HZ BOOKS
华章科技

智能科学与技术丛书

ELSEVIER

DATA MINING

Practical Machine Learning Tools and Techniques, Fourth Edition

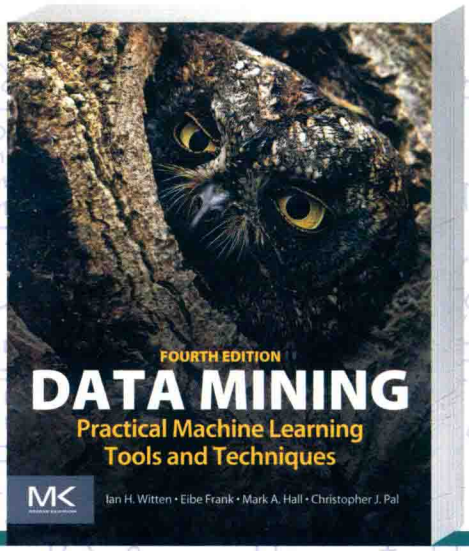
数据挖掘

实用机器学习工具与技术

（原书第4版）

伊恩 H. 威腾（Ian H. Witten）
[新西兰] 埃贝·弗兰克（Eibe Frank）
马克 A. 霍尔（Mark A. Hall）
[加] 克里斯多夫 J. 帕尔（Christopher J. Pal）
著

李川 郭立坤 彭京 蔡国强 任艳 等译



机械工业出版社
China Machine Press

智能科学与技术丛书

DATA MINING

Practical Machine Learning Tools and Techniques, Fourth Edition

数据挖掘

实用机器学习工具与技术

（原书第4版）

伊恩 H. 威腾（Ian H. Witten）
[新西兰] 埃贝·弗兰克（Eibe Frank）
马克 A. 霍尔（Mark A. Hall） 著
[加] 克里斯多夫 J. 帕尔（Christopher J. Pal）

李川 郭立坤 彭京 蔡国强 任艳 等译

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 4 版）/（新西兰）伊恩 H. 威腾（Ian H. Witten）等著；李川等译．—北京：机械工业出版社，2018.1
（智能科学与技术丛书）
书名原文：Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition

ISBN 978-7-111-58916-7

I. 数… II. ① 伊… ② 李… III. 数据采集 IV. TP274

中国版本图书馆 CIP 数据核字（2018）第 004330 号

本书版权登记号：图字 01-2017-0492

ELSEVIER
Elsevier(Singapore) Pte Ltd.
3 Killiney Road, #08-01 Winsland House I, Singapore 239519
Tel: (65) 6349-0200; Fax: (65) 6733-1817

Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition
Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal

ISBN-13: 978-0-12-804291-5

This translation of Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition by Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal was undertaken by China Machine Press and is published by arrangement with Elsevier (Singapore) Pte Ltd.

Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition by Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal 由机械工业出版社进行翻译，并根据机械工业出版社与爱思唯尔（新加坡）私人有限公司的协议约定出版。

《数据挖掘：实用机器学习工具与技术》（原书第 4 版）（李川 郭立坤 彭京 蔡国强 任艳 等译）

ISBN: 978-7-111-58916-7

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：曲 熠　　责任校对：李秋荣
印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2018 年 3 月第 1 版第 1 次印刷
开　　本：185mm×260mm 1/16　　印　　张：27.25
书　　号：ISBN 978-7-111-58916-7　　定　　价：99.00 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

译 者 序

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

随着大数据时代的到来，数据的汇聚、融合、开放、应用在热烈、纷扰、踌躇的节拍中坚定前行。数据挖掘的深度发展为机器学习提供了丰富的实用工具与技术，并在悄无声息中急剧地改变着人们的生活方式。随着 2017 年年初 Master 击败中日韩的超一流围棋选手，大数据分析技术终于突破了所谓的“技术临界点”。科大讯飞的语音精准识别更是打开了数据挖掘在虚拟空间、人机交互、机器人等领域的应用之门。同真实的自然界一样，新兴的“数据自然界”中潜藏着无尽的奥秘和巨大的财富，吸引着大批来自自然科学、人文学科以及商界的精英投身其中。新技术革命时代，正确地解读数据、有效地利用数据，是指引人类前行的灯塔。

本书的几位作者在业内大名鼎鼎，其中 Ian H. Witten 和 Eibe Frank 共同设计了影响深远的 Weka 系统。Weka 的设计集合了前人工作的大成，正如 Google 一样，它也是通过简单思想的迅速实现给所有人带来了前所未有的不同感受。完美的图形界面、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，对于专业人士的探索也能时常予以灵感。而且，Weka 系统为高校的数据挖掘教学提供了实验环境，可谓施惠于众人。

Ian H. Witten 和 Eibe Frank 研发出 Weka 系统后，将他们在开发过程中积累的经验、实际的数据挖掘项目以及教学过程中的体会集结成册，即本书的第 1 版。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 软件日趋成熟。2005 年，本书推出第 2 版。第 2 版最大的变化是增加了专门介绍 Weka 系统的内容。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 系统在过去的十余年里焕然一新，增加了大量的数据挖掘功能，集成了非常丰富的机器学习算法和相关技术。2011 年，本书第 3 版面世，该版介绍了大量新涌现的数据挖掘算法和诸如 Web 数据挖掘等新领域。第 4 版则针对当下数据挖掘的深度发展，着重增加了深度学习的有关内容，详细介绍了概率算法与深度学习的基本理论。

本书的翻译工作主要由四川大学李川副教授负责。郭立坤、彭京、蔡国强和任艳协助进行了后期统稿工作。参与翻译的还有四川大学计算机科学与技术专业的研究生们，他们是冯冰清、刘光明、缪杨帆、蒋志恒、胡代艳、潘科学、张若愚、李晓娟、李茜锦等。他们在紧张的学习之余认真负责地翻译本书，在此对他们表示感谢！此外，也感谢机械工业出版社各位编辑在本书的翻译过程中给予的大力支持。

尽管译者心正意诚，然受限于自身水平，本书的翻译仍有可能存在不足之处，敬请各位读者给予批评、指正，以使本书更趋完善。

李 川

2017 年 11 月

前　言

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

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，即以数据形式存在的状态。假如我们将数据定义为被记录下来的事实，那么“信息”就是隐藏于这些记录事实的数据中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信息释放出来。

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。如果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价值不大的或者没有实际意义的，还有一些可能是虚假的，或者是由于某些具体数据集的巧合而产生的。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，能够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，并可用于多种用途。这是一种抽象化过程：如实地全盘接收现有数据，然后在此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为了发现和描述数据中的结构模式而采用的机器学习工具与技术。

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可谓是铺天盖地。夸张的报道宣称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，其实机器学习绝没有什么魔法，也没有什么隐藏的力量，更没有什么巫术，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。

在许多应用中，机器学习使得从数据样本中获取结构描述成为可能。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，即从数据所描述的过去预测将来在新情况下会发生什么，通常是预测新的样本分类。但也许人们更感兴趣的是，“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。根据经验，在绝大多数数据挖掘实践应用中，用户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。

本书诠释了多种多样的机器学习方法。其中部分出于方便教学的目的而仅仅罗列了一些简单方案，以清楚解释基本思想如何实现。其他则更多考虑到具体实现而列举了很多应用于实际工作中的真实系统。在这些方法中，有很多都是近几年发展起来的。

我们创建了一套综合软件以说明书中的思想。软件名称是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），简称 Weka[⊖]，它的 Java 源代码参见 www.cs.waikato.ac.nz/ml/weka。Weka 几乎可以完整地、产业化地实现本书中所包含的所有技术。它包括了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，它提供了清楚而简洁的实

⊖ 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。

例，以帮助理解机器学习中的相关机理。Weka 还提供了一个工作平台，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。最后，它还包括了一个形如 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价值。如果你希望用自己的数据进行实验，用 Weka 就能轻松做到。但 Weka 绝不是唯一的选择，例如，免费统计计算环境 R 就包含许多机器学习算法。Python 编程语言的爱好者可能更喜欢流行的 scikit-learn 库。用于分布式计算的现代“大数据”框架也支持机器学习，如 Apache Spark。在实际应用中，部署机器学习的选择有很多。本书仅讨论基本的学习算法，没有深入研究特定软件的实现细节，但会在恰当的位置指出所讨论的算法可以在 Weka 软件的什么位置找到。本书还简要介绍了其他机器学习软件，如用于高维数据的“深度学习”。不过，大多数具体软件的信息被归纳到了附录中。

提供数据挖掘案例研究的商业书籍中往往涉及一些非常实用的方法，这些方法与当前机器学习教材中出现的更理论化、更原则化的方法之间存在鸿沟，本书跨越了这个鸿沟。这个鸿沟相当大，为了让机器学习技术应用得到成果，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便期待好结果出现的技术。不同的问题需要用不同的技术解决，但是根据实际问题来选择合适的技术并非易事，你需要知道到底有多少种可能的解决方案。本书所论及的技术范围相当广泛，并不囿于某种特定的商业软件或方案。书中给出了大量实例，但是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（而且真实数据集的获取常受限于商业机密）。本书所选择的数据集并非用来说明那些大型数据中的实际问题，而是要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；本书同样适用于需要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，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、业余爱好者以及学生和教授。他们需要这样一本书：拥有大量实例且简单易读，向读者阐释机器学习相关的主要技术是什么、它们做什么、如何运用它们以及它们是如何工作的。本书面向实际，倾向于告诉读者“如何去做”，同时包括许多算法和伪代码。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隐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。对于本书的大部分内容，我们避免对特定的理论或数学知识做要求。然而，随着其逐渐成熟，我们认识到这门学科的复杂性，所以我们在第 9 章和第 10 章给出了实质性的理论材料，它们是全面理解最近的实践技术尤其是深度学习所必需的。

本书分为几个层次，不管你是想走马观花地浏览一下基本概念，还是想深入详尽地掌握所有技术细节，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。人们常常发现，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。本书共分为两部分，第一部分是关于数据挖掘中机器学习的简单介绍，读者将首先在前三章学习机器学习的基本思想。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，并给出了一些现实中的实际应用。第 2 章和第 3 章给出了不同的输入和输出，或者称之为知识表达（knowledge representation），不同的输出要用到不同的算法。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，便于读者理解。其中的相关原理通过各种具体算法来呈现，但并未包含复杂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，必须对机器学习的效果进行评估。第 5 章可以单独阅读，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问题。

第二部分介绍数据挖掘中机器学习的一些高级技术。在最底层和最详细的层次上，第 6 章和第 7 章详尽地揭示了实现一系列机器学习算法的步骤，以及在实际应用中为了更好地完成数据挖掘任务所必需的、较为复杂的部分（但忽略了某些算法对复杂数学原理的要求）。一些读者也许想忽略这部分的具体内容，但只有到这一层，我们才涉及完整且可运作的机器学习的 Weka 实现方案。第 8 章讨论了一些涉及机器学习输入和输出的实际问题，例如选择属性和离散化属性。第 9 章和第 10 章分别为机器学习和深度学习提供了对概率方法的严谨描述。第 11 章除了介绍有监督学习和无监督学习外，还介绍了半监督学习和多实例学习，而第 12 章主要介绍集成学习技术，这种技术综合了不同学习技术的输出。第 13 章展望未来的发展趋势。

本书描述了在实际机器学习中所使用的大多数方法，但是没有涉及强化学习（reinforcement learning），因为它仅仅是一种优化技术，在实际的数据挖掘中极少应用；也没有包括遗传算法（genetic algorithm）、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。

附录 A 介绍了在第 9 章和第 10 章需要用到的一些数学知识。附录 B 介绍了 Weka 数据挖掘工作平台，该平台给出了第一、二部分中所描述的大部分思想的实现。之所以这样安排，是为了清晰地将概念与实践层面区分开。在第一、二部分，每章的末尾都给出了相关的 Weka 算法。你可以忽略它们或浏览阅读，或者当你急于继续分析数据而不想被算法是如何工作的技术细节所打扰时，选择直接开始 Weka 实践。

更新与修改

我们于 1999 年完成本书的第 1 版，分别于 2005 年和 2011 年完成第 2 版和第 3 版。经过精心修改、润色的第 4 版于 2016 年同读者见面。这个世界在过去 20 年间可谓沧海桑田！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。当然，我们也对第 3 版中出现的错误进行了校正，并将这些错误集中放到公开的勘误文件里。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 版最主要的改变是增加了专门的篇章来介绍 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。在第 1 版中广泛使用和普及的 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，这使得读者用起来更加得心应手。其中最基本的界面是 Explorer 界面，通过这个界面，所

有 Weka 功能都可以通过菜单选择和表单填写的方式完成。另一个界面是 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，你可以使用它对语料库进行设置，使其自动运行已选定的机器学习算法，这些算法都带有不同的参数，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 版还包括一些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，我们增加了一些神经网络方面的内容：感知器和相关的 Winnow 算法、多层感知器和 BP 算法，以及 logistic 回归。我们介绍了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，以及用于回归分析的支持向量机。另外，应读者的要求并考虑到 Weka 新特性的更新，我们还加入了有关贝叶斯网络的新章节，其中介绍了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，文本数据挖掘受到了极大的关注，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。我们还介绍了用于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），以及新型组合模型技术（如累加回归、累加 logistic 回归、logistic 模型树以及选择树等），还讨论了利用无标签数据提高分类效果的最新进展，包括协同训练（cotraining）和 co-EM 方法。

第 3 版

第 3 版在第 2 版的基础上进行了彻底革新，大量新方法、新算法的引入使得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。这一版中 Weka 的版本已经涵盖本书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。在第 3 版中，我们还添加了大量参考文献——引用数量是第 1 版的 3 倍之多。

Weka 变得焕然一新，易于使用，并且在数据挖掘能力上有很大提高。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，它使得我们对用户的数据挖掘需求了如指掌，在充分借鉴发展经验的同时又能很好地选择本书内容。

第 3 版中增加了一些重要的材料，包括 Web 挖掘和对个人如何经常从所谓的匿名数据中“重新识别”的讨论。其他的扩充技术包括多实例学习、互动成本效益分析（cost-benefit analysis）的新材料、成本复杂度（cost-complexity）剪枝、使用扩展前缀树在内存中存储压缩版本的数据集的高级关联规则算法、核岭回归以及随机梯度下降和层次聚类方法。我们增加了新的数据转换：偏最小二乘回归、蓄水池抽样、一分类学习、分解多类分类问题为嵌套二分法的集成以及校准类概率。我们还在集成学习技术中增加了新的信息：随机化与装袋以及旋转森林。此外，还增加了数据流学习和 Web 挖掘的新章节。

第 4 版

编写第 4 版的主要原因是为了增加深度学习方面的综合材料，本质上是由于领域内真正庞大的数据资源（如图片和语音处理）的出现，以及真正庞大的计算资源的可利用性，包括服务器集群和图形处理单元，这些激发了新的发展。然而，深度学习技术是建立在理论和实

践有力结合的基础之上的。而且我们还收到其他请求，要求我们加入更多的、更严谨的、更理论化的材料。

这迫使我们重新思考书中理论的作用。我们深思熟虑后添加了两个新的理论指导章节：第 10 章的深度学习以及第 9 章的概率方法。第 10 章涵盖深度学习本身以及它的前身；第 9 章给出了概率方法原则性的理论发展，这对于了解其他新算法是非常必要的。我们意识到很多读者并不愿意学习这些理论，我们保证本书的其余部分将是简单、易理解的。但是，新增的理论基础对于想快速理解研究界的先进技术的读者而言将是关键的材料。

Weka 的发展非常迅速。它现在提供使用其他语言和系统的方法，例如流行的 R 统计计算语言、Spark 和 Hadoop 分布式计算框架、Python 和 Groovy 脚本语言，以及面向流学习的 MOA 系统等。鉴于在一本纸质书中记录如此全面而快速发展的系统是不可能的或者说是不可取的，为此，我们创建了一系列的在线开放课程，例如用 Weka 进行数据挖掘。更多用 Weka 进行的数据挖掘以及用 Weka 进行的高级数据挖掘见 https://weka.waikato.ac.nz。

第 4 版包含许多其他更新和补充以及更多的参考文献。这里不再一一介绍，你不妨试着进一步阅读。

书写致谢部分常常是最美好的时候！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。我们还极大地受益于与后加入项目团队的 Michal Mayo 和 Robert Durrant 的交流。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 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。相对于本书第 1 版时期，Weka 团队有了极大的扩充，做出贡献的成员非常多，因此对每个人都表达充分的感谢不太实际。这里我们要感谢：Chris Beckham 为 Weka 提供的包，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，Nikhil Kishore 的 elastic net 回归，Niels Landwehr 的 logistic 模型树，Chi-Chung Lau 的所有知识流界面图标，Abdelaziz Mahoui 的 K* 实现，Jonathan Miles 的核滤波实现，Stenfan Mutter 的关联规则挖掘，Malcolm Ware 大量各方面的贡献，Haijian Shi 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，Benjamin Weber 对 Weka 语法解析模块的统一，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。

我们生活在南半球一个偏远（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 在本书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。最近也有许多访问学者，包括 Arie Ben-David、Carla Brodley、Gregory Butler、Stefan Kramer、Johannes Schneider、Jan van Rijn 和 Michalis Vlachos，还有很多学者来到系里为我们作学术报告。特别感谢 Albert Bifet 对第 3 版草稿的详细反馈意见，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。Kathryn Hempstalk 研究了一分类学习方法，Richard Kirkby 研究了数据流。Gabi Schmidberger 研究了密度估计树，Lan Huang 研究了基于概念的文本聚类，Alyona Medelyan 研究了关键短语提取。最近，Felipe Bravo 研究了针对 Twitter 的情感分类，Mi Li 研究了快速聚类方法，Tim Leathart 研究了集成嵌套二分法。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 和 Xiaofeng Yu。同事 Steve Jones 和 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 希望感谢他之前在卡尔加里大学的学生所做的重要工作，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，让他对能学习的机器如此着迷。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 Srinivasan 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机器学习领域时仍有极大耐心，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。

Chris Pal 感谢其他几位作者邀请他撰写第 4 版，感谢家人为了支持他的写作而尽量不占用他的时间。感谢 Polytechique Montréal 提供的休假机会，使得他有时间前往新西兰，同时也感谢怀卡托大学计算机科学系的接待，是他们使得这次合作更加富有成效。还要感谢他的许多导师、学术团队、合著者和同事，多年来，他们的见解一直在影响着他，包括 Brendan Frey、Geoff Hinton、Yoshua Bengio、Sam Roweis、Andrew McCallum 和 Charles Sutton 等。特别感谢 Hugo Larochelle 对深度学习的教学建议。Chris 感谢蒙特利尔研究所的朋友和同事对学习算法的贡献，还要感谢 Theano 开发团队，正在这里学习和已经毕业的学生共同创造了研究机器学习的绝佳氛围。特别感谢 Chris Beckham，他对本版新章节的早期草稿提供了出色的反馈意见。

Morgan Kaufmann 出版公司的 Charlie Kent 以及 Tim Pitts 非常努力地工作才有了本书的出版，项目经理 Nicky Carter 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，他们虽然辛苦劳累但是依旧保持着积极向上的态度。本书第 2 版的部分章节是在莱斯布里奇大学访问期间所写的，感谢加拿大 iCORE 对我们这次前往阿尔伯塔省南部地区访学的支持。

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、Kyle 和 Francesca 不那么吵闹，让 Mark 得以集中精力。我们来自加拿大、英国、德国、爱尔兰、新西兰以及萨摩亚：新西兰将我们聚在一起，感谢这个充满田园风光的完美国度。

目　录

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