新工科×新商科·大数据与商务智能系列

R 语言大数据分析与挖掘

谢笑盈　金康伟　主　编

陈海城　副主编

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

内 容 简 介

本书首先简要介绍了大数据分析与挖掘的相关概念，以及 R 语言的基础知识，以此来帮助读者了解、使用 R 语言；其次详细介绍了探索性数据分析、数据采集，以此来帮助读者了解数据的基本分析方法和数据的获取方法；然后着重介绍了目前主流的数据挖掘算法——时间序列算法、线性回归算法、分类算法、关联算法、聚类算法，从算法的原理到如何使用 R 语言进行算法实现都进行了详细的介绍并提供了实操代码，以此帮助读者学习数据挖掘及使用 R 语言完成数据挖掘任务；最后通过 6 个旅游行业的实际案例来帮助读者将学习到的知识运用到真实的业务场景中，并融会贯通整个知识体系。

本书无须读者具备 R 语言和大数据分析与挖掘的基础知识。无论是 R 语言初学者，还是熟练的 R 语言用户，都能从本书中找到有用的内容。本书既可以作为一本学习 R 语言的教材，也可以作为大数据分析与挖掘的工具书。

图书在版编目（CIP）数据

R 语言大数据分析与挖掘 / 谢笑盈，金康伟主编. —北京：电子工业出版社，2023.3
（大数据与商务智能系列）
ISBN 978-7-121-45238-3

Ⅰ. ①R… Ⅱ. ①谢… ②金… Ⅲ. ①程序语言－应用－数据处理－高等学校－教材 Ⅳ. ①TP274

中国国家版本馆 CIP 数据核字（2023）第 046101 号

责任编辑：王二华　　文字编辑：张天运
印　　刷：
装　　订：
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：17.5　　字数：448 千字
版　　次：2023 年 3 月第 1 版
印　　次：2023 年 3 月第 1 次印刷
定　　价：55.00 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：（010）88254532。

前　言

2015 年 8 月 31 日国务院印发了《促进大数据发展行动纲要》，同年党的十八届五中全会首次提出了“国家大数据战略”；2016 年 9 月 19 日，国务院出台了《政务信息资源共享管理暂行办法》，同年 12 月，工业和信息化部印发了《大数据产业发展规划（2016—2020 年）》；2021 年 11 月 30 日，工业和信息化部再次发布了《“十四五”大数据产业发展规划》，提出到 2025 年，我国大数据产业测算规模要突破 3 万亿元，年均复合增长率保持 25%左右，创新力强、附加值高、自主可控的现代化大数据产业体系要基本形成。国家对大数据战略的空前重视，更凸显了大数据分析与挖掘的巨大价值，伴随着数据在各行业、各领域的深层渗透及应用，大数据已经成为影响竞争和发展的重要因素，而对大数据的探索、分析、挖掘已经成为了大数据分析领域的基本技能之一。

2017 年 3 月 26 日，在浙江义乌正阳博瑞旅游产业发展有限公司与浙江师范大学联合成立的正阳旅游研究院的支持下，浙江师范大学经济与管理学院和经管之家 CDA 数据分析研究院合作举办了首期“正阳旅游大数据创新创业培训班”，以培养当前互联网经济背景下旅游业发展急需的兼具理论知识和实战经验的旅游大数据分析师。由于是首次开设旅游大数据创新创业培训班，市场上没有任何相应的教材可以借鉴，而培训对象是没有编程基础并且统计知识较为薄弱的旅游管理专业的学生，为了使他们能在较短的时间内理解大数据分析的真实价值，掌握大数据分析的过程，并能快速成长为兼有理论知识和实战经验的旅游大数据分析师，开发和出版大数据系列教材迫在眉睫。此时，具备强大统计能力的免费数据分析软件 R 语言进入了大家的视野，R 语言具有免费、开源、资源丰富、简单易学、可视化优、兼容性好等优势，对于大数据的收集、转换、探索、建模、可视化方面的工作都能够完全胜任，因此，授课老师一致决定选择 R 语言作为旅游大数据创新创业培训班的数据分析工具。经过两轮的教学实践，这本《R 语言大数据分析与挖掘》初步成形。

本书的目的是让读者掌握如何用 R 语言实现大数据分析与挖掘，秉持理论与实践相结合的原则，书中不仅提供了深入浅出的理论阐述及细致入微的思路剖析，还提供了大量的 R 语言操作代码，以达到引领读者迅速进入大数据分析与挖掘领域的目的。为了让读者提高解决实际问题的能力并对大数据分析与挖掘的各种方法融会贯通，书中配备了 6 个旅游行业的实际案例，包括旅游数据的采集、探索分析、挖掘建模等，这些案例中运用的方法同样适用于其他应用领域。

本书得益于很多人的帮助和支持。

首先，感谢上官诚兴先生，他通过出色的沟通工作促成了浙江师范大学经济与管理学院和经管之家 CDA 数据分析研究院的合作，并为首期“正阳旅游大数据创新创业培训班”的开班做了大量具体而细致的工作，这为本教材的出版提供了必要条件。

其次，非常感谢参与授课的零一老师、董雪婷老师、覃智勇老师为本教材的出版无偿提供了大量的素材和案例。

再次，由衷感谢浙江师范大学经济与管理学院旅游管理专业的同仁们，特别是龚海珍老师和马骏老师，因为他们的积极参与，正阳旅游大数据创新创业培训班才能得以正常运转，并良性循环。

另外，本教材的出版还受到了国家社科基金项目“基于抽样学习的非平衡数据分类方法研究（17BTJ028）”的资助，在此一并感谢。

编　者

目　录

第 1 章　大数据分析与挖掘概论 …… 1
1.1　大数据分析与挖掘 …… 1
1.1.1　大数据定义 …… 1
1.1.2　大数据分析与挖掘的概念 …… 2
1.2　大数据分析与挖掘流程 …… 3
1.2.1　数据获取 …… 3
1.2.2　数据预处理 …… 3
1.2.3　数据分析 …… 4
1.2.4　数据解释 …… 5
1.3　大数据分析与挖掘应用 …… 5
1.3.1　优化任务 …… 5
1.3.2　预测任务 …… 5
1.3.3　分类任务 …… 5
1.3.4　识别任务 …… 6
第 2 章　R 语言编程基础 …… 7
2.1　R 语言的安装及配置 …… 7
2.1.1　R 语言的获取和安装 …… 8
2.1.2　RStudio 的获取和安装 …… 9
2.2　界面与菜单 …… 12
2.2.1　RGui 界面 …… 12
2.2.2　RStudio 界面 …… 12
2.3　变量与数据类型 …… 15
2.3.1　变量 …… 15
2.3.2　数据类型 …… 15
2.4　数据结构 …… 17
2.4.1　向量 …… 17
2.4.2　数组 …… 19
2.4.3　矩阵 …… 21
2.4.4　列表 …… 22
2.4.5　数据框 …… 23
2.4.6　因子 …… 25
2.5　控制语句 …… 26
2.5.1　条件语句 …… 26
2.5.2　循环语句 …… 28

2.6 函数 …… 30
2.6.1 内置函数 …… 30
2.6.2 自定义函数 …… 33
第 3 章 数据预处理 …… 35
3.1 数据表的基本操作 …… 35
3.1.1 数据表保存 …… 35
3.1.2 数据表读取 …… 37
3.1.3 选取子集 …… 40
3.1.4 连接数据库 …… 42
3.2 数据分组、分割、合并和变形 …… 44
3.2.1 数据分组 …… 44
3.2.2 数据分割 …… 46
3.2.3 数据合并 …… 47
3.2.4 数据变形 …… 49
3.3 缺失值、异常值、重复值处理 …… 52
3.3.1 缺失值 …… 52
3.3.2 异常值 …… 59
3.3.3 重复值 …… 61
3.4 数据类型的转换 …… 62
3.4.1 判断数据类型函数 …… 62
3.4.2 转换数据类型的函数 …… 63
3.5 提取字符 …… 64
3.5.1 截取字符 …… 64
3.5.2 正则表达式 …… 65
第 4 章 探索性数据分析 …… 69
4.1 描述性统计方法 …… 69
4.1.1 常用统计指标 …… 69
4.1.2 数据总结 …… 70
4.2 数据可视化 …… 72
4.2.1 箱线图 …… 72
4.2.2 直方图 …… 74
4.2.3 散点图 …… 75
4.2.4 饼图 …… 77
第 5 章 数据采集 …… 80
5.1 网络数据采集的原理 …… 80
5.1.1 网页通信的过程 …… 80
5.1.2 请求数据的方法 …… 87
5.1.3 网页的组成元素 …… 88
5.2 数据采集入门 …… 90
5.2.1 数据采集常用包概述 …… 90
5.2.2 数据采集前的准备 …… 91
5.2.3 编写第一个数据采集 …… 92

5.3　使用常用的 R 包采集数据 …… 96
5.3.1　使用 RCurl 包获取网络数据 …… 96
5.3.2　使用 rvest 包获取网络数据 …… 98
5.3.3　使用 httr 包获取网络数据 …… 106
5.4　爬虫限制处理 …… 107
5.4.1　解决 IP 限制问题 …… 108
5.4.2　验证码处理 …… 109
5.4.3　登录问题处理 …… 110
第 6 章　时间序列算法 …… 111
6.1　时间序列算法概述 …… 111
6.1.1　时序对象 …… 113
6.1.2　时序平滑处理 …… 113
6.1.3　时序季节性分解 …… 116
6.2　时序指数模型 …… 118
6.3　时序 ARIMA 模型 …… 122
第 7 章　线性回归算法 …… 129
7.1　一元线性回归模型 …… 129
7.2　多项式回归模型 …… 131
7.3　多元线性回归模型 …… 133
第 8 章　分类算法 …… 136
8.1　Logistic 回归 …… 136
8.1.1　Logistic 回归算法原理 …… 137
8.1.2　逻辑回归算法应用 …… 139
8.2　决策树 …… 143
8.2.1　决策树算法原理 …… 143
8.2.2　决策树算法应用 …… 147
8.3　支持向量机 …… 150
8.3.1　支持向量机算法原理 …… 151
8.3.2　支持向量机算法应用 …… 155
8.4　朴素贝叶斯 …… 157
8.4.1　贝叶斯定理 …… 157
8.4.2　最大似然估计 …… 157
8.4.3　朴素贝叶斯分类算法原理 …… 158
8.4.4　朴素贝叶斯分类算法应用 …… 159
8.5　人工神经网络 …… 166
8.5.1　人工神经网络的基本概念 …… 166
8.5.2　感知器和人工神经元模型 …… 167
8.5.3　前馈神经网络 …… 168
8.5.4　人工神经网络算法应用 …… 169
8.6　随机森林 …… 177
8.6.1　随机森林算法原理 …… 177
8.6.2　随机森林算法应用 …… 178

8.7 XGBoost 算法……183
8.7.1 XGBoost 算法的原理……184
8.7.2 XGBoost 算法应用……187
第 9 章 关联算法……189
9.1 关联算法概述……189
9.1.1 相关名词……190
9.1.2 关联规则及频繁项集的产生……190
9.2 Apriori 算法……191
9.2.1 Apriori 算法概述……191
9.2.2 先验原理……191
9.2.3 连接步和剪枝步……192
9.2.4 Apriori 算法流程……193
9.2.5 Apriori 算法实例……193
9.3 ECLAT 算法……196
9.3.1 ECLAT 算法概述……196
9.3.2 ECLAT 算法流程……198
9.3.3 ECLAT 算法实例……198
第 10 章 聚类算法……202
10.1 聚类算法概述……202
10.1.1 聚类算法的类型……202
10.1.2 聚类算法评估的特点……202
10.2 K 均值聚类算法……203
10.2.1 划分方法概述……203
10.2.2 K 均值聚类算法的优缺点……203
10.2.3 K 均值聚类算法的流程……203
10.2.4 K 均值聚类分析案例……204
10.3 凝聚式层次聚类算法……205
10.3.1 凝聚式层次聚类概述……205
10.3.2 凝聚式层次聚类算法流程……207
10.3.3 凝聚式层次聚类算法实例……209
【应用案例 1】景点舆情数据采集……211
【应用案例 2】旅游电商平台数据采集……218
【应用案例 3】旅游网站景点路线推荐……233
【应用案例 4】旅游城市和景点的负荷预测……236
【应用案例 5】精品旅行服务成单预测……239
【应用案例 6】航班延误预测……259

第 1 章　大数据分析与挖掘概论

【内容概述】

1）了解大数据分析与挖掘的概念。
2）了解大数据挖掘与大数据分析的区别。
3）了解大数据分析与挖掘的应用。

1.1　大数据分析与挖掘

“大数据”这一概念最早公开出现于 1998 年，美国高性能计算公司 SGI 的首席科学家约翰·马西（John Mashey）在一个国际会议报告中指出：随着数据量的快速增长，必将出现数据难理解、难获取、难处理和难组织等 4 个难题，并用“Big Data（大数据）”来描述这一挑战，在计算领域引发思考。

2007 年，数据库领域的先驱人物吉姆·格雷（Jim Gray）指出大数据将成为人类触摸、理解和逼近现实复杂系统的有效途径，并认为在实验观测、理论推导和计算仿真等三种科学研究范式后，将迎来第四范式——“数据探索”，后来同行学者将其总结为“数据密集型科学发现”，开启了从科研视角审视大数据的热潮。

大数据于 2012 年、2013 年达到宣传高潮，2014 年后概念体系逐渐成形，对其认知亦趋于理性。大数据相关技术、产品、应用和标准不断发展，逐渐形成了由数据资源与 API、开源平台与工具、数据基础设施、数据分析、数据应用等板块构成的大数据生态系统，并持续发展和不断完善，其发展热点呈现了从技术向应用再向治理的逐渐迁移。经过多年的发展和沉淀，人们对大数据已经形成基本共识：大数据现象源于互联网及其延伸所带来的无处不在的信息技术应用及信息技术的不断低成本化。

1.1.1　大数据定义

大数据（Big Data）是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

大数据的 5V 特点为大量（Volume）、高速（Velocity）、多元（Variety）、价值（Value）、真实（Veracity）。

- 大量（Volume）：数据量大。数据量的大小决定所考虑数据的价值和潜在的信息。

- 高速（Velocity）：获得数据的速度快。
- 多元（Variety）：数据类型多样。
- 价值（Value）：合理运用大数据，以低成本创造高价值。
- 真实（Veracity）：数据准确可依赖。

1.1.2 大数据分析与挖掘的概念

从概念上可以认为，大数据分析是大数据挖掘的一个子项。在通常的概念下，它们之间是有差别的，但是严格意义下，大数据的所有成果都可以纳入大数据挖掘的成果范畴。大数据技术首先提供存储和计算能力，其次洞察数据中隐含的意义。大数据依赖硬件设备的升级，洞察数据的意义依赖大数据挖掘算法的不断优化创新。

1．大数据分析

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。这一过程也是质量管理体系的支持过程。在实际应用中，数据分析可帮助人们做出判断，以便采取适当行动。

大数据分析是在数据分析概念的基础上发展而来的。两者分析的数据对象不同，在大数据分析过程中不用对数据进行随机抽样分析，而是对所有数据进行分析处理。大数据分析在已定的假设、先验约束上处理原有计算方法、统计方法，将数据转化为有用的信息。

2．大数据挖掘

大数据挖掘又称为资料探勘、数据采矿。它是数据库知识发现中的一个步骤。大数据挖掘一般指从大量的数据中通过算法搜索隐藏于其中的信息的过程。大数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现搜索隐藏于数据中的信息。

3．两者的联系与区别

大数据分析与大数据挖掘既有联系也有区别，两者在协作上有以下联系。

（1）需要对大数据分析得到的信息进一步挖掘，将其转化为有效的预测和决策，这时就需要大数据挖掘。

（2）大数据挖掘进行价值评估的过程也需要调整先验约束而再次进行大数据分析。

两者在算法、数据和运行环境三个方面的区别如下。

（1）算法：大数据分析对算法的要求随着数据量的增加而降低，大数据挖掘则对算法要求更高，复杂度更大。

（2）数据：大数据分析的对象多为动态增量数据和存量数据，大数据挖掘则大多使用存量数据。

（3）运行环境：大数据分析对运行环境要求较高，多为云计算和云存储环境，而大数据挖掘则没有特定的要求，单机环境也是允许的。

1.2　大数据分析与挖掘流程

从大数据的特征和产生领域来看，大数据的来源相当广泛，由此产生的数据类型和应用处理方法千差万别。但是总的来说，大数据分析流程可划分为数据获取、数据预处理、数据分析和数据解释 4 个步骤。

1.2.1　数据获取

大数据的“大”，原本就意味着数量多、种类复杂，因此，通过各种方法获取数据信息便显得格外重要。数据获取是大数据分析流程中最基础的一步，目前常用的数据获取手段有传感器、射频识别、数据检索分类工具（如百度和谷歌等搜索引擎）、行业论坛或平台等商业网站及条形码技术等。

数据的类别主要分为线下数据和线上数据，线下数据主要依托硬件，如红外传感器、高清摄像头等设备来获取，线上数据主要依托互联网获取，如互联网舆情信息、商务平台商品信息等。

数据获取无时无刻不在进行中。例如，对于大型商场，在商场门口可以使用红外传感器记录进入商场的客户数量；客户连接商场提供的 Wi-Fi，可以通过 Wi-Fi 定位客户在商场中的行为轨迹；通过 POS 机或收银系统可以跟踪客户在商场中的消费情况。这些获取到的数据可经过处理后用于商场的精准营销、路径优化等。

1.2.2　数据预处理

数据预处理是非常重要的环节，对数据使用的一致性、准确性、完整性、时效性、可信性、可解释性提供了基本保障。现实中的数据避免不了“脏”数据，“脏”数据主要是指具备以下特征的数据。

（1）不完整：缺少属性值或仅包含处理后结果（没有源数据）的数据。

（2）包含噪声：存在错误或偏离期望值的数据。

（3）不一致：前后存在矛盾、差异的数据。

由于获取的数据规模庞大，存在大量的“脏”数据，因此在一个完整的大数据挖掘过程中，数据预处理是必不可少的环节，大约要花费 60%～70%的时间。

数据预处理有 4 种方法：数据清洗、数据集成、数据变换和数据规约。

（1）数据清洗：空缺值处理、格式标准化、错误纠正、异常数据和重复数据的清除。

（2）数据集成：将多个数据源中的数据结合起来并统一存储，建立数据仓库，并消除数据冗余。

（3）数据变换：平滑、聚集、数据概化、规范化、属性构造等。

（4）数据规约：数据立方体聚集、维度规约（删除不相关的属性）、数据压缩（用 PCA、

LDA、SVD、小波变换等方法进行数据降维）、数值规约（线性回归、对数线性模型、直方图、聚类、抽样）。

1.2.3 数据分析

数据分析是整个大数据分析流程里最核心的部分，因为在数据分析的过程中，会发现数据的价值所在。传统的数据分析方法已经不能满足大数据时代数据分析的需求。在数据分析技术方面，Google 公司于 2006 年率先提出了“云计算”的概念，其内部各种数据的应用都依托 Google 公司内部研发的一系列云计算技术，如分布式文件系统 GFS、分布式数据库 BigTable、批处理技术 MapReduce 及开源实现平台 Hadoop 等。这些技术平台对大数据进行处理和分析提供了很好的手段。

数据分析有如下 6 个基本方面。

1）可视化技术

不管是对数据分析专家来说还是对普通用户来说，数据可视化是数据分析工具最基本的要求。可视化技术可以直观地展示数据，让数据自己说话，让观众看到结果。

2）数据挖掘算法

如果说可视化是给人看的，那么数据挖掘就是给机器看的。集群、分割、孤立点分析，还有其他的算法深入数据内部，挖掘价值。这些算法不仅要处理大数据的量，还要面对处理大数据的速度问题。

3）预测性分析能力

数据挖掘可以让分析员更好地理解数据，而预测性分析可以让分析员根据可视化分析和数据挖掘的结果做出一些预测性的判断。

4）语义引擎

非结构化数据的多样性给数据分析带来了新的挑战，因此需要一系列的工具去解析、提取、分析数据。语义引擎需要被设计成能够从“文档”中智能提取信息。

5）数据质量和数据管理

数据质量（数据的真实性、准确性、完整性、时效性）和数据管理（如何有效保障数据质量）是管理方面的最佳实践。通过标准化的流程和工具对数据进行处理可以保证一个预先定义好的高质量的分析结果。

6）数据仓库

数据仓库是为了便于多维分析和多角度展示数据，按特定模式存储数据所建立起来的关系型数据库。在商业智能系统的设计中，数据仓库的构建是关键，是商业智能系统的基础，承担对业务系统数据整合的任务，为商业智能系统提供数据抽取、转换和加载（ETL），并按主题对数据进行查询和访问，为联机数据分析和数据挖掘提供数据平台。

1.2.4　数据解释

在一个完善的大数据分析流程中，数据解释至关重要。但随着数据量的加大，数据分析结果往往也越复杂，用传统的数据解释方法已经不足以满足数据分析结果输出的需求。因此，为了提升数据解释、展示能力，必须对数据进行可视化操作。通过可视化分析，可以形象地向用户展示数据分析结果，更方便用户理解和接受结果。常见的可视化技术有基于集合的可视化技术、基于图标的可视化技术、基于图像的可视化技术、面向像素的可视化技术和分布式可视化技术等。

1.3　大数据分析与挖掘应用

在大数据时代的背景下，数据资产如何被有效地利用起来成为了一个热门话题。数据最终是要为商业、民生、国防等方面的运作与优化提供支撑的。近年来，由于大数据技术的发展，大数据分析与挖掘的应用如雨后春笋遍及各个领域，例如，今日头条的个性推荐、高德地图的拥堵预测、公安机关的网络舆情巡查系统等。各种应用可谓是琳琅满目，但总结起来可以将大数据应用的场景分为优化、预测、分类和识别 4 个方面，这 4 个方面也是大数据分析的主要任务。

1.3.1　优化任务

优化是大数据分析的主要任务，通过数据反馈了解哪些方面需要改进从而制定相关的决策。优化任务还需要更多的技术手段。

在人们的生活中，大数据分析产生了许多便利的应用场景，具体如下。

（1）在出行方面：通过交通数据，交通实时预测算法可以改善人们的出行。

（2）在购物方面：通过用户行为和基础数据，个性化推荐算法可以改善人们的网上购物体验。

（3）在疫情防控方面：通过出行大数据创造的五色管理方法可以有效地对高风险人群进行预警，在降低疫情传播速度的同时，也方便了低风险人群的出行。

1.3.2　预测任务

预测是大数据分析和挖掘的最终目的，这是由于预测可以提前洞察到事物未来的趋势，掌握信息差，而信息差是制胜的关键，无论是商业上、政治上还是军事上，比竞争对手提前预知事物的发展态势是十分重要的。

1.3.3　分类任务

分类任务包含分类算法和聚类算法，分类和聚类有明显区别，分类是把现有事物打上

已知标签，聚类是把相似的事物放在一起。对事物进行分类或聚类后，可以了解每个现有事物的特征，或者预估新兴事物的特征。例如，医学上的自动诊断，通过对大量的检验报告及病症的分类训练，实现对新的检验报告的分类预测。网络舆情监控系统也是如此，通过分类任务来感知敏感信息，从而实现自动监控。

1.3.4 识别任务

识别是人工智能的范畴，对人或物的识别可以改变人们的生活方式，也可以提升社会各个生产环节的效率。例如，刷脸支付技术通过人脸的识别实现“无密”支付，这里的密码就是所有者的脸；停车场感应系统，通过识别车牌号码实现无须停车取卡，配合无感支付更实现了停取车的便捷化体验。