HZ BOOKS
华章教育

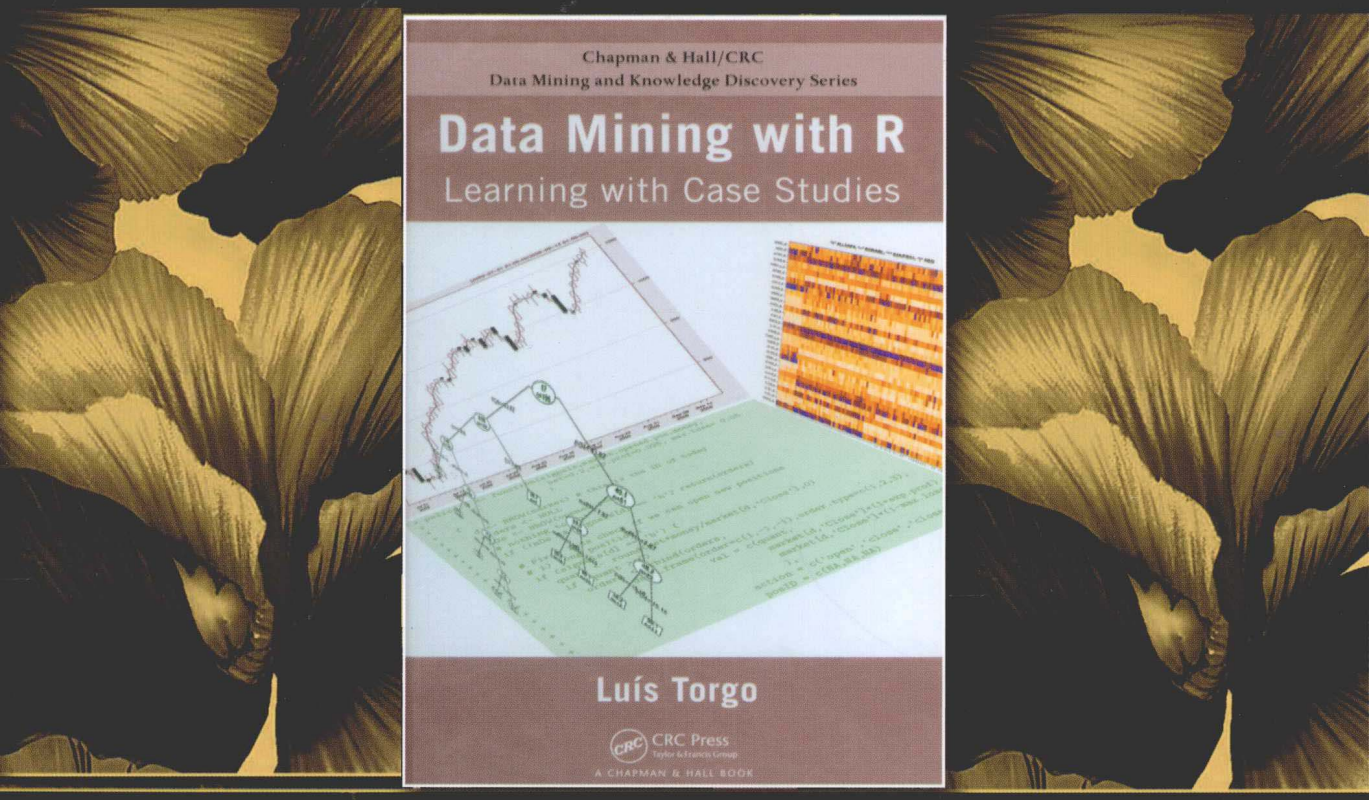
计 算 机 科 学 丛 书

CRC Press
Taylor & Francis Group

数据挖掘与R语言

（葡）Luís Torgo 著
李洪成 陈道轮 吴立明 译

Data Mining with R
Learning with Case Studies



机械工业出版社
China Machine Press

数据挖掘与R语言

Data Mining with R Learning with Case Studies

“如果你想学习如何用一款统计专家和数据挖掘专家所开发的免费软件包，那就选这本书吧。本书包括大量实际案例，它们充分体现了R软件的广度和深度。”

—— **Bernhard Pfahringer**, 新西兰怀卡托大学

本书利用大量给出必要步骤、代码和数据的具体案例，详细描述了数据挖掘的主要过程和技术，广泛涵盖数据大小、数据类型、分析目标、分析工具等方面的各种具有挑战性的问题。

本书的支持网站（http://www.liaad.up.pt/~ltorgo/DataMiningWithR/）给出了案例研究的所有代码、数据集以及R函数包。

本书特色

- 通过仔细选择的案例涵盖了主要的数据挖掘技术。
- 给出的代码和方法可以方便地复制或者改编后应用于自己的问题。
- 不要求读者具有R、数据挖掘或统计技术的基础知识。
- 包含R和MySQL基础知识的简介。
- 提供了对数据挖掘技术的特性、缺点和分析目标的基本理解。

作者简介

Luís Torgo 葡萄牙波尔图大学计算机科学系副教授，现在在LIAAD实验室从事研究工作。他是APPIA会员，同时还是OBEGEF的创办会员。

CRC Press
Taylor & Francis Group

客服热线：(010) 88378991 88361066
购书热线：(010) 68326294 88379649 68995259
投稿热线：(010) 88379604

读者信箱：hzjsj@hzbook.com
华章网站：www.hzbook.com
网上购书：www.china-pub.com

上架指导：计算机/数据挖掘

ISBN 978-7-111-40700-3

9 787111 407003 >

定价：49.00元

计 算 机 科 学 丛

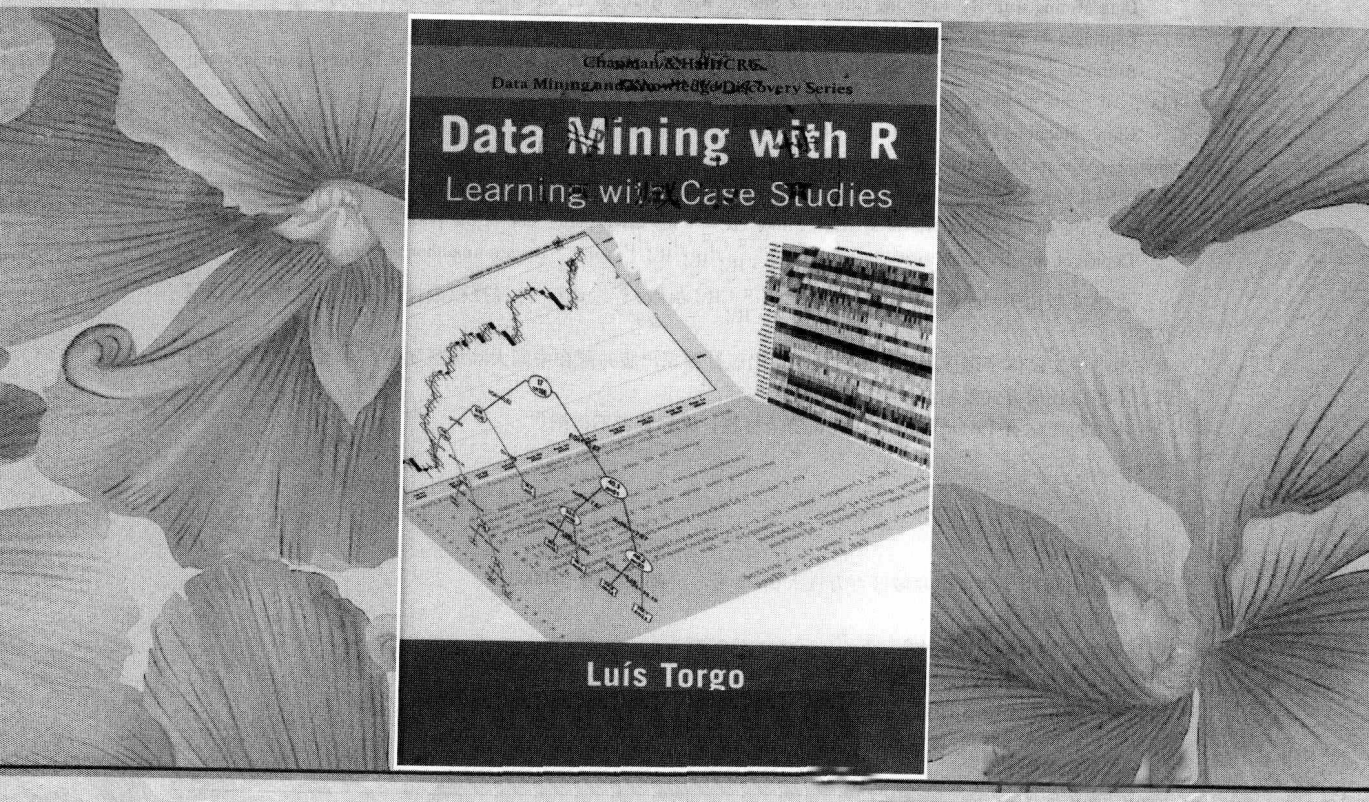
数据挖掘与R语言

（葡）Luís Torgo 著

李洪成 陈道轮 吴立明 译

Data Mining with R

Learning with Case Studies



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据挖掘与R语言/（葡）托尔戈（Torgo，L.）著；李洪成等译. —北京：机械工业出版社，2013.2
（计算机科学丛书）
书名原文：Data Mining with R：Learning with Case Studies

ISBN 978-7-111-40700-3

Ⅰ. 数… Ⅱ. ①托… ②李… Ⅲ. ①数据采集 ②程序语言－程序设计 Ⅳ. ①TP274 ②TP312

中国版本图书馆CIP数据核字（2012）第302931号

本书版权登记号：图字：01-2012-0226

本书首先简要介绍了R软件的基础知识（安装、R数据结构、R编程、R的输入和输出等）。然后通过四个数据挖掘的实际案例（藻类频率的预测、证券趋势预测和交易系统仿真、交易欺诈预测、微阵列数据分类）介绍数据挖掘技术。这四个案例基本覆盖了常见的数据挖掘技术，从无监督的数据挖掘技术、有监督的数据挖掘技术到半监督的数据挖掘技术。全书以实际问题、解决方案和对解决方案的讨论为主线来组织内容，脉络清晰，并且各章自成体系。读者可以从头至尾逐章学习，也可以根据自己的需要进行学习，找到自己实际问题的解决方案。

本书不需要读者具备R和数据挖掘的基础知识。不管是R初学者，还是熟练的R用户都能从书中找到对自己有用的内容。读者既可以把本书作为学习如何应用R的一本优秀教材，也可以作为数据挖掘的工具书。

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
北京瑞德印刷有限公司印刷
2013年4月第1版第1次印刷
185mm×260mm·13.5印张
标准书号：ISBN 978-7-111-40700-3
定 价：49.00元

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

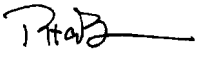
推 荐 序

Data Mining with R: Learning with Case Studies

数据挖掘正在改变着企业和其他大型组织与客户的互动方式，同时也改变着它们管理复杂过程的方式。大量的数据正在很好地用于预测客户行为和结果。在软件方面，R 以其强大的功能和诱人的价格（免费）正在改变着定量分析的“生态系统”。

本书的目的是引领读者迅速地进入这两个世界。本书以实际案例的方式介绍数据挖掘和 R 软件，这样读者就可以在真实情境中进行学习，而不会迷失在统计理论的细节讨论或者计算机科学的基础概念中。本书中用到的工具全部是免费的：MySQL 数据库（用于数据库操作）和 R 软件（用于分析）。因此，本书教给你的是如何动手的知识。通过学习本书，你将体验到数据挖掘和 R 的强大功能。如果你能安装这些工具，并通过应用这些工具来详细地学习书中的案例研究，你将收获颇丰。本书逐步地通过案例研究来介绍 R 的概念，如果你还不熟悉 R 或者 MySQL，你可以按章节顺序来学习这些案例。

本书的原作者 Luís Torgo，根据他在葡萄牙波尔图大学丰富的教学经验、在其他国家讲授数据挖掘课程的经验，以及聚集了世界各地专业人士的 Statistics. com 在线课程中的教学经验，精心地写作了本书。



2012 年 12 月 17 日

Statistics. com 在线课程网站总裁 Peter Bruce

目前，数据挖掘和 R 是学术界及工业界中的两个关键技术。丰富的传感器机制使得自动收集数据成为可能后，产生了非常大的数据集，这需要自动化的机制来将这些数据转化为有用的信息，以供决策者使用和参考。R 是一个开发这些自动化机制的很好选择。R 提供的大量算法和方法，以及它的自由和开放源码特性，使得 R 成为数据挖掘的最佳选择之一。本书的目的是向读者介绍数据挖掘和 R 的知识。本书的写作思路是给读者介绍一系列有代表性的研究案例，通过这些案例，读者不仅从中学到主流的数据挖掘方法，同时也可以学习本书所提供的 R 代码，并最终把这些代码应用到他们自己的数据挖掘项目中。

随着中文版的出版，我希望我能说服更多的人认识 R 和数据挖掘的优势。得知我的书得到世界各地读者的关注，对我而言是一项伟大的荣誉。我相信本书中文版的发行将有助于中国的 R 社区。对所有的中国读者，我真诚地希望，在读完本书后，你们发现它不仅有助于你们的工作，同时你们将和我自己一样增加了对数据挖掘和 R 的热情。

Luís Torgo

2012 年 12 月 16 日于葡萄牙，波尔多

译 者 序

Data Mining with R: Learning with Case Studies

本书是2011年查普曼和霍尔公司（Chapman & Hall/CRC）出版的《Data Mining with R: Learning with Case Studies》一书的中文版。英文版从出版后就在亚马逊美国网站上得到了极高的评价，是2011年亚马逊网站上数据挖掘类书籍销量最好的一本。机械工业出版社以极快的速度引进这本书的中文版，使国内读者在原版出版一年左右的时间里读到本书，不得不赞扬他们独到的眼光。本书翻译完稿的时候（2012年10月），其英文版的销量还是排在专业书籍的前列，原作者为本书维护了一个网站，读者可以访问该网站查看这些信息。

本书的作者Luís Torgo是一位数据挖掘专家，同时也是一位R开发者。本书给出了四个数据挖掘的实际案例，它们分别是藻类频率的预测、证券趋势预测和交易系统仿真、交易欺诈预测，以及微阵列数据分类。这四个案例基本覆盖了常见的数据挖掘技术，从无监督的数据挖掘技术、有监督的数据挖掘技术到半监督的数据挖掘技术。同时这四个案例从数据量、分析目标和数据类型方面引出了各种各样的挑战性问题，本书给出了克服这些挑战的方法和技巧。阅读本书不需要具备R和数据挖掘的基础知识。为了便于读者阅读，本书第1章给出了R软件的基础知识（安装、R数据结构、R编程、R的输入和输出等）。全书以实际问题、解决方案和对解决方案的讨论为主线来组织内容。读者既可以把本书作为学习如何应用R的一本优秀教材，也可以作为数据挖掘的工具书。读者可以根据自己的需要参考书中的某些具体方法，找到自己实际问题的解决方案。

R本身是一款十分优秀的统计分析和数据挖掘软件，有关R的书籍和文档也是相当多的。但是系统地讲解用R进行数据挖掘的书籍目前还没有。本书以四个案例研究的形式组织内容，脉络清晰，并且各章自成体系。读者可以从头逐章学习，也可以根据自己的需要进行学习。不管是R初学者，还是熟练的R用户都能从书中找到对自己有用的内容。

本人在2011年年初学习作者Luís Torgo在Statistics. com上的在线课程，深感本书的内容极具实用价值，萌生了把本书翻译为中文的念头。2011年年末，恰逢机械工业出版社华章公司引进了本书的版权，在王春华编辑的支持下，我承担了本书的翻译工作。由于英文的习惯和汉语有较大的不同，对于一些特别长的句式，译者按照原文的意思进行了分解处理。关于书中的术语，译者尽量采用中文已有的对应术语，如果中文没有对应术语，译者尽力采用贴切的名称来反映原文中的术语。

本书的翻译工作由李洪成、陈道轮和吴立明共同完成。另外，许金玮、朱振兴、陈冰、汤静文、瞿秋霞、张潇予等也对本书的部分翻译提供了帮助。在本书的翻译过程中，原作者Torgo博士多次就译者提出的问题进行耐心而细致的解答。这里对他的帮助表示由衷的谢意。另外，感谢美国统计教育学院Peter Bruce为本书中文版写的推荐序。由于水平所限，书中可能会有翻译不当之处，希望读者多加指正。

李洪成

本书的主要目的是向读者介绍如何用 R 进行数据挖掘。R 是一个可以自由下载[1]的语言，它提供统计计算和绘图环境，其功能和大量的添加包使它成为一款优秀的、多个已有（昂贵）数据挖掘工具的替代软件。

数据挖掘的一个关键问题是数据量。典型的数据挖掘问题包括一个大的数据库，需要从中提取有用的信息。在本书中，我们用 MySQL 作为核心数据库管理系统。对多个计算机平台，MySQL 也是免费的[2]。这意味着，我们可以不用付任何费用就可以进行“重要的”数据挖掘任务。同时，我们希望说明解决方案质量上并没有任何损失。昂贵的工具并不意味着一定更好！只要你愿意花时间来学习如何应用它们，R 和 MySQL 就是一对很难超越的工具。我们认为这是值得的，希望在读完本书之后，你也相信这点。

本书的目的不是介绍数据挖掘的各个方面。许多已有的书籍覆盖了数据挖掘领域。我们用几个案例来向读者介绍 R 的数据挖掘能力。显然，这几个案例不能代表我们在现实世界中碰到的所有数据挖掘问题。同时，我们给出的解决方案也不是最完全的方案。我们的目的是通过这些实际案例向读者介绍如何用 R 进行数据挖掘。因此，我们案例分析的目的是展示用 R 进行信息提取的例子，而不是提供数据挖掘案例的完整分析报告。它们可以作为任何数据挖掘项目的可能思路，或者作为开发数据挖掘项目解决方案的基础。尽管如此，我们尽力尝试覆盖多方面的问题，展示数据大小、不同数据类型、分析目标和进行分析所必需的工具所带来的挑战。然而，这里的实践方式也是有代价的。实际上，作为具体案例研究的一种形式，为了让读者在自己的计算机上执行我们所描述的步骤，我们也做了某些妥协。也就是说，我们不能处理太大的问题，这些问题要求的计算机资源不是每个人都具备的。尽管这样，我们认为本书涵盖的问题也不算小，并对不同的数据类型和维度给出了解决方案。

这里并不要求读者具有 R 的先验知识。没有学过 R 和数据挖掘的读者应该可以学习书中的案例。书中的各个案例相互独立，读者可以从书中任何一个案例开始。在第一个简单案例中，给出了一些基本的 R 知识。这意味着，如果你没有学过 R，至少应该从第一个案例开始学习。而且，第 1 章给出了 R 和 MySQL 的简介，它可以帮助你理解后面的章节。我们也没有假设你熟悉数据挖掘和统计技术。在每个案例的必要地方，都对不同的数据挖掘技术进行了介绍。本书的目的不是向读者介绍这些技术的理论细节和全面知识，我们对这些工具的描述包括了它们的基本性质、缺点和分析目标。如果需要进一步了解技术细节，可以参考其他书籍。在某些节的末尾，我们提供了“参考资料”，如果需要，可以参考它们。总之，本书的读者应该是数据分析工具的用户，而不是研究人员或者开发人员。同时，我们希望后者把本书作为进入 R 和数据挖掘“世界”的一种方式，从而发现本书的用途。

---

[1] 下载网址：http://www.r-project.org。

[2] 下载网址：http://www.mysql.com。

本书有一个免费的 R 代码集，可以从本书网站下载[⊖]。其中含有案例研究中的所有代码，这可以帮助你的实践学习。我们强烈建议读者在阅读本书时安装 R 并实验书中的代码。而且，我们创建了一个名为 DMwR 的 R 添加包，它包含本书用到的多个函数和以 R 格式保存的案例数据集。你应该按照本书的指示，安装并加载该添加包（第 1 章给出了细节）。

⊖ 下载网址：http://www.liaad.up.pt/~ltorgo/DataMiningWithR/。