处于它的原始状态：数据。假如数据被定性为记录下的事实，那么信息就是构成数据基础的一系列模式或期望。在数据库中有大量信息被锁定，即那些具有潜在重要性，但尚未被发现和表达出来的信息。我们的任务就是要将它们揭示出来。

数据挖掘是将隐含的、尚不为人所知的，同时又是潜在有用的信息从数据中提取出来，建立计算机程序，自动在数据库中细察，以发现规律或者模式。假如有明显的模式被发现，将可能被归纳以对未来的数据做出准确的预测。当然，问题还是会有的，许多模式可能是陈腐或是没有意义的。另有一些是虚假的，是由于某些具体数据集的偶然巧合而产生的。在现实生活中数据是不完美的：有些部分遭到篡改，有些会丢失。所有发现的东西都是不精确的：任何规律都有例外、任何事例都有规律所不能覆盖到的情况。算法必须稳健得足以应付不完美的数据，提取出不精确但有用的规律。

机器学习为数据挖掘提供了技术基础，用于将信息从数据库的原始数据中提取出来，以可以理解的形式表达、并可以用作多种用途。这是一个抽象的过程：取得数据，据实地推导出数据的结构。本书将介绍在数据挖掘实践中，用以发现和描述数据中的结构模式而采用的机器学习工具和技术。

就像所有新萌发的技术会得到强烈的商业关注一样，数据挖掘的运用也受到大量的技术或大众出版社的追捧。夸大其实的报道声称可以建立学习算法在数据的海洋中遨游来发现秘密。但机器学习中绝没有什么神奇、隐藏的力量，没有炼金术。相反，有的只是一些确实的、简单实用的技术，能将有用的信息自原始数据中提取出来。本书介绍了这些技术并展示了它们是如何工作的。

我们将机器学习解释为从样本中得到结构性的描述。这种描述用于预测、解释和理解。有些数据挖掘应用着重于预测：从数据所描述的过去，预测将来在新的情况下将会发生什么，通常是猜测新的样本分类。但我们同样感兴趣，也许更感兴趣的是“学习”的结果是一个可以用来对样本进行分类的真实结构描述。这种结构描述不仅支持预测，也支持解释和理解。根据我们的经验，在绝大多数数据挖掘实践应用中，用户最感兴趣的莫过于取得对样本实质的把握。事实上，这是机器学习优于传统的统计模型的一个主要优点。

本书解释了许多种机器学习方法，有些是出于教学目的，仅仅罗列了纲要以解释清楚基本的原理。其他的则是实际中的、现在正运用于实际工作的系统。许多方法是时新的，在近几年中发展起来的。

我们创建了一个以Java语言编写的完整的软件资源，用以说明本书中的思想。软件的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），简称Weka[⊖]，它的源代码可通过国际互联网*www.cs.waikato.ac.nz/ml/weka*得到。它几乎完全产业化地实现了本

⊖ Weka（发音与Mecca押韵）是一种具有奇特天性、不会飞的鸟，只在新西兰发现。

书中介绍的所有技术。它包括机器学习方法的说明性代码和实现技术。针对一些简单技术，提供了清楚而简洁的实现，这些简单技术的设计目的是为了帮助理解机器学习中的运作机制。Weka还提供了一个工作平台，完整、实用、高水准地实现了许多流行的学习方案，这些方案能够直接运用于实际的数据挖掘或研究领域。最后它提供了一个Java类库形式的框架，这个框架支持嵌入式机器学习的应用，乃至新的学习方案的实现。

本书的目的是介绍用于数据挖掘的机器学习工具和技术。读完本书后，你将了解这些技术，并体会它们的功效和可应用性。如果你希望用自己的数据加以实践，用Weka 软件能轻易地做到。

本书跨越了商业书籍上提供的一些在数据挖掘研究案例里非常实际的方法和当前机器学习教科书中的以理论为主的阐述之间存在的鸿沟（在第1章最后的补充读物里简要介绍了这些书）。这个鸿沟是相当大的。为了让机器学习的技术运用得富有成果，需要理解它们是如何工作的。这不是一种可以盲目应用而后期待好结果出现的技术。不同的问题要应用不同的技术，但是哪些技术更适合某种特定情况却不明显：需要知道有可能的解决方案的范围。我们所论及的技术范围非常广泛，这是因为和其他商业书籍不同，本书无意推销某种特定的商业软件或方案。我们列举了大量的例子，但是在例子中所应用的数据集却小得足以跟踪每一步的进展。真实的数据集太大，不能做到这一点（而且通常包含商业秘密）。我们选择的数据集没有展示大型数据集的实际问题，但能够帮助你理解不同技术的作用，它们是如何工作的，以及它们的应用范围是什么。

本书面向对实际工作中的数据挖掘所包含的原理和方法感兴趣的、并有技术基础的普通读者，同样适用于需要获得这方面新技术的信息专家，以及所有希望从技术层面具体理解机器学习包含什么的读者。本书也是为有着一般兴趣的信息系统的实际工作者们所写，例如：程序员、咨询顾问、开发人员、信息技术管理员、规范编写者、专利审核者、好奇的业余爱好者，以及学生和教授。他们需要一本拥有大量实例且简单易读的有关机器学习主要技术是什么，它们做什么，如何运用它，以及它们是如何工作的书。本书面向实际，重点突出“如何做”，同时包括许多算法、代码和实现。所有在实际工作中进行数据挖掘的读者将直接得益于书中叙述的技术。本书目的是帮助试图穿越夸大其词的宣传找出机器学习真谛的人们，也将帮助需要可行的、非学术的、值得信赖的方案的人们。我们避免了对理论和数学知识方面的特殊要求，在页边用浅灰色条形所标记的内容是可选部分，通常是为了照顾对理论和技术感兴趣的读者，跳过这部分内容不会对整体的连贯性有任何影响。

本书分为几个层次，使其既适合对基本概念感兴趣的读者的需要，也能适合想深入详尽地了解、掌握所有技术细节的读者的需要。我们相信机器学习的消费者们需要更多地了解他们运用的算法如何工作。我们常常可以看到，优秀的数据模型与它的诠释者是分不开的，诠释者需要知道模型是如何产生的，从而感知这个模型的长处和局限性。当然，并非所有的使用者都有必要对算法的精妙之处有很深入的理解。

我们把机器学习方法的详尽阐述处理为几个连续部分。最顶层，也就是前三章，读者将学习到基本理论。第1章通过一些例子来说明机器学习是什么，它能用在什么地方。同时提供了一些真实的实际应用。第2、3章叙述了不同的输入和输出，或者称之为知识表达（knowledge representation）。不同的输出要求不同的算法。在下一层的第4章，介绍了机器学习的基本方法，已简化以方便理解。这里原理的给出牵涉到许多算法，但没有涉及到算法所

包含的复杂的细节和精妙的实现方案。为了从机器学习技术的应用升级到解决具体的数据挖掘问题上，必须对机器学习的效果有一个评估。第5章可以单独进行阅读。它能帮助读者评估从机器学习中得到的结果，解决性能评估中出现的某些复杂的问题。

在最底层也是最详细的一层，第6章完全详尽地揭示了实现整系列机器学习算法的步骤，以及在实际应用中为了更好工作所必需的、较为复杂的部分。尽管一些读者也许想忽略这部分的具体内容，但只有到这一层，我们才涉及到完整的、可运作的、并经过测试的机器学习的Weka实现方案。第7章讨论了一些涉及机器学习输入的实际问题。例如：选择和离散属性，这里我们谈到几个更高级的技术来提炼和组合从不同学习技术得出的结果。第一部分的最后一章是展望未来的发展趋势。

本书叙述了在实际工作中的绝大多数机器学习方法。但是，没有涉及到加强学习（reinforcement learning），因为这在实际的数据挖掘中极少应用，也没有包括遗传算法（genetic algorithm），因为它仅仅是一个优化技术；同样也没有包含关系学习（relational learning）和归纳逻辑程序设计（inductive logic programming），因为它们很少被主流数据挖掘应用采纳。

阐明本书思想的数据挖掘系统在第二部分，这是为了清楚地把理论部分和从实践角度如何使用区分开来。如果你急于分析你的数据，不想被技术细节所困扰，可以直接从第4章跳到第二部分。

选定Java来实现本书的机器学习技术，是因为作为面向对象的编程语言，它允许用统一的界面进行学习方案和方法的前期和后期处理。用Java取代C++，Smalltalk 或者其他面向对象的语言，是因为用Java编写的程序能够运行在大部分计算机上，而不需要重新进行编译，或复杂的安装过程，或者最坏的情形，需要修改源代码。Java程序编译成字节码后，能运行在任何安装了适当解释器的计算机上。这个解释器称为Java虚拟机。Java虚拟机和Java编译器能免费用于所有重要平台。

像所有广泛运用的编程语言一样，Java也受到过批评。虽说本书无意对此说三道四，有些批评无疑是正确的。无论如何，在所有的现有可供选择的、能得到广泛支持的、标准化的及拥有详尽文档的编程语言中，Java似乎是本书的最佳选择。它主要的不足是它的运行速度，或者说它在速度上有缺陷。由于执行前，虚拟机先要把字节码翻译成机器编码，执行一个Java程序要比相应的用C语言编写的程序慢好几倍。如果虚拟机用一个即时编译器，根据我们的经验，结果要慢三到五倍。即时编译器将整个字节码块翻译成机器编码，而不是一个接一个地翻译字节码，所以它的运行速度能够得到大幅度的提高。如果对于你的应用来说，这个速度依然很慢，有些编译器能够跳过字节码这一步，直接将Java程序转换成机器编码。当然这种编码不能运行在其他平台上，以至于牺牲了Java的一个最大的优势。

更新和改写过的部分

我们在1999年完成了本书的第一版。现在，2005年4月，我们正在对本书的第二版进行润饰。数据挖掘和机器学习领域在这些年中已臻成熟。虽说在这一版中基本的核心材料没有发生变化，但我们还是尽可能地利用这个机会更新内容，使它能反映过去五年中的变化。当然，也有错误要更正，我们已将错误收在可以公开查询的勘误表中。事实上，虽说发现的错误少得惊人，我们还是希望在第二版中错误会更少（第二版的勘误表可以从本书的主页

http://www. cs.waikato.ac.nz/ml/weka/book.html 中得到)。我们已从整体上对材料进行了编辑，使这本书跟上时代，征引文献的数量翻番。那些最吸引人的部分也已加入了新的材料，下面是重点介绍。

应广泛要求，我们对神经网络部分进行了全面的补充，在4.6节里增加了感知器（perceptron）以及和它密切相关的 Winnow 算法；6.3节里增加了多层感知器（multilayer perceptron）和反向传播（backpropagation）算法。我们对使用核感知器（kernel perceptron）和径向基函数网络（radial basis function network）得到非线性决策边界的实现方法补充了最新的内容。同样应读者的要求，对贝叶斯网络部分增加了一个新的章节。描述了如何基于这些网络来学习分类器，以及如何用 AD树来有效地实现。

本书第一版中所采用的Weka工作平台已经广泛地使用并受到普遍的欢迎。它现在有了一个交互式界面，令人耳目一新，或者说是由三个独立的交互界面组成，使用方便许多。其中最主要的交互界面是Explorer界面。用户可以通过菜单选择和表单填写形式来访问Weka的所有工具。另外一个是Knowledge Flow界面，用户可以设计对数据流处理的配置。还有一个是Experimenter界面，用户可以对一组数据集应用不同参数设置的机器学习算法来建立自动测试、收集测试结果的统计数据，并对结果做重要性测试。这些界面为成为一个数据挖掘实际工作者降低了门槛。在书中我们给出了如何运用这些界面的详细说明。当然，这本书可以离开Weka软件单独使用。为了强调这一点，我们把所有有关工作台的内容单独放在本书最后的第二部分介绍。

Weka数据挖掘的功效在过去五年里不仅变得日益方便，也日渐强大和成熟。如今，它集合了空前数量的机器学习算法和相关技术。它的发展既是被这个领域的发展所推动，也受到了用户和需求的推动。我们相当了解数据挖掘的实际工作者们在想什么。新版本应该包括哪些内容，我们在这方面的体验是值得和盘托出的。

前几章介绍了一些概要的基础内容，相对来说内容改动较少。在第1章里，我们增加了一些具体领域的应用例子。在第2章里增加了有关新的稀疏数据和一些有关字符串属性和日期属性的内容。第3章增加了对交互式决策树结构的介绍，这是一种有用并有启发性的技术，教你如何对自己的数据手工建立一个决策树。

第4章除了介绍分类的线性决策边界，神经网络的基础知识外，还包括一个新的内容：应用于文档分类的多项贝叶斯模型和 logistic回归。在过去的五年，很多人对基于文本的数据挖掘很感兴趣，我们在第2章介绍字符串属性，在第4章介绍了用于文档分类的多项贝叶斯法，在第7章介绍了文本的转换。第4章介绍了许多关于寻找实例空间的高效数据结构的新知识，*k*D树和新近开发的球树。这两个数据结构用来高效地找出最近邻数据，同时能加速基于距离的聚类。

第5章叙述了机器学习的统计评估原则，没有什么改变。主要的增加，除了用于评估预测功效的Kappa统计量的一个注解外，就是成本敏感学习的更多详细对策。我们描述如何运用分类器，在不考虑成本的情况下做成本敏感的预测；或者在训练过程中将成本考虑进来，建立一个成本敏感模型。我们也论及了流行的成本曲线的新技术。

第6章有几处增加，除了上面提到的有关神经网络和贝叶斯网分类器的内容外，也为成功的RIPPER规则学习器所运用的试探法提供了更多的活生生的细节。我们描述了如何利用模型树为数值预测建立规则。展示了如何将局部加权回归运用到分类问题中。最后，叙述了*X*均值

聚类算法，这是对传统k均值算法的一个巨大改进。

有关输入输出的第7章改动最多，因为这方面是近来机器学习实践发展的重点。我们描绘了新的属性选择方案，例如特征搜索和支持向量机的使用；以及新的组合模型技术如叠加回归、叠加logistic回归、logistic模型树以及选择树等等。完整地阐述了LogitBoost（在首版中提到过但未能解释）。新增加了一节关于有用的数据转换的内容，包括重要部分分析以及文本挖掘和时间序列的转换。我们还涉及了在利用没有标签数据来改进分类，包括联合训练和co-EM方法方面的最新进展。

新版重写了第一部分的最后一章有关新的发展方向和不同的远景，以跟上时代的步伐，现在它包括了诸如对抗性学习及无处不在的数据挖掘等新挑战。

致谢

写致谢部分永远是最令人愉悦的！许多人曾帮助过我们，我们很高兴借这个机会来感谢他们。本书发轫自新西兰怀卡托大学计算机科学系的机器学习小组。我们从这个小组的教职成员处得到了慷慨的鼓励和帮助：John Cleary、Sally Jo Cunningham、Matt Humphrey、Lyn Hunt、Bob McQueen、Lloyd Smith及Tony Smith。特别感谢Mark Hall、Bernhard Pfahringer，尤其是Geoff Holmes、小组的领导者和激励的源泉。所有在机器学习小组工作过的成员都对我们的思想有过贡献：特别要提一下Steve Garner、Stuart Inglis 和 Craig Nevill-Manning，为他们在小组工作初创时的帮助，因为那时成功还显得很渺茫，诸事都那么困难。

用于展示书中思想的Weka系统是本书的一个重要组成部分。它是由两位作者共同构思，并由Eibe Frank及Len Trigg、Mark Hall设计和实现的。机器学习小组的许多成员都在早期做了许多贡献。自从第一版以来，Weka小组已经有了可观的扩大，如此多的人做过贡献以致无法在此向每位一一致谢。我们为Remco Bouckaert 在贝叶斯网络的实现，Dale Fletcher在数据库的诸多方面、Ashraf Kibriya 以及 Richard Kirkby 在举不胜举方面的贡献，Niels Landwehr在logistic模型树方面，Abdelaziz Mahoui在K*的实现，Stefan Mutter在关联规则方面，Gabi Schmidberger和Malcolm Ware在许多不同方面的贡献，Tony Voyle在最小均方回归方面，Yong Wang在Pace回归和M5'的实现，Xin Xu 在*JRip*，logistic回归以及其他许多方面的贡献而表示感谢。我们为所有那些献身于我们小组的人们表示真诚的感激，也同样感谢所有来自怀大机器学习小组以外的帮助。

我们的弱点是身处南半球的一个遥远的角落（但很美），我们珍视每一位我们系的来访者所起的重要作用，他们既是我们的义务宣传员，也启迪了我们的思想。我们要特别提到Rob Holte、Carl Gutwin以及Russell Beale，他们每一位都曾访问过几个月；而David Aha虽然只来过几天，却用他的热情和鼓励在小组的初创期起到了大作用；Kai Ming Ting和我们在第7章的许多主题上一起工作了两年，帮助我们的机器学习走上了正轨。

怀大的学生们在这个项目的发展中起到了极大的作用。Jamie Littin在ripple-down规则和关系学习方面工作。Brent Martin探索基于实例的学习和嵌套的基于实例表达。Murray Fife也为关系学习做了艰辛的工作。Nadeeka Madapathage探究了运用功能语言来表达机器学习算法。其他研究生也给了我们数不胜数的影响，尤其是Gordon Paynter，YingYing Wen以及Zane Bray，他们都曾和我们一同进行文本挖掘工作。怀大的同事们Steve Jones和Malika Mahoui 也为这个方面以及其他机器学习课题做了大量贡献。近期，我们从来自Freiburg的交流学生处得

益颇多，其中有Peter Reutemann和Nils Weidmann。

Ian Witten还想感谢他以前在卡尔加里大学的学生们，尤其是Brent Krawchuk、Dave Maulsby、Thong Phan以及Tanja Mitrovic，他们都对他在机器学习早期的思想形成有所帮助。同样还有卡尔加里大学的教师Bruce MacDonald、Brian Gaines以及David Hill，坎特伯雷大学的教师John Andreae。

Eibe Frank感激他以前在卡尔斯鲁厄大学的导师Klaus-Peter Huber（现在SAS学院），是Klaus将机器奇妙的学习能力展现给Eibe。在他的学术旅程中，Eibe还得益于和他的加拿大同行Peter Turney、Joel Martin、Berry de Bruijn，以及德国的Luc de Raedt、Christoph Helma、Kristian Kersting、Stefan Kramer、Ulrich Rückert和Ashwin Srinivasan的学术交流。

Morgan Kaufmann公司的Diane Cerra和Asma Stephan为本书的构架做了辛勤的努力，我们的责任编辑Lisa Royse使整个出版过程极为顺畅。Bronwyn Webster则对怀卡托大学这方面的工作给予了出色的支持。

感谢那些默默无闻的书评家的努力，特别是其中一位对本书提出了好几处中肯及富有建设性的意见、帮助我们做了极大的改进。此外，我们还要向负责加利福尼亚大学，Irvine分校的机器学习数据知识库的图书馆工作人员表示感谢，经他们仔细挑选的数据集在我们的研究里是无价之宝。

我们的研究得到了新西兰研究、科学和技术基金以及新西兰皇家Marsden基金协会的赞助。怀卡托大学计算机科学系也通过各种方法予以了慷慨帮助，我们特别感激Mark Apperley富于启迪的领导和温暖的鼓励。第一版的一部分是两位作者在访问加拿大卡尔加里大学期间完成的，我们感谢那里的计算机科学系的支持，还要感谢参与我们机器学习课程教学实验的学生们在冗长课上的积极和有益的态度。

在写第二版时，Ian得到了加拿大杰出信息学研究学会和南阿尔伯特的Lethbridge大学的帮助，他们提供了所有的作者们梦寐以求的东西——一个有着舒心及快乐环境的安静场所。

最后，也是最重要的，我们要感谢我们各自的家庭和伴侣。Pam、Anna和Nikki都十分明白家里有一个写书的人意味着什么（“不要再干了！”），但还是让Ian继续干下去并写完了这本书。Julie总是那么支持，即便Eibe有时要在机器学习实验室中开夜车，而Immo和Ollig则是闲暇时的最大快乐。不管我们源自何方，加拿大、英格兰、德国、爱尔兰、萨摩亚，新西兰将我们聚在一起并给了我们一个理想的、甚至可以说是诗情画意的场所来完成这项工作。