HZ BOOKS
华章教育

计 算 机 科 学 丛 书

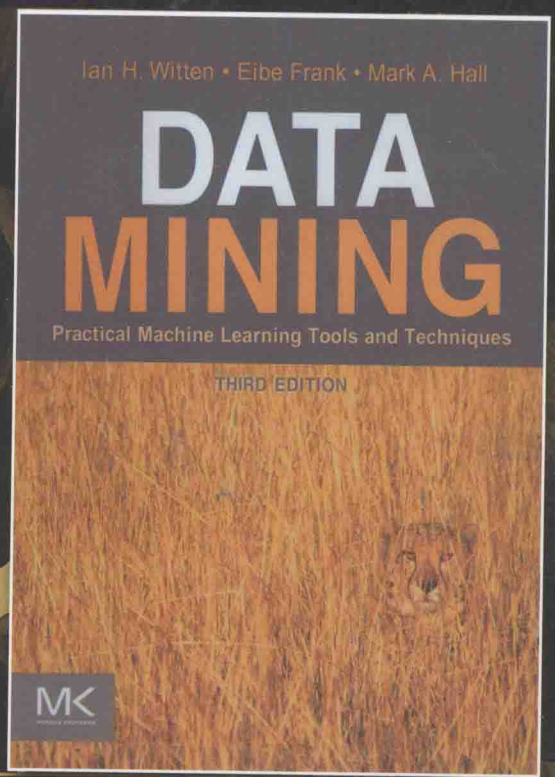
ELSEVIER
爱思唯尔

原书第3版

数据挖掘
实用机器学习工具与技术

(新西兰) Ian H. Witten Eibe Frank Mark A. Hall 著
怀卡托大学
李川 张永辉 等译
四川大学

Data Mining
Practical Machine Learning Tools and Techniques Third Edition



机械工业出版社
China Machine Press

计 算 机 科 学 丛

原书第3版

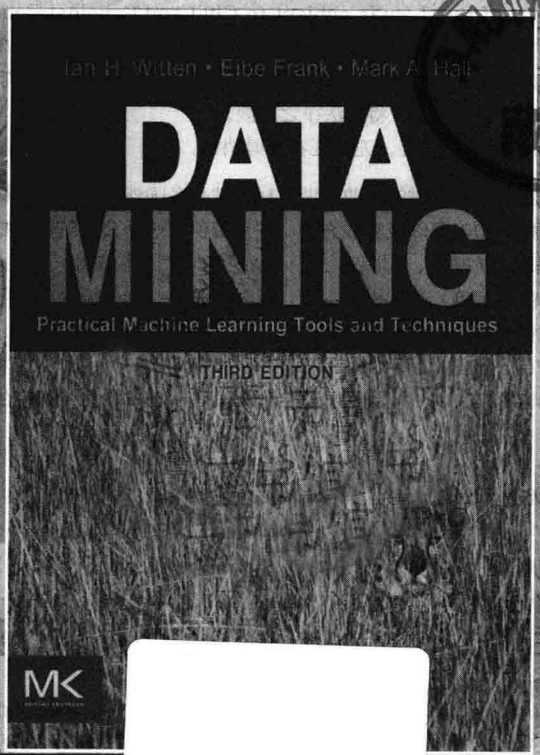
数据挖掘

实用机器学习工具与技术

(新西兰) Ian H. Witten Eibe Frank Mark A. Hall 著
怀卡托大学
李川 张永辉 等译
四川大学

Data Mining

Practical Machine Learning Tools and Techniques Third Edition



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据挖掘：实用机器学习工具与技术（原书第 3 版）/（新西兰）威滕（Witten，I.H.），（新西兰）弗兰克（Frank，E.），（新西兰）霍尔（Hall，M.A.）著；李川等译．—北京：机械工业出版社，2014.4
（计算机科学丛书）
书名原文：Data Mining: Practical Machine Learning Tools and Techniques, Third Edition

ISBN 978-7-111-45381-9

I. 数… II. ①威… ②弗… ③霍… ④李… III. 数据采集 IV. TP274

中国版本图书馆 CIP 数据核字（2014）第 035715 号

本书版权登记号：图字：01-2011-4804

本书是机器学习和数据挖掘领域的经典畅销教材，被众多国外名校选为教材。书中详细介绍用于数据挖掘领域的机器学习技术和工具以及实践方法，并且提供了一个公开的数据挖掘工作平台 Weka。本书主要内容包括：数据输入 / 输出、知识表示、数据挖掘技术（决策树、关联规则、基于实例的学习、线性模型、聚类、多实例学习等）以及在实践中的运用。本版对上一版内容进行了全面更新，以反映自第 2 版出版以来数据挖掘领域的技术变革和新方法，包括数据转换、集成学习、大规模数据集、多实例学习等，以及新版的 Weka 机器学习软件。

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

出版发行：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：盛思源

印　刷：北京瑞德印刷有限公司		版　次：2014 年 5 月第 1 版第 1 次印刷	
开　本：185mm × 260mm　1/16		印　张：30	
书　号：ISBN 978-7-111-45381-9		定　价：79.00 元	

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

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，还揭示了学术的源变，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都既是机遇，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和发展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到积极的推动作用，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公司较早意识到“出版要为教育服务”。自 1998 年开始，我们就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与 Pearson，McGraw-Hill，Elsevier，MIT，John Wiley & Sons，Cengage 等世界著名出版公司建立了良好的合作关系，从他们现有的数百种教材中甄选出 Andrew S. Tanenbaum，Bjarne Stroustrup，Brain W. Kernighan，Dennis Ritchie，Jim Gray，Afred V. Aho，John E. Hopcroft，Jeffrey D. Ullman，Abraham Silberschatz，William Stallings，Donald E. Knuth，John L. Hennessy，Larry L. Peterson 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、研究及珍藏。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”已经出版了近两百个品种，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。其影印版“经典原版书库”作为姊妹篇也被越来越多实施双语教学的学校所采用。

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。随着计算机科学与技术专业学科建设的不断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将步入一个新的阶段，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。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：

华章网站：www. hzbook. com
电子邮件：hzjsj@ hzbook. com
联系电话：（010）88379604
联系地址：北京市西城区百万庄南街 1 号
邮政编码：100037

华章教育
华章科技图书出版中心

译者序

Data Mining: Practical Machine Learning Tools and Techniques, Third Edition

信息技术正以惊人的速度将现实世界中的信息转化为数据，存储到各类计算机系统中，且这一过程的发展态势可能超出人类的有限预想。其中蕴含着的，不仅是自然和生命，还有人类的行为、情感和历史。同我们生存其中的真实自然界一样，新兴的数据中潜藏着无尽的奥秘和巨大的财富，因此吸引着大批来自自然科学、人文科学以及商界的学者和技术人员投身其中。正确地解读、有效地利用这些数据是新技术革命时代照亮人类前行的灯塔。

本书前两位作者是大名鼎鼎的 Ian H. Witten 和 Eibe Frank，他们共同设计了影响广远的 Weka 系统。Weka 的设计与提出正如谷歌一样，它通过将单纯思想迅速实现给人们带来前所未有的不同感受，完美的图形界面、感性直观的可视化呈现、友好的用户界面消除了初学者的陌生感，为同行的探索时常予以灵感，又集合了前人工作的大成。而且，实验系统为高校的数据挖掘教学提供了实验环境，施惠于众人。

两位作者研发 Weka 后，将他们开发过程中的经验、实际的数据挖掘项目以及教学过程中的体会融为一体，形成此书的第 1 版。Weka 此后经历多次版本更新。1999 年 Weka 的第 1 版是 Witten 教授和 Frank 博士开发的。后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 软件日趋完善，2005 年本书推出第 2 版。第 2 版最大的变化是加入了一个专门介绍 Weka 系统的部分。得益于数据挖掘领域的飞速发展和用户日新月异的需求引导，Weka 系统在过去的十年里焕然一新，增加了大量数据挖掘功能，集成了非常丰富的机器学习算法和相关技术，于是催生了本书第 3 版的问世。第 3 版在前面两版基础上增加了大量近年来最新涌现的数据挖掘算法和诸如 Web 数据挖掘等新领域的介绍，所介绍的 Weka 系统较第 2 版增加了大约 50% 的算法及大量其他新内容。

本书的翻译是在极其紧张的条件下，经过所有团队成员的艰辛拼搏最终杀青的，其中凝聚着所有参与者的真诚与责任。本书的翻译工作由李川副教授统一协调负责，参与的人员有四川大学计算机科学与技术专业的研究生吴诗极、张永辉、李艳梅、谢世娜，他们在节假日、寒夜里加班工作，对译文字雕句琢最终有了本书的诞生。于中华副教授协助进行了本书的最终统稿。机械工业出版社的王春华、盛思源老师在本书的译著过程中给予了大力的支持和关心帮助。没有这些幕后的无私奉献，不可能有本书的面世。

尽管译者心正意诚，然则受限于自身的水平，本书一定存在不少问题，还期望各位读者给予批评、指正，各位的反馈将使本书更趋完善。最后，真诚期望本书对大家有益，这是对我们翻译工作的最大认可！

译者

2014 年 1 月 9 日夜

四川大学 DB&KE 实验室

计算和通信的结合建立了一个以信息为基础的新领域。但绝大多数信息尚处于原始状态，即以数据的形式存在的状态。假如我们将数据定义为被记录下的事实，那么信息就是在这些记录事实的数据中所隐藏的一系列模式或预期。在数据库中蕴藏了大量具有潜在重要性的信息，这些信息尚未被发现和利用，我们的任务就是将这些数据释放出来。

数据挖掘是将隐含的、尚不为人知的同时又是潜在有用的信息从数据中提取出来。为此我们编写计算机程序，自动在数据库中筛选有用的规律或模式。假如能发现一些明显的模式，则可以将其归纳出来以对未来的数据进行准确预测。当然，数据挖掘结果中肯定会出现一些问题，比如许多模式可能是不言自明的或者没有实际意义的。另一些还有可能是虚假的，或者由于某些具体数据集的偶然巧合而产生的。在现实世界中，数据是不完美的：有些被人为篡改，有些会丢失。我们所观察到的所有东西都不是完全精确的：任何规律都有例外，并且总会出现不符合任何一个规律的实例。算法必须具有足够的健壮性以应付不完美的数据，并能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，可用其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，并可用做多种用途。这是一种抽象化过程：如实地全盘接收现有数据，然后在其基础上推导出所有隐藏在这些数据中的结构。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。

就像所有新兴技术都会受到商界的强烈关注一样，关于数据挖掘应用的报道正淹没在那些技术类或大众类出版社的大肆宣扬中。夸张的报道向人们展示了通过设立学习算法就能从浩瀚的数据汪洋中发现那些神秘的规律。但机器学习中绝没有什么魔法，没有什么隐藏的力量，也没有什么巫术，有的只是一些能将有用信息从原始数据中提取出来的简单和实用的技术。本书将介绍这些技术并展示它们是如何工作的。

我们将机器学习理解为从数据样本中获取结构描述的过程。这种结构描述可用于预测、解释和理解。有些数据挖掘应用侧重于预测：从数据所描述的过去预测将来在新情况下会发生什么，通常是猜测新的样本分类。但同样令我们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掌握样本的本质。事实上，这是机器学习优于传统统计模型的一个主要优点。

本书向我们诠释多种机器学习方法。其中一部分出于方便教学的目的而仅仅罗列一些简单方案，以便清楚解释基本思想如何实现。其他则考虑到具体实现而列举很多应用于实际工作中的真实系统。很多都是近几年发展起来的新方法。

我们创建了一套综合的软件资源以说明本书中的思想。软件名称是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis, Weka[⊖]），它的 Java 源代码可以在 www. cs. waikato. ac. nz/ml/weka 中得到。Weka 几乎可以完善地实现本书中包含的所有技术。它包括机器学习方法的说明性代码以及具体实现。针对一些简单技术，它提供清楚而

⊖ Weka（发音与 Mecca 类似）是一种天生充满好奇心的不会飞的鸟，这种鸟仅在新西兰的岛屿上出现过。

简洁的实现，以帮助理解机器学习中的相关机制。Weka 还提供一个工作平台，完整、实用、高水准地实现了很多流行的学习方案，这些方案能够运用于实际的数据挖掘项目或学术研究中。最后，本书还包括一个形如 Java 类库的框架，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。

本书旨在介绍用于数据挖掘领域的机器学习工具和技术。读完本书后，你将对这些技术有所了解并能体会到它们的功效和实用性。如果你希望用自己的数据进行实验，用 Weka 就能轻易地做到。

提供数据挖掘案例研究的商业书籍中往往涉及一些非常具有实用性的方法，这些方法与当前机器学习教材中出现的更理论化、原则化的方法之间存在巨大鸿沟，本书跨越了这个鸿沟（关于本书的一些简介将出现在后面第 1 章的末尾）。这个鸿沟相当大，为了让机器学习技术应用富有成果，需要理解它们是如何工作的。这不是一种可以先盲目应用而后期待好结果出现的技术。不同的问题需要不同的技术来解决。但是如何根据实际问题来选择合适的技术并不是那么容易的事情：你需要知道到底有多少可能的解决方案。我们在本书中所论及的技术范围相当广泛，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。我们列举大量实例，但为展示实例所采用的数据集却小得足以让你搞清楚实例的整个过程。真实的数据集太大，不能做到这一点（而真实数据集的获取常受限于商业机密）。我们所选择的数据集并不是用来说明那些拥有大型数据的真实问题，而是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。

本书面向对实际数据挖掘技术所包含的原理和方法感兴趣的“技术敏感型”普通读者。本书同样适用于需获得这方面新技术的信息专家，以及所有希望了解机器学习领域技术细节的人。本书也是为有着一般兴趣的信息系统实际工作者所写的，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、业余爱好者，以及学生和专家教授。他们需要拥有这样一本书：拥有大量实例且简单易读，向读者阐释与机器学习相关的主要技术是什么、做什么、如何运用它们，以及它们是如何工作的。本书面向实际，告诉读者“如何去做”，同时包括许多算法、代码以及具体实例的实现。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。本书旨在帮助那些希望找到掩藏在天花乱坠广告宣传下的机器学习真谛的人们，以及帮助那些需要实际可行的、非学术的、值得信赖的方案的人们。我们避免对特定的理论或数学知识做过分要求。在某些涉及特定知识的地方，我们会将相关文本框起来，这些内容是可选部分，通常是为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，不管你是想走马观花地浏览基本概念，还是想深入详尽地掌握技术细节，阅读本书都可以满足你的要求。我们相信机器学习的使用者需要更多地了解他们运用的算法如何工作。我们常常可以发现，优秀的数据模型是与它的诠释者分不开的，诠释者需要知道模型是如何产生的，并且熟悉模型的长处和局限性。当然，并不要求所有的用户都对算法的细节有深入理解。

根据上述考量，我们将对机器学习方法的描述分为几个彼此承接的层次。本书共分为三部分，第一部分是关于数据挖掘的介绍，读者将在这一部分学习数据挖掘的基本思想，这一部分包括书中的前五章。第 1 章通过实例说明机器学习是什么，以及能用在什么地方，并附带提供一些实际应用。第 2、3 章给出不同的输入和输出，或者称为*知识表达*（knowledge representation）。不同的输出要求不同的算法。第 4 章介绍机器学习的基本方

法，这些方法都以简化形式出现以方便读者理解。其中的相关原理通过各种具体算法来呈现，这些算法并未包含复杂细节或精妙的实现方案。为从机器学习技术的应用升级到解决具体的数据挖掘问题，必须对机器学习的效果有一个评估。第5章可以单独阅读，它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问题。

第二部分介绍数据挖掘的一些高级技术。在最低同时也是最详细的层次上，第6章详尽揭示实现整系列机器学习算法的步骤，以及在实际应用中为更好工作所必需的、较为复杂的部分（但忽略某些算法对复杂数学原理的要求）。尽管有些读者也许想忽略这部分的具体内容，但只有到这一层，才能涉及完整、可运作并经过测试的机器学习的Weka实现方案。第7章讨论一些涉及机器学习输入/输出的实际问题，例如，选择属性和离散化属性。第8章主要介绍“集成学习”技术，这种技术综合来自不同学习技术的输出。第9章展望发展趋势。

本书阐述了在实际机器学习中所使用的大多数方法，但未涉及强化学习（reinforcement learning），因为它在实际数据挖掘中极少应用；未包含遗传算法（genetic algorithm），因为它仅仅是一种优化技术；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所采纳。

第三部分介绍Weka数据挖掘平台，它提供在第一部分和第二部分中所描述的几乎所有思想的实例。我们将那些概念性的材料从如何使用Weka的实际操作材料中清楚地分离出来。在第一、二部分每一章的结尾会给出指向第三部分中相应Weka算法的索引。读者可以忽略这些部分，或者如果你急于分析你的数据并且不愿意纠结于说明算法的技术细节，可以直接跳到第三部分。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，它允许通过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编写的程序能运行在大部分计算机上而不需要重新进行编译，不需要复杂的安装过程，甚至不需要修改源代码。Java程序编译成字节码后，能运行于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上。

在当前所有的可能选择中，能得到广泛支持的、标准化的、拥有详尽文档的编程语言，Java似乎是最佳选择。但是，由于在执行前要通过虚拟机将字节码编译为机器代码，所以Java程序的运行速度比用C或C++语言编码的相应程序慢。这个缺陷在过去看来很严重，但在过去二十年间，Java的执行效率有了大幅度提升。依我们的经验，如果Java虚拟机采用即时编译器，那么Java运行慢这个因素几乎可以忽略不计。即时编译器将整个字节码块翻译成机器代码，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。如果对你的应用来说，这个速度依然很慢，还可以选择采用某些编译器，跳过字节码这一步，直接将Java程序转换成机器代码。当然这种代码不能跨平台使用，这样牺牲了Java的一个最大优势。

更新与修改

1999年，我们完成本书的第1版，2005年初完成第2版，经过我们精心修改润色的本书第3版在2011年同读者见面。这个世界过去二十年间可谓是沧海桑田！在保留前版基本核心内容的同时，我们增加了很多新内容，力图使本书与时俱进。本书第3版较前

版接近翻倍的文字量可以反映出这种变化。当然，我们也对前版中出现的错误进行了校正，并将这些错误集中放到我们的公开勘误文件里（读者可以通过访问本书主页 http：//www. cs. waikato. ac. nz/ml/weka/book. html 得到勘误表）。

第2版

本书第2版中最主要的改变是增加了一个专门的部分来介绍 Weka 机器学习工作平台。这样做可以将书中的主要部分独立于工作平台呈现给读者，我们将在第3版中沿用这个方法。在第1版中广为使用和普及的 Weka 工作平台在第2版中已经改头换面，增加了新的图形用户界面或者说是三个独立的交互界面，这使读者使用起来更得心应手。其中最基本的界面是 Explorer 界面，通过该界面，所有 Weka 的功能都可以经由菜单选择和表单填写的方式完成；另一个界面称为 Knowledge Flow 界面，它允许对流数据处理过程进行设置；第三个界面是 Experimenter 界面，可以使用它对某一语料库设置自动地运行选定的机器学习算法，这些算法都带有不同的参数设置，Experimenter 界面可以收集性能统计数据，并在所得实验结果的基础上进行有意义的测试。这些界面可以降低数据挖掘者的门槛。第2版中包括一套如何使用它们的完整介绍。

此外，第2版还包括如下我们前面曾大致提及的新内容。我们对介绍规则学习和成本敏感评估的章节进行了扩充。为了满足普遍需求，我们增加了一些有关神经网络方面的内容：感知器及相关的 Winnow 算法，以及多层感知器和 BP 算法，Logistic 回归也包含在内。我们介绍如何利用核感知器和径向基函数网络来得到非线性决策边界，还介绍用于回归分析的支持向量机。另外，应读者要求和 Weka 新特性的加入，我们还融入了有关贝叶斯网络的新章节，其中介绍如何基于这些网络来学习分类器以及如何利用 AD 树来高效地应用这些分类器。

在过去的五年（1999—2004）中，文本数据挖掘得到极大的关注，这样的趋势反映在以下方面：字符串属性在 Weka 中的出现、用于文本分类的多项式贝叶斯以及文本变换。我们还介绍用以搜寻实例空间的高效数据结构：为高效寻找最近邻以及加快基于距离的聚类而采用的 *k*D 树和球形树。我们给出新的属性选择方案（如竞赛搜索和支持向量机的使用），以及新组合模型技术（如累加回归、累加 Logistic 回归、Logistic 模型树以及选择树等），还讨论利用无标签数据提高分类效果的最新进展，包括协同训练（co-training）和 co-EM 方法。

第3版

第3版在第2版基础上进行彻底革新，大量新方法、新算法的引入使本书在内容上与时俱进。我们的基本理念是将本书和 Weka 软件平台更紧密地融合。Weka 现在的版本已经涵盖本书前两部分绝大多数思想的实现，同时你也能通过本书获取关于 Weka 的几乎所有信息。第3版中，我们还添加了大量文献的引用：引用数量达到第1版的3倍多。

Weka 在过去十年中变得焕然一新，也变得易于使用，并且在数据挖掘功能方面有很大提高。它已经集成了无比丰富的机器学习算法和相关技术。Weka 的进步部分得益于数据挖掘领域的近期进展，部分受惠于用户引导以及需求驱动，它使我们对用户的数据挖掘需求了若指掌，充分地借鉴发展中的经验又能很好地选择本书内容。

如前文所述，新版本分为三个部分，其中章节内容有部分调整。更重要的是，增加了很多新内容，以下列举部分重要的改动：

第1章包含了一小节有关 Web 挖掘的内容，并且从道德角度探讨据称是匿名数据中的

个体再识别问题。另外一个重要的补充是关于多实例学习（multi-instance learning），这方面内容出现在两个新增小节中：4.9 节介绍基本方法，6.10 节介绍一些更高级的算法。第 5 章包含有关交互式成本－收益分析（interactive cost-benefit analysis）的新内容。第 6 章也有大量新增内容：成本－复杂度剪枝（cost-complexity pruning）、高级关联规则算法（这种算法利用扩展前缀树将压缩版本的数据集存储到主存）、核岭回归（kernel ridge regression）、随机梯度下降（stochastic gradient descent），以及层次聚类方法（hierarchical clustering method）。旧版中关于输入/输出的章节被分为两章：第 7 章讲述数据转换（主要与输入有关），第 8 章是集成学习（输出）。对于前者，我们增加了偏最小二乘回归（partial least-squares regression）、蓄水池抽样算法（reservoir sampling）、一分类学习（one-class learning）——将多分类问题分解为集成嵌套二分法问题，以及校准类概率（calibrating class probabilities）。对于后者，我们增加了新内容以比较随机方法与装袋算法和旋转森林算法（rotation forest）。而关于数据流学习和 Web 挖掘的内容则增添到第二部分的最后一章。

第三部分主要介绍 Weka 数据挖掘工作平台，也加入大量新内容。Weka 中添入多种新的过滤器、机器学习算法、属性选择算法、如多种文件格式转换器一样的组件以及参数优化算法。实际上，第 3 版中介绍的新版本 Weka 比第 2 版中的 Weka 增加了 50% 的算法。所有这些变化都以文档形式保存。为了满足一些常见要求，我们给出关于不同分类器输出的细节并解释这些输出所揭示的意义。另一个重要的变化是我们新增了一个崭新的章节——第 17 章，在这一章中给出一些关于 Weka Explorer 界面的辅导练习（其中的部分练习颇具难度），这些练习我们建议 Weka 新用户都能尝试着做一遍，这有助于你了解 Weka 究竟能做些什么。

致 谢

Data Mining: Practical Machine Learning Tools and Techniques, Third Edition

书写致谢部分常常都是最美好的时候！许多人给了我们帮助，我们非常享受这个机会来表达谢意。本书源于新西兰怀卡托大学计算机科学系的机器学习研究项目，项目早期科研人员给了我们极大的鼓励与帮助，他们是：John Cleary、Sally Jo Cunningham、Matt Humphrey、Lyn Hunt、Bob McQueen、Lloyd Smith 以及 Tony Smith。特别感谢项目经理 Geoff Holmes 带来了极其丰富的灵感与鼓励，同时还要特别感谢 Bernhard Pfahringer，他们两位在 Weka 软件部分做了重要的工作。机器学习项目所有相关的科研人员都给了我们思考上的帮助，这里特别提到几位学生：Steve Garner、Stuart Inglis 以及 Craig Nevill-Manning，他们帮助我们一起度过了希望渺茫、万事艰难的项目启动初期。

Weka 系统证明了本书的许多想法，Weka 是本书非常重要的部分。该部分的构思由作者完成，设计与实现主要由 Eibe Frank、Mark Hall、Peter Reutemann 以及 Len Trigg 完成，怀卡托大学机器学习实验室的诸多成员都做了很重要的初期工作。相对于本书的第 1 版，Weka 团队有了极大的扩充，做出贡献的成员如此之多因此对每个人都表达充分的感谢不太实际。这里感谢 Remco Bouckaert 提供的 Bayes net 包等一系列贡献，Lin Dong 实现的多实例学习方法，Dale Fletcher 在有关数据库方面提供的帮助，James Foulds 的多实例过滤，Anna Huang 的信息瓶颈聚类，Martin Gütlein 的特征选择，Kathryn Hempstalk 的一分类分类器，Ashraf Kibriya 和 Richard Kirkby 多到难以列举的贡献，Niels Landwehr 的 Logistic 模型树，Chi-Chung Lau 的所有知识流界面图标，Abdelaziz Mahoui 实现的 K *，Stenfan Mutter 的关联规则挖掘，Makcolm Ware 大量各方面的贡献，Haijian Shi 实现的树学习器，Marc Sumner 的快速 Logistic 模型树，Tony Voyle 的最小中值二乘回归，Yong Wang 的 Pace 回归以及 M5′的最初实现，Xin Xu 的多实例学习包 JRip 以及 Logistic 回归等诸多贡献。对所有这些努力工作的人，我们在此一并表示最真诚的感谢，同时也感谢怀卡托大学之外的相关人员对 Weka 部分所做的贡献。

我们隐匿在南半球一个偏远（但十分漂亮）的角落，非常感激那些来我们系的访问学者，他们带给我们非常重要的反馈，帮助我们拓展思路。我们尤其希望提到 Rob Holte、Carl Gutwin 以及 Russell Beale，他们三位的访问都长达数月；David Aha 虽然仅造访了几天，但同样在项目最脆弱的初期阶段给了我们极大的热情与鼓励；Kai Ming Ting 在第 8 章所述的许多主题上与我们有长达两年的合作，他带领我们进入机器学习的主流中。最近还有许多访问学者，包括 Arie BenDavid、Carla Brodley 以及 Stefan Kramer。特别感谢 Albert Bifet 对第 3 版草稿给了我们详细的反馈意见，大部分我们已经采纳并且做了修改。

怀卡托大学的学生对这个项目的开展和推进起到了非常重要的作用，他们当中的许多人已经在上述 Weka 贡献者之列，实际上他们在其他部分同样做了很多工作。早期 Jamie Littin 研究了链波下降规则以及关联学习，Brent Martin 探索了基于实例的学习方法以及基于实例的嵌套表示，Murray Fife 刻苦钻研关联学习，Nadeeka Madapathage 调查了表示机器学习算法的函数式语言的使用。最近，Kathryn Hempstalk 研究了一分类学习方法，她的研究反映在 7.5 节中；同样，Richard Kirkby 关于数据流方面的研究反映在 9.3 节中。第 17 章中的部分练习是 Gabi Schmidberger、Richard Kirkby 以及 Geoff Holmes 设计的。其他研究

生也在很多方面影响了我们，尤其是 Gordon Paynter、YingYing Wen 以及 Zane Bray 三位与我们一起研究了文本挖掘，还有 Quan Sun、Xiaofeng Yu。同事 Steve Jones、Malika Mahoui 一起为本项目及其他机器学习项目做了深入的研究贡献。我们也从许多来自 Freiburg 的访问学生身上学到了很多，这其中就包括 Nils Weidmann。

Ian Witten 希望感谢他之前卡尔加里大学的学生承担的重要角色，尤其是 Brent Krawchuk、Dave Maulsby、Thong Phan 以及 Tanja Mitrovic，这些学生帮助他形成机器学习方面的初期想法，同时还有卡尔加里大学的老师 Bruce MacDonald、Brain Gaines 和 David Hill 以及坎特伯雷大学的老师 John Andreae。

Eibe Frank 感谢他之前在卡尔斯鲁厄大学的主管 Klaus-Peter Huber 对他的影响，让他对所学机器着迷。在他的旅途中，与加拿大的 Peter Turney、Joel Martin、Berry de Bruijn 以及德国的 Luc de Raedt、Christoph Helma、Kristian Kersting、Stefan Kramer、Ulrich Rückert、Ashwin Srinivasn 的交流同样让他获益良多。

Mark Hall 感谢现在就职于密苏里州立大学的前主管 Lloyd Smith 在他论文偏离了原有主题而进入到机器学习领域时仍有着对工作的极大耐心，感谢包括访问学者在内的所有工作人员，尤其感谢多年来怀卡托大学机器学习小组的全体人员极具价值的见解以及鼓舞人心的讨论。

Morgan Kaufmann 出版社的 Rick Adams 以及 David Bevans 非常努力地工作才有了本书的出版，项目经理 Marilyn Rash 让进展变得如此顺利。感谢加州大学欧文分校的机器学习数据库储藏室的图书管理员仔细搜集的数据集，这些数据集对研究工作价值巨大。

我们的研究由新西兰科研、科技、技术基金以及新西兰皇家学会马斯登基金资助。怀卡托大学计算机科学系为我们提供了大量的帮助，同时我们还要特别感谢 Mark Apperley 的英明领导和温暖人心的鼓励。本书第 1 版的部分章节是两位作者在加拿大卡尔加里大学访问时所写，感谢卡尔加里大学计算机科学系所给予的支持，同时还要感谢用本书上机器学习课程的学生虽然辛苦劳累但是依旧保持着积极向上的态度。

最后，最重要的是感谢我们的家人和同事。Pam、Anna 以及 Nikki 对家里有一个作家有何影响了然于心（“没有下次了!”），但依然接受 Ian 在家里任何一个地方写书。Julie 总是非常支持 Eibe，即使在 Eibe 不得不在机器学习实验室挑灯夜读的时候也不例外。Immo 以及 Ollig 让我们愉悦和放松。Bernadette 十分支持 Mark，用尽各种办法让 Charlotte、Luke、Zach 和 Kyle 不那么吵闹，让 Mark 得以集中精力。我们在加拿大、英国、德国、爱尔兰、新西兰以及萨摩亚各地欢庆：新西兰将我们聚在一起，提供了一个充满田园风光几乎完美的地方让我们完成这项工作。

目 录

Data Mining: Practical Machine Learning Tools and Techniques, Third Edition

第二部分 高级数据挖掘