数据分析与决策技术丛书

Machine Learning with R Cookbook

机器学习与R语言实战

丘祐玮（Yu-Wei Chiu）著

潘怡 叶晖 译

涵盖100多种分析数据和构建预测模型的实用方法

提供简单易实现的R源码

机械工业出版社
China Machine Press

数据分析与决策
技术丛书

Machine Learning with R Cookbook

机器学习与R语言实战

丘祐玮（Yu-Wei Chiu）著

潘怡 叶晖 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

机器学习与 R 语言实战 / 丘祐玮著 . —北京：机械工业出版社，2016.4
（数据分析与决策技术丛书）

ISBN 978-7-111-53595-9

I. 机… II. 丘… III. ① 机器学习 ② 程序语言 - 程序设计 IV. ① TP181 ② TP312

中国版本图书馆 CIP 数据核字（2016）第 085266 号

本书版权登记号：图字：01-2015-7672

机器学习与 R 语言实战

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：佘 洁　　责任校对：殷 虹

印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2016 年 5 月第 1 版第 1 次印刷

开　　本：186mm × 240mm 1/16　　印　　张：22

书　　号：ISBN 978-7-111-53595-9　　定　　价：69.00 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）88379426 88361066　　投稿热线：（010）88379604

购书热线：（010）68326294 88379649 68995259　　读者信箱：hzit@hzbook.com

The Translators' Words 译 者 序

数据的采集、聚集以及可视化仅仅是数据分析整体工程的一部分，要从海量数据中抽取出有价值的信息是目前大数据应用领域一项新的并且有挑战性的工作。作为大数据的技术基石，机器学习这一新兴学科虽然已经被越来越多的人们所认识，但由于学科自身的交叉性，许多算法往往让人觉得复杂和难以理解。本书作者作为一名资深的数据科学家，借助当前机器学习和数据分析领域最常用的工具R语言，分享了其在数据分析领域实践机器学习算法的诸多心得。

本书内容全面，深入浅出地介绍了采用R语言实现包括分类、回归、聚类、关联分析等常用的机器学习算法的知识，每一个算法都通过案例详细说明了构建模型、实现模型以及评价模型的过程。同时，为了照顾初学者，本书也涵盖了R语言的基础知识，包括环境准备、数据转换、分析和结果可视化的方法。本书最后抛砖引玉，展示了使用RHadoop处理和分析海量数据的过程。

阅读完本书并亲自动手完成作者所有算法案例后，您将对机器学习和R语言都有更深入的了解，设计学习算法来发现隐藏在数据中有价值的模式也不再是遥不可及的目标。

本书能够得以出版，要感谢机械工业出版社的缪杰、佘洁编辑，他在翻译过程中给予了我们很多建设性的指导意见。其次，还要感谢吴怡编辑，是她让我们与机械工业出版社结缘。

由于教学科研需要，译者很早就已经接触了机器学习这一领域，但由于学科发展速度日新月异，在翻译过程中我们仍然遇到了一些问题，尽管我们在此期间查阅了大量的文献及网络资源，并逐字逐句地对译稿进行了反复推敲和琢磨，还是不可避免地存在错误和疏漏之处，还望各位读者不吝指正。

前　言 Preface

如今，大数据在诸多领域已经成为一个时髦的热门词汇，越来越多的人开始接触并考虑引入这一技术以促进公司产品的销售获得更多利润。然而，数据的采集、聚集以及可视化仅仅是数据分析整体工程的一部分，要从数据中抽取出有价值的信息才是一项有挑战性的新工作。

大多数研究人员习惯依据历史样本数据进行统计分析，这种处理方法的弊端在于从统计分析中能够获得的信息十分有限。事实上，科学家们经常要解决从目标数据中发现被隐藏的模式以及探索未知关系的问题。目前，机器学习已经逐渐成为除统计分析以外的一种新的分析方法，它使用学习算法，结合输入的样本数据，能够得到更加精确的预测模型。通过机器学习，商业操作及其发展趋势的分析不再局限于人脑层面的思考，机器层面的分析使企业能够在大数据中发现潜在价值。

R 语言是目前机器学习和数据分析领域最常用的工具，开源和免费的优势使得它成为最受数据科学家们欢迎的主流语言。R 语言为用户提供了丰富的学习包和可视化函数，用户不需要掌握任何分析过程背后数学模型的细节就能很简单地通过 R 语言在数据集上执行机器学习算法，快捷地完成数据分析任务。

本书采取了务实的方法介绍如何使用 R 语言来实践机器学习。全书共 12 章，每章包含若干小节，当读者循序渐进地学习完每一小节后，将能够使用数目繁多的机器学习包构建自己的预测模型。

本书首先引导读者学会搭建一个 R 语言环境并使用简单的 R 命令来观察数据。接下来读者将学习利用机器学习算法进行统计分析并评价生成模型，以及如何使 R 语言与 Hadoop 结合以构建大型数据分析平台。本书所涉及的全部机器学习案例都附带了详细的说明。

我们相信，读完这本书你将发现机器学习从来没有这样容易。

章节内容

第 1 章介绍了如何创建一个可用的 R 环境和基本的 R 命令，包括数据读取、数据操纵、简单的统计分析以及数据的可视化。

第 2 章介绍了如何使用 R 语言进行探索性数据分析，以 Titanic 数据为例，探讨了数据的转换、分析以及结果的可视化。我们建立了一个预测模型，来判断泰坦尼克号可能的幸存者。

第 3 章首先重点探讨了数据采样和概率分布的概念，然后演示了对数据进行统计描述和统计推断性统计的过程。

第 4 章探讨一个因变量（响应变量）和一组或多组独立的（预测量）解释变量之间的线性关系。读者将学习使用各类回归模型来解释数值间的关联，同时还将学习运用合适的模型对连续变量进行预测。

第 5 章介绍基于树的分类器、k 近邻分类器、逻辑回归分类器以及朴素贝叶斯分类器。为了帮助读者们能够更好地理解分类器的工作方式；这一章提供了一个基于电信数据集的用户分类实例。

第 6 章介绍了两种复杂但功能强大的分类算法：神经网络和支持向量机。尽管这些方法从根本而言难度都较大，但通过这一章的学习，读者会发现在 R 语言里使用这些算法做出精确的预测是一件非常容易的事情。

第 7 章展示一些评估模型性能的方法，通过这些检验方法，我们能够从中挑选出最优化的模型应用于预测。

第 8 章探讨集成分类器，相对于单一分类器，集成分类器在分类和回归处理方面具有更多优势。而鉴于其在很多数据预测比赛中的良好表现，读者更应该了解在项目中如何使用集成分类器。

第 9 章讨论多种聚类算法。通过聚类，我们能够发现对象间的共性，该章使用聚类算法对顾客进行划分，同时比较了不同聚类算法之间的差异。

第 10 章讨论了如何发现事务数据中所隐含的常见模式和关联项。

第 11 章介绍如何从原始变量中选择和抽取特征。借助降维，我们能够消除冗余特征对分析结果的影响，并降低计算的代价以避免模型的过度适应。该章将借助一个具体的图像压缩和存储案例解释降维方法。

第 12 章介绍 RHadoop 处理和海量数据分析，以及如何使用 RHadoop。该章依次介绍了 RHadoop 环境的构建，使用机器学习方法处理实际的海量数据集，最后该章探讨了使用亚马逊弹性计算云（Amazon EC2）服务来部署 RHadoop 集群。

附录 A 提供 R 和与机器学习相关的所有资源。

附录 B 提供泰坦尼克号幸存者的数据集。

学习指南

如果希望实践本书中的案例，你需要一台安装了 R 语言包并且能够访问 Internet 的计算机。读者可以从 http://www.cran.r-project.org/ 下载安装程序，详细的安装说明可以在本书第 1 章中找到。

本书所提供的全部示例程序都已经在 R 3.1.2 版本 +Windows 环境下测试成功，这些示例也同样适用于安装在 Mac OS X 以及类 UNIX OS 系统上的最新版本的 R 语言包。

本书面向的读者

本书适合那些希望了解并掌握 R 语言实践机器学习完成数据观察的读者，我们在书中介绍了 R 语言的基础知识，那些具备基本编程能力或了解机器学习算法的读者们能够在学习本书后有所收获，但如果读者没有任何 R 语言的基础也没有关系。

About the Author 作者简介

Yu-Wei，Chiu (David Chiu) 是 LargitData 公司（www.LargitData.com）的创始人。David 曾是 Trend Micro 公司的软件工程师，负责构建商务智能大数据平台以及客户关系管理系统。除了是一名创业者和数据科学家之外，David 还专注于利用 Spark 和 Hadoop 来处理海量数据，并使用数据挖掘技术来进行数据分析。他还是一名专业的讲师，在很多会议上做过关于 Python、R 以及 Hadoop 方面的技术报告。

2013 年，Yu-Wei 审读了《Bioinformatics with R Cookbook》(Packt 出版社)。更多内容请参考他的个人网站 www.ywchiu.com。

我要衷心感谢我的家人和朋友，是他们支持和鼓励我完成了本书。我要诚挚地向我母亲 Ming-Yang Huang（Miranda Huang）、我的良师 Man-Kwan Shan、本书的校对 Brendan Fisher，中国台湾的 R 用户组，数据科学项目（Data Science Program，DSP），以及其他支持过我的朋友表示感谢。

审校者简介 *About the Reviewers*

Tarek Amr 目前在荷兰工作的数据科学家，于东安格利亚大学获得知识发现和数据挖掘硕士学位，开放知识基金会和数据学院的志愿者，负责与开放数据相关的项目，以及数据新闻和数据可视化领域的培训工作。Tarek 还是另外一本书《Python Data Visualization Cookbook》（Packt 出版社）的评审人，目前正致力于撰写一本有关使用 D3.js 实现数据可视化的书籍。有关他的更多信息请参考：http://tarekamr.appspot.com/。

Abir Datta Cognizant Technology Solutions 公司的数据科学家，专注于保险、金融服务和数字化纵向分析。Abir 主要负责分析、预测建模，为不同行业用户提供商务智能 / 分析领域端到端的海量数据集成解决方案，从而为用户解决商务分析问题。Abir 也开发了一些算法来识别顾客潜在的特征以形成战略决策通道，从而获得更大的商业成功。

Abir 对风险模型也有所研究，是当前他所服务的公司内负责开发风险控制平台小组的一员，该平台已经被众多银行和金融服务机构认可。

Saibal Dutta 目前在印度卡哈拉格普尔的印度理工学院从事数据挖掘及机器学习领域的研究工作。他同时还拥有印度奥里萨邦国家技术研究所电子与通信工程硕士学位。Saibal 担任了 HCL 有限公司和诺基亚公司的软件开发顾问。在长达 4 年的顾问工作中，他与宜家（瑞典）、培生（美国）等国际大公司都有过合作，而他对创业的热情也引导他在数据分析领域创办了属于自己的企业，目前该企业正处于 bootstraping 阶段。Saibal 熟悉数据挖掘、机器学习、图像处理和商务咨询。

Ratanlal Mahanta 拥有计算金融硕士学位，目前在 GPSK 投资集团担任高级量化策略分析师。拥有 4 年为投资银行及风险管理公司提供量化交易及战略研究的经验。他也是高频及算法交易方面的专家，拥有以下领域的从业经验：

- 量化交易：FX、股票、期货、买卖以及金融衍生品技术。
- 算法：偏微分方程、随机微分方程、有限差分法、蒙特卡罗算法以及机器学习。
- 编码：R 编程、C++、MATLAB、HPC 以及科学计算。
- 数据分析：海量数据分析 (EOD 到 TBT)、Bloomberg、Quandl 以及 Quantopian。
- 策略研究：Vol 套利、常规及奇异期权操作建模、趋势跟踪、均值回归、协整、蒙特卡罗仿真、风险价值、压力测试、高夏普率买方交易战略、信用风险建模以及信用

评级。

Ricky Shi　量化交易员和研究者，专注于大规模机器学习以及稳健预测技术。他拥有机器学习及海量数据挖掘方面的博士学位。目前，Ricky 正负责一项应用数学方面的研究，希望将学术研究成果推广至现实领域。他与众多研究机构和公司都有合作，包括雅虎实验室、AT&T 实验室、Eagle Seven、摩根斯坦利股权交易实验室（ETL），以及由 Philip S. Yu 教授领导的 Engineers Gate Manager LP。

他的研究内容包括：

- 异构数据相关性分析，例如从用户的人口统计特征和用户社交网络进行社交广告分析。
- 时序对象关联分析，例如动态相关性分析，寻找当前最有影响力的金融产品（震荡检验、叠加图），并将其用于套期保值和投资组合管理中。
- 学习任务关联分析，例如传递学习。

Jithin S.L　于洛约拉科技学院获得信息技术学士学位，从分析领域起步到各应用领域大数据分析，Jithin 与许多知名的机构都合作过，包括汤森路透、IBM、Flytxt 等，完成了不同的任务，涉足领域包括银行、能源、医疗健康以及通信等，并解决过全球性大数据应用项目。

Jithin 在许多国内外会议都发表过有关技术和商务方面的研究论文。

他的人生格言是学习是永无止境的过程，它对我们理解、抽象现实世界并为这个世界带来新生事物都有帮助。

目　录 *Contents*

第 1 章 Chapter 1

基于 R 实践机器学习

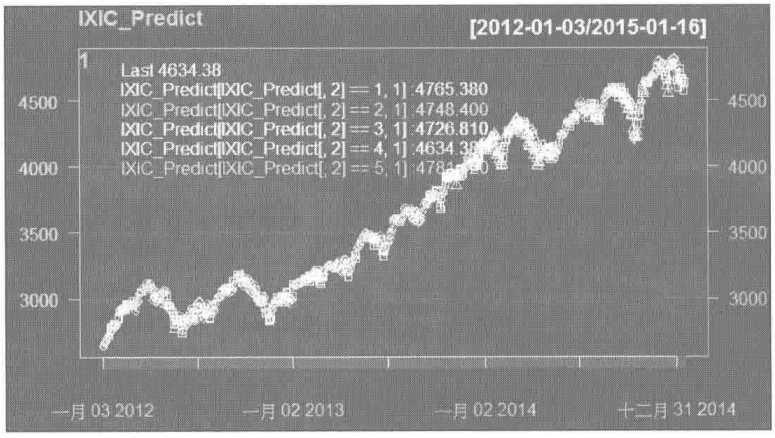
1.1 简介

机器学习的主要目标包括发现隐藏在数据中的模式、未知关联以及有价值的信息。除此之外，机器学习结合数据分析技术也可以应用于预测分析。有了机器学习，对商业活动的分析和处理就不再局限于人工处理，而是可以借助机器的分析发现海量商业数据中所隐藏的价值。

机器学习和人类思维模式有共通之处，传统数据分析方法无法应对由于数据累积更新而对分析模型带来的影响，而机器学习可以不断地从正在被处理和分析的数据中获得信息，也就是说，算法处理的数据越多，其建模能力就越强。

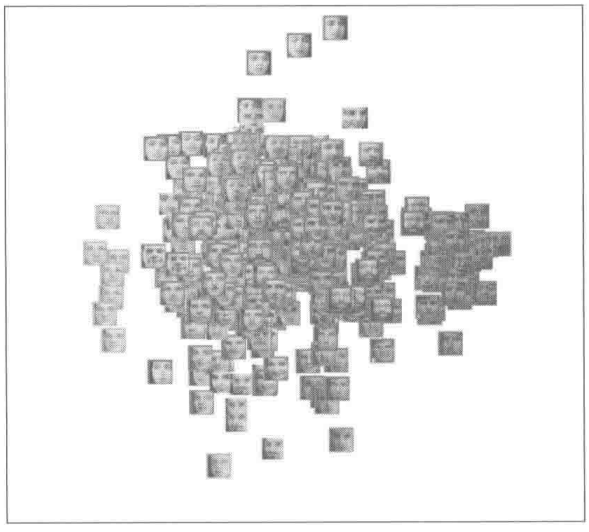
作为 GNU-S 语言的一个分支，R 是一种功能强大的统计语言，被广泛应用于数据的处理和分析。另外，R 还提供了很多有关机器学习的包和数据可视化的函数，使得用户能够简单快速地完成数据分析。当然，最重要的是，R 还是一个免费的开源工具。

R 在很大程度上降低了实践机器学习的复杂度，我们仅需要了解哪一个算法可以解决问题，利用已经写好的包和简单的几行命令，就能针对数据构建相应的预测模型。例如，我们既可以利用朴素贝叶斯方法来进行广告垃圾邮件的筛选，也可以基于 k 均值算法来对顾客类别进行划分，还可以借助线性回归模型来预测未来的房价，或者就像下面这个图一样，通过隐马尔可夫模型来预测未来股票市场。



使用 R 预测股票涨跌

更进一步地，我们还可以利用非线性降维来计算图像数据之间的相异性，并如下图所示那样，通过图形展示聚类结果。具体的操作会在书中接下来的章节中提及。



人脸图像聚类结果可视化

本章将从整体上对机器学习及 R 语言进行一个概要介绍，第一小节包括如何搭建 R 及其集成开发环境 RStudio，配置环境后，接下来的一小节说明安装和导入 R 的算法包。为了更好地了解如何使用 R 来完成数据的分析，后面的 4 小节将探讨包括数据的读写、数据操作、基本统计方法以及数据的可视化。本章最后一节将列出有用的数据来源和其他资源清单。

1.2　下载和安装 R

要使用 R，当然需要首先在机器上安装 R。本节将详细介绍下载和安装 R 的过程。

1. 准备

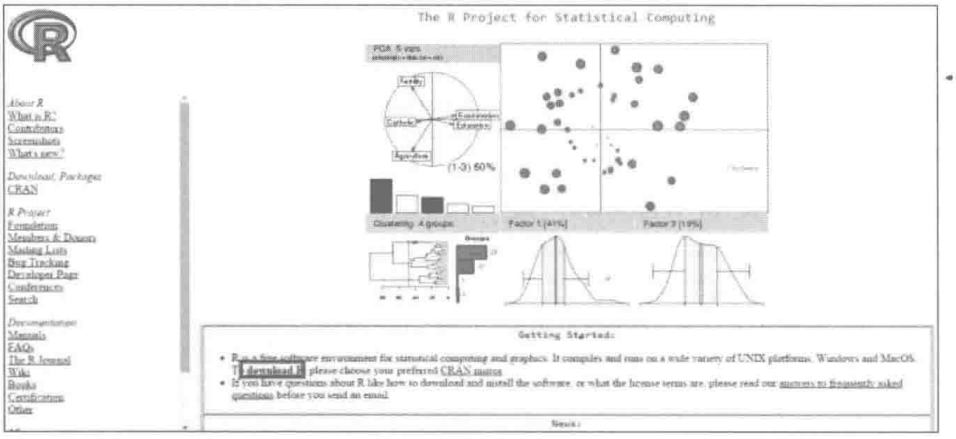
如果读者是 R 初学者，可以在 R 的官方网站（http://www.r-project.org/）找到详细的介绍，包括 R 语言的发展历史以及它的功能。如果已经准备好下载和安装 R，可以访问以下链接：

```
http://cran.r-project.org/
```

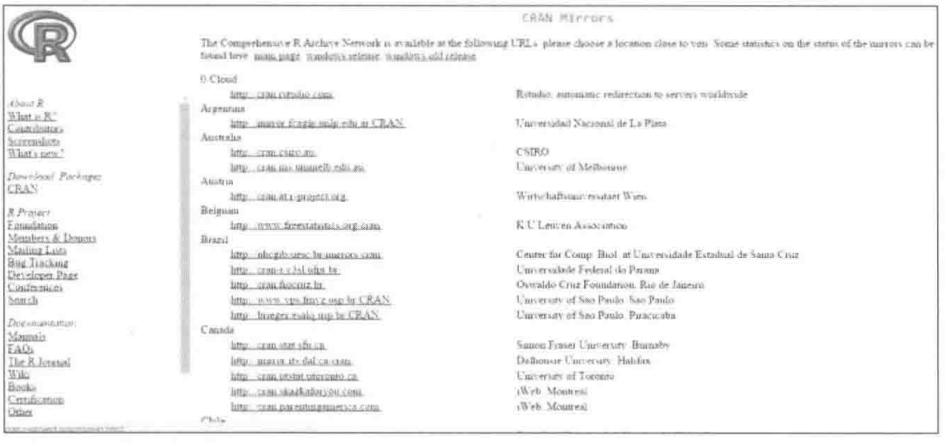
2. 操作

执行以下操作，在 Windows 或 Mac 环境下完成 R 的下载及安装工作：

1）访问 R CRAN 网站（http://www.r-project.org/），单击 download R 链接，指向 http://cran.r-project.org/mirrors.html）：



2）选择离自己最近的镜像网站：



CRAN 的镜像网站