高级大数据人才培养丛书

大数据时代，R语言让你从数据中发掘价值

R语言与大数据编程实战

李倩星 ◎ 编著

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

高级大数据人才培养丛书

R语言与大数据编程实战

李倩星　编著

電子工業出版社
Publishing House of Electronics Industry
北京·BEIJING

内容简介

本书是一本R语言入门读物，它旨在帮助读者迅速构建起与数据分析相关的知识体系，并学习如何使用R软件实现数据分析方法。无论有无深厚的编程基础或数学基础，本书都能帮助读者成长为一名合格的数据分析师。

本书全面介绍了来自统计分析、机器学习、人工智能等领域的多种数据分析算法，在讲解与之相关的R代码时，还讨论了这些算法的原理、优缺点与适用背景。本书按照由易到难的原则组织章节主题，读者将获得最好的阅读体验。通过阅读本书，读者将对R语言在数据分析领域的应用有一个全面的认识。这种认识不被特定行业所局限，任何行业的读者都能利用本书介绍的数据分析方法解决本行业的数据分析问题。

图书在版编目（CIP）数据

R语言与大数据编程实战 / 李倩星编著. —北京：电子工业出版社，2017.9
（高级大数据人才培养丛书）

ISBN 978-7-121-32634-9

Ⅰ. ①R… Ⅱ. ①李… Ⅲ. ①程序语言－程序设计Ⅳ. ①TP312

中国版本图书馆CIP数据核字（2017）第215708号

策划编辑：李　冰
责任编辑：李　冰
特约编辑：彭　瑛　赵海军等
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱　　邮编：100036
开　　本：787×1092　1/16　　印张：20　　字数：512千字
版　　次：2017年9月第1版
印　　次：2017年9月第1次印刷
定　　价：59.00元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

本书咨询联系方式：libing@phei.com.cn。

前　言

R 语言是如今最热门的编程语言之一，它由统计学家开发，在解决数据分析问题时具有先天优势。R 语言是一门新兴的语言，掌握它，就是掌握了一门高效的数据分析软件。随着大数据概念的普及，R 语言能够实现的功能越来越丰富，越来越多的数据分析从业人员产生了学习 R 语言的需求。本书迎合时代潮流，讲解了大数据时代下 R 语言渗透最广泛的几个领域，全面介绍了如何使用 R 语言完成数据挖掘工作。对 R 语言编程人员来说，本书是一本不可或缺的工具书。

本书特色

1. 通俗易懂、实用性强，适合各层次读者学习

本书对读者的数学基础或编程基础不做任何要求。在讲解知识点时，本书采用了平实的语言，对每个疑难点都加以详细解释。此外，本书以实用为主旨，秉承“看得懂、学得会、用得上”的编写原则，精心选取了流行于行业前沿的 18 个主题，不仅通俗易懂，还确保读者所学的知识具有实际应用价值。通过阅读本书，读者都能迅速掌握 R 语言的编程技巧及相关的数据分析知识，并在实际工作中立刻应用它们。

2. 条理清晰、结构巧妙，全面盘点数据分析常用算法

数据分析是一个涉及多领域的交叉学科，R 软件的触角同样也能伸展到多个领域。本书选取了统计分析、机器学习、人工智能等多个学科的流行算法作为主题，讲解了如何使用 R 语言实现它们。这些算法有些偏重数学思维，有些偏重编程技巧，本书主要遵循由易到难的顺序排列主题，并尽量把起源于同一学科的算法放在一起。读者可以按照顺序阅读本书，也可以优先选择感兴趣的部分。此外，本书还穿插介绍了与 R 软件相关的一些其他编程主题，这些主题共同形成知识网络，帮助读者迅速成长为能够独当一面的数据科学家。

3. 知识点丰富、可拓展性强，满足读者的多重需求

本书涉及多个学科，全面介绍了 R 软件能够实现的多种算法，满足了读者的三大需求：首先，使用通俗易懂的语言介绍 R 软件，帮助读者实现零基础入门；其次，囊括多种数据分析算法，带领读者全面认识 R 软件的强大之处，帮助读者成长为合格的数据科学家；最后，本书具备较强的可拓展性，从事任何行业的读者都能够从本书中获取适合其行业的知识。本书还给出了 R 语言进阶的线索，无论想向哪一方面进阶，本书都能为读者打造最坚实的基础。

本书内容及体系结构

本书总共 18 章，分别为 R 的基本介绍、原始数据的探索与预处理、R 的数据可视化、R 中参数的估计和检验、R 中的方差分析、R 中的相关分析和回归分析、更高级的数据可视化、R 中的聚类分析和判别分析、R 中的主成分分析和因子分析、R 中的广义线性回归模型、R 中的时间序列模型、R 中的最优化问题、使用 R 绘制地理信息图形、使用 R 构建支持向量机、实现更高效的流程控制和高级循环、R 代码的调试与优化、构建电影评分预测模型、贝叶斯垃圾邮件过滤器模型。这 18 章进一步又分为 5 个部分。

第一部分为本书的第 1~6 章。其中前 3 章展示了 R 软件的一些入门功能，如数据预处理和数据可视化等，后 3 章则介绍了三种基础的统计分析方法，即参数的估计和检验、方差分析、相关分析、回归分析。这 6 章围绕初级的统计方法展开，是数据分析师必备的基本知识。

第二部分为本书的第 7~11 章，这 5 章介绍了更高级的统计方法。其中，第 7 章为第 3 章的延伸，介绍了数据可视化的高级方法，第 8~11 章则介绍了 6 种高级统计分析方法，这部分的内容与第一部分互为补充。

第三部分为本书的第 12~14 章，这部分内容围绕机器学习展开。第 12 章的主题为最优化，是机器学习的基本理论。第 13 章介绍了如何使用矢量化的思想绘制地图。第 14 章则介绍了支持向量机，它是最典型的机器学习算法之一。这部分讲解了更高深的 R 语言编程技巧，讨论了一些 R 软件能够解决的高难度问题。

第 15、16 章可视为本书的第四部分。这两章围绕如何优化 R 代码展开，系统地讨论了如何写出错误较少的、运行速度较快的代码。这部分内容帮助读者建立良好的编程习惯，以及与其他 R 用户更好地协同工作。

第 17、18 章则为本书的最后一部分，这两章分别讨论了一个完整的数据挖掘项目。其中电影评分预测的案例着重于表现数据挖掘的完整流程，包括繁复的数据预处理与反复的模型比较等工作；垃圾邮件过滤的案例则引出 R 软件能够处理的另一个主题——文本分析。

上述划分方法仅为一个参考，本书的 18 章既互相联系又彼此独立，读者可按照上述划分方法阅读本书，也可优先阅读某些章节，如将第 3、7、13 章等与数据可视化相关的三个章节放在一起阅读。

本书读者对象

- 想要了解 R 语言的数据分析从业人员。
- 统计学、金融学、计算机技术与科学等专业的学生。
- 想要提高 R 语言编程能力的数据分析师。
- 希望系统学习统计分析方法的从业人员。
- 其他对 R 语言有兴趣的各类人员。

目　录

第1章　R的基本介绍

作为一门新兴的编程语言，R是如今值得学习的语言。由统计学家开发出的R语言具有许多奇特性质，本章将较为全面地介绍R的特性和用途，并讲解R的安装方法、变量类型、从其他数据源读取数据、程序包等基本知识。本章帮助读者对R形成整体印象，同时本章内容也是后续章节的基石。

1.1　强大的R

R语言脱胎于S语言，是一门专门用于处理数据探索、统计分析等任务的编程语言。它由统计学家开发完成，在数据分析方面具有天然的优势，运行R程序的R软件是如今最流行的统计软件之一。

与其他统计软件相比，R软件最特别的地方在于它是开源的。这同时意味着：第一，R是免费的；第二，R的用户能够自由地参与到R的开发中。R社区将它的忠实用户聚合在一起，这些用户主要由统计学家、计算机学家、数据分析师等组成，不同领域的用户在R社区中交流碰撞，协助R核心团队丰富和完善R的功能。

R的用户之间具有非常紧密的联系，他们最大的贡献是创建了形形色色的程序包，这些程序包分别封装了一些具有特定作用的函数。如今，R软件已经内置了非常丰富的各类函数库，能够满足绝大多数统计人员的各类需求，它的制图功能也远超其他统计软件。

R的另一个特点在于它支持混合型的编程范式。R是一种解释型的语言，当用户在R软件中编写好一条代码后，R会立即执行它。这种做法的好处在于用户可以即时地看到程序的返回结果，在作图时尤其方便。R是一种面向对象的语言，同时它也支持函数式编程，即用户可以在R中调用现成的或自己编写的函数，这一点与C语言较为相似，但R要比C语言更加灵活。

尽管R的优点很突出，但它也同样具有局限之处。首先，R语言的编程原理较为传统，在处理数据时，R需要将数据全部载入内存，这一点极大地影响了R的运行效率，尽管如今的计算机内存做得越来越大，但在有些大规模数据集的处理工作中，R还是会显得不够得力。其次，R软件的保密性不如SAS等统计软件好，这限制了R在大型商业项目中的应用。最后，由于R软件是由统计学家开发的，因此其语法设计并不特别严谨，有时它会出现一些奇怪的错误。

随着大数据时代的到来，R语言正被越来越多的人关注，不仅是统计分析和数据挖掘，一些研究机器学习和模式识别的专家同样关注到R的发展。根据TIOBE提供的编程语言排行榜，R语言的流行程度在近几年内已经飙升至前十名，其火爆程度只有Python才能与其比肩，而同为统计软件的SAS和MATLAB则一直徘徊于二三十名的位置。

R 的优点使它广泛地流行于统计人员和中小型商业公司中。Google、百度等互联网巨头则将 R 语言看作一个沙盘，使用软件验证各种数据模型的可行性，并最终使用其他语言实现。随着 R 的用户越来越多样化，其可扩展能力进一步强化，能够解决的问题也越来越丰富。如今，金融、医药、教育、社会科学等每一个需要数据分析的领域都需要精通 R 的人才。

1.2 R 语言在大数据中的应用

R 语言的起源是统计学家为解决数据分析领域问题而开发的语言，所以和 MATLAB、Python 等可用于数据处理的语言相比，在数据分析处理方面具有一些独特优势，本节将讲述 R 语言在大数据领域中的典型应用。

1.2.1 R 语言用户行为分析

近几年，淘宝、京东等几家电商的价格战打得不亦乐乎，而从电商发布的战果来看，几败具伤的价格战已经无法保证电商的利润，他们开始转向利用大数据分析工具对用户行为进行分析，通过对大数据的充分使用和挖掘在商战中获胜。

何为“用户行为分析”？简单的理解就是对用户在网站上发生的所有行为进行分析，找到里面的规律和用户感兴趣的信息。这些行为如搜索商品信息、浏览新闻、购物评价和打分、美团点评、加入收藏列表、加入购物车、购买、使用特价购物券、换货和退货等；除此之外，还包括在第三方网站上的相关行为分析，如比价、看相关评测、参与讨论、社交媒体上的交流、与好友互动等。电子商务的突出特点就是可以通过后台收集到大量客户在购买前的行为信息，而这些信息对于分析用户行为至关重要。

全球电子商务的创始者之一亚马逊公司结合大数据分析工具，以迅雷不及掩耳之势，彻底颠覆了很多行业的市场规则及竞争关系。亚马逊取胜的根本原因在于对数据的战略性认识和使用，亚马逊通过传统门店无法比拟的互联网手段，获取了极其丰富的用户行为信息，并且进行深度分析与挖掘。

电商通过对用户行为的分析，可以制定更加贴心的服务。例如，当客户浏览了多款手机而没有做购买的行为时，系统会把适合客户的品牌、价位和类型的多款手机信息推送到客户的账户，只要客户登录系统，就可以看到这些推送信息。这样的个性化推荐服务往往会起到非常好的效果，不仅可以提高客户购买的意愿，缩短购买的路径和时间，还可以在比较恰当的时机捕获客户的最佳购买冲动，提升用户体验，是一个一举多得的好方式。

在电商领域中，用户的行为信息量之大令人难以想象，据不完全统计，用户在电商网站购买一个商品前，平均会浏览 3 ～ 5 个网站（淘宝、京东、1 号店等）、30 ～ 36 个页面，统计起来对于一个一天有近百万访问量的中型电商，一天就会有 1TB 左右的活跃数据。

除了电商，爱奇艺、优酷、土豆网等各类视频网站已成为人们娱乐和学习的重要平台，视频网站成功与否的最重要衡量标准就是用户的满意度。由于 Web 应用能够以很细的粒

度、很高的频度不断记录用户的行为轨迹，这些数据中隐藏着用户的访问习惯、兴趣偏好及情绪变化等信息，同时也隐含着用户群体行为的规律和发展趋势。挖掘深藏在数据背后的知识，能够发现用户习惯的观看流程，访问网站的时间及喜好的视频，各视频间存在的关联关系等。掌握了这些知识，就能科学解决用户跳出等问题，根据用户访问习惯改进网站服务流程，以及针对用户进行个性化服务，变革传统的网站管理和运营模式，主动提升用户的体验以促进视频网站的快速发展。

综观国内外成功的电商和视频网站等互联网企业，对用户行为信息的分析和使用，无不在这个必争之地做了大量投入。他们对数据战略性的高度认识和使用，非常值得国内的电商学习和借鉴。

R 语言作为能够进行交互式数据分析和探索的强大平台，具备一套完整的数据处理、计算和制图软件系统，在电商用户行为数据分析和挖掘领域应用广泛。基于 R 语言对视频网站的用户在线行为数据进行分析挖掘，通过对采集到的用户行为数据进行引流指标、黏性指标的分析和挖掘，可以得到网页被访问的频繁度、停留时间、用户观看视频的喜好等信息。将分析的结果应用到网站运营与管理中，不仅能够为网站个性化服务、精准推销和开发新型业

务模式提供技术和理论支撑，而且能够把握当前网络关注的热点问题，从而正确引导用户的网络舆论方向。

1.2.2 R 语言处理金融大数据

最近几年，数据分析方法在商业和金融市场上的重要性持续增加，因为我们有丰富的数据环境，经济和金融市场的数据相比以前更加综合和完整，在许多国家成百上千个变量的数据可以更系统、更精确地搜集，计算处理上的便利和统计软件包的使用使得对复杂的高维金融数据的分析成为可能，通过互联网可以很容易地应用开源软件包下载公开的金融数据，如 R 语言和软件开发环境，所有这些软件的特性和功能免费公开，因而被广泛使用。

2008 年的金融危机在某种程度上是由错误的金融模型造成的，既有模型过于简单的原因，又有模型过于复杂的原因，房地产经纪人和买家依赖于一个隐式模型，它表明价格已经在高位，且还会继续上涨。贷款人使用统计模型来对打包的按揭产品进行分析设计，这似乎可以奇迹般地降低风险，然而最后的结果是灾难性的，在 4 年之后仍然可以感受到房地产泡沫的影响。

那么，如何进行有用的并且没有危险的金融分析呢？首先应该对金融数据有一个基本的理解，尤其是时间序列数据，因为不确定性是主要的影响因素，比如可以用概率模型来描述资产收益率的频率分布，利用时间序列模型对数据进行描述平滑和季节调整。

使用 R 语言进行金融分析时，可以利用 R 语言的优势，研究分析基于 Hadoop 存储证券的日内交易数据，通过 RHive 连接 R 语言与 Hive，建立相关性算法模型，在历史数据中回测，构建投资决策组合，并生成可视化结果用于展示。

R 语言 quantmod 包是股市金融建模应用比较多的一个包。该包从多个数据源获取历史数据，绘制金融数据图表，以及在金融数据图表中添加各种技术指标，通过多种金融

模型分析，辅助股票筛选和判断。该包获取数据的来源主要有两个：Yahoo! 和 Google。最常用的是 Yahoo! 中的各种数据。但该包只能获取股票的历史交易记录信息，如最大值、最小值、开盘价、收盘价及成交量。在此基础上利用股票的历史数据，通过 R 语言建立模型，并对数据进行分析，从模型的检验决定未来的交易行为。

1.2.3 R 语言天气数据可视化

R 语言天气数据可视化，就是通过获取天气的历史大数据，使用 R 语言和可视化技术，展示中国每个省份的天气情况，给准备旅游的朋友提供一种出行提示。除了旅游，在传统零售行业，雨天大概会影响相对于晴天 30% ～ 40% 的销售业绩，所以从网上获取天气数据进行分析，并根据天气数据做出预测，提前做好预防措施和提醒业务人员，把损失减少到最低就显得十分重要。要实现天气数据可视化，需要实现的功能和遇到的问题如下。

- 天气数据：数据从哪里找到；如何下载；如何存储。
- 定时任务：天气数据需要每日更新，图片需要每日新生成。
- 地图和天气可视化：要把中国行政区划图和天气数据（包括风力方向可视化）结合在一起绘图，让用户一眼就能看明白。
- Web 展示：通过可视化技术，我们生成的只是一张静态图片，要如何发布到 Web 端进行展示。
- 微博：结合新浪微博，让更多的用户看到并使用这个应用。
- 用户交互：用户可以查看不同日期、不同类型的图片，用户还可以通过微博分享。

从上面的描述中，单独使用一种语言不容易实现这些功能。单独用 PHP 开发，做一个 Web 网站非常容易，连接新浪微博也有现成的 SDK 可以调用，获取数据及存储也不麻烦，但是如何实现地图和天气数据的可视化？这个是 R 语言的强项，所以，可以将 R 语言和 PHP 语言相结合，发挥 R 语言的优势，用 R 语言的 rvest 包就可以方便获取天气数据，并实现天气数据的可视化。

1.2.4 R 语言医疗大数据分析

医疗大数据是相对于一般数据而言的，指的是人们从医疗系统的软件系统中捕捉大容量数据，通过大数据分析获得新的认知，从而创造新的价值来源。医疗大数据几乎包含公民所有个人信息，包括医疗、饮食、住所、旅行登记等。在临床操作方面，有 5 个主要场景的大数据应用。根据麦肯锡公司的估计，如果这些应用被充分采用，针对美国一个国家的医疗健康开支一年就将减少 165 亿美元。主要场景的大数据应用包括 5 个方面。

（1）比较效果研究。

通过全面分析病人特征数据和疗效数据，然后比较多种干预措施的有效性，可以找到针对特定病人的最佳治疗途径。世界各地的很多医疗机构，如德国 IQWIG（德国医疗质量和效率研究所）、加拿大普通药品检查机构等，已经开始了 CER（比较效果研究）项目，并取得了初步成功。