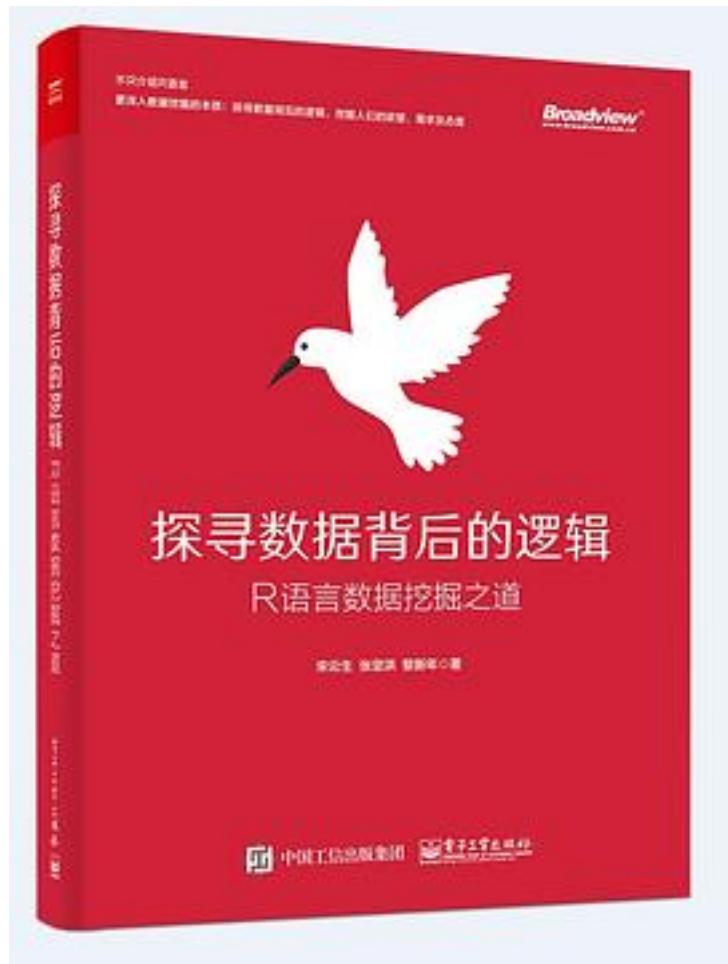


# 探寻数据背后的逻辑：R语言数据挖掘之道



[探寻数据背后的逻辑：R语言数据挖掘之道 下载链接1](#)

著者:宋云生

出版者:电子工业出版社

出版时间:2018-8

装帧:平装

isbn:9787121338618

数据分析、数据挖掘的本质是探寻数据背后的逻辑，挖掘人们的欲望、需求、态度等。《探寻数据背后的逻辑：R语言数据挖掘之道》不仅仅教会读者如何掌握数据挖掘相关技能，更教会读者如何从数据挖掘结果中分析出更深层次的逻辑。

《探寻数据背后的逻辑：R语言数据挖掘之道》主要介绍使用R语言进行数据挖掘的过程。具体内容包括R软件的安装及R语言基础知识、数据探索、数据可视化、回归预测分析、时间序列分析、算法选择流程及十大算法介绍、数据抓取、社交网络关系分析、情感分析、话题模型、推荐系统，以及数据挖掘在生物信息学中的应用。另外，《探寻数据背后的逻辑：R语言数据挖掘之道》还介绍了R脚本优化相关内容，使读者的数据挖掘技能更上一层楼。

《探寻数据背后的逻辑：R语言数据挖掘之道》适合从事数据挖掘、数据分析、市场研究的工作者及学生群体，以及对数据挖掘和数据分析感兴趣的初级读者

## 作者介绍：

宋云生，中山大学生命科学学院硕士毕业，混迹于医药商业、汽车制造等多个行业，先后从事市场研究、BI（商业智能）、质量控制等多个领域的数据研究和落地应用，现主攻自然语言理解领域的实际应用。

张坚洪，华南农业大学数学与应用数学本科毕业，先后从事汽车、金融等行业，主要工作方向为数据仓库、数据挖掘在互联网金融领域的应用。

黎新年，中山大学生命科学学院博士毕业，主要研究方向为基因组的进化、群体演化和系统发育。

## 目录: 第1章 万事不只开头难 1

1.1 工欲善其事，必先利其器：安装 1

1.1.1 安装R和RStudio 1

1.1.2 安装数据包 3

1.1.3 数据包加载、卸载、升级，查看帮助文档 5

1.1.4 什么样的R包值得相信 7

1.2 了解R的对象 8

1.2.1 如何进行常见的算术运算 8

1.2.2 R语言的三大数据类型 10

1.2.3 向量及其运算 12

1.2.4 因子变量鲜有人知的秘密 15

1.2.5 矩阵相关运算及神奇的特征值 17

1.2.6 数据框及其筛选、替换、添加、排序、去重 18

1.2.7 与数组（array）相比，表单（list）的用处更加广泛 22

1.2.8 如何进行数据结构之间的转化 23

1.3 R语言的重器：函数 26

1.3.1 自编函数 26

1.3.2 有用的R字符串函数 29

1.4 控制流在R语言里只是一种辅助工具 31

1.4.1 判断 32

1.4.2 循环 33

1.5 数据的读入与输出 35

1.5.1 常见数据格式的输入／输出（CSV、TXT、RDATA、XLSX） 35

1.5.2 数据库连接：Oracle、MySQL及Hive 37

1.5.3 乱码就像马赛克一样让人讨厌 39

第2章 数据探索，招招都是利器 41

2.1 不要在工作后才认识“脏数据” 41

2.1.1 以老板信服的方式处理缺失数据 42

2.1.2 异常值预警 48

2.1.3 字符处理正则表达式不再是天书 49

2.2 数据透视、数据整形、关联融合与批量处理	50
2.2.1 还忘不掉Excel的数据透视表吗	50
2.2.2 你能给数据做整形手术吗：long型和wide型	52
2.2.3 关联合并表	54
2.2.4 数据批处理：R语言里最重要的一个函数家族：*ply	55
2.3 一招完成数据探索报告	58
2.4 拯救你的很多时候是基础理论	61
2.4.1 参数检验及非参检验	62
2.4.2 学了很多算法却忘了方差分析	68
2.4.3 多因素方差分析及协方差作用	70
2.4.4 很多熟悉的数据处理方法已经成笑话，工具箱该换了	73
第3章 从商务气质的数据可视化说起	84
3.1 说说数据可视化的专业素养	84
3.1.1 数据可视化历史上有多少背影等你仰望	84
3.1.2 商务图表应该具有哪些素质	87
3.1.3 那些你不知道的图表误导性伎俩	94
3.1.4 如何快速解构著名杂志的图表	98
3.2 ggplot2包：一个价值8万美元的态度	103
3.2.1 一张图学会ggplot2包的绘图原理	105
3.2.2 基础绘图科学：ggplot2包的主题函数继承关系图（关系网络图）	127
3.2.3 基础图表一网打尽	132
3.2.4 古老的地图焕发新颜	151
3.3 将静态图转为D3交互图表：plotly	156
3.4 从基础到进阶的变形图表	157
3.4.1 马赛克图（分类变量描述性分析）	157
3.4.2 Sankey图和chordDiagram图	158
第4章 分位数回归模拟股票指数风险通道	163
4.1 用线性回归预测医院的药品销售额	163
4.2 多项式回归及常见回归方程的书写	168
4.3 Lasso回归和回归评价的常见指标	170
4.4 分位数回归拟合上证指数风险通道	175
第5章 时间序列分析	181
5.1 时间序列分析：分析带有时间属性的数列	181
5.2 不是所有序列都叫时间序列	181
5.3 时间序列三件宝：趋势、周期、随机波动	183
5.3.1 趋势	183
5.3.2 周期	184
5.3.3 随机波动	186
5.4 预测分析	186
5.4.1 指数平滑法	186
5.4.2 ARIMA模型预测	188
第6章 选择什么算法也有一套流程	192
6.1 重新审视一下这几个模型	192
6.1.1 Logistic回归	192
6.1.2 我要的不是一棵树，而是整座森林：随机森林	195
6.1.3 神奇的神经网络	196
6.2 银行信用卡评估模型之变量筛选	197
6.2.1 变量构建	197
6.2.2 Logistic回归变量筛选	198
6.2.3 随机森林变量筛选	203
6.2.4 人工神经网络建模	204
6.3 必须面对的模型评估	204
第7章 深入浅出十大算法	208
7.1 C5.0算法	208

7.1.1 一个重要的概念：信息熵	208
7.1.2 非列变量选择的实例	209
7.1.3 C5.0算法的R实现	210
7.2 K-means算法	212
7.2.1 K-means算法的R实现	212
7.2.2 怎么确定聚类数	213
7.3 支持向量机（SVM）算法	213
7.3.1 通俗理解SVM	214
7.3.2 SVM的R实现	216
7.4 Apriori算法	216
7.4.1 举例说明Apriori	217
7.4.2 Apriori算法的R实现	219
7.5 EM算法	220
7.5.1 举例说明EM算法	221
7.5.2 EM算法的R实现	222
7.6 PageRank算法	223
7.7 AdaBoost算法	224
7.8 KNN算法与K-means算法有什么不同	226
7.9 Naive Bayes（朴素贝叶斯）算法	227
7.10 CART算法	228
第8章 数据抓取	231
8.1 数据挖掘工程师不可抱怨“巧妇难为无米之炊”	231
8.2 抓取股市龙虎榜数据，碰碰运气	232
8.2.1 了解XML和Html树状结构，才能庖丁解牛	233
8.2.2 了解RCurl包和网页解析函数	234
8.2.3 抓取股票龙虎榜	235
8.2.4 资金流入分析	237
8.3 抓取某家医药信息网站全站药品销售数据	240
8.3.1 所有医药公司名称一网打尽	240
8.3.2 为什么抓取数据时可以使用For循环	242
8.3.3 不要把代码写复杂	244
8.3.4 用Sankey数据流描绘医药市场份额流动	248
第9章 不可不说的社交网络关系	254
9.1 社交网络图	254
9.1.1 社交网络图告诉你和谁交朋友	254
9.1.2 这几个基本概念你需要抓牢	256
9.1.3 还有比本章任务更有趣的数据挖掘吗	259
9.2 你还要装备几个评价指标	260
9.2.1 社交网络大小	260
9.2.2 社交网络关系的完备性	261
9.2.3 节点实力评价	262
9.3 全球某货物贸易中的亲密关系	263
9.3.1 全球某货物贸易数据整合清洗	263
9.3.2 分组和社交网络中心	267
9.3.3 全球某货物交易圈：寻找各自的小伙伴	270
9.4 中国电影演艺圈到底有没有“圈”	276
9.4.1 数据清洗与整形	276
9.4.2 看看演艺圈长什么样	279
9.4.3 谁才是演艺圈的“关系户”	281
9.4.4 用Apriori算法查查演艺圈合作的“朋友”关系	283
9.4.5 给范冰冰推荐合作伙伴	284
第10章 情感分析：一种准确率高达90%的新方法？	287
10.1 情感分析及其应用：这是老生常谈	287
10.1.1 情感分析的用途	287

10.1.2 情感分析的方法论	288
10.1.3 有关情感分析的一些知识和方向	289
10.2 文本分析的基本武器：R	290
10.2.1 RJava包配置	290
10.2.2 Rwordseg包安装	291
10.2.3 jieba分词包安装	291
10.3 基于词典的情感分析的效果好过瞎猜吗	292
10.3.1 数据整理及词典构建	292
10.3.2 分词整理	297
10.3.3 情感指数计算	299
10.3.4 方法评价：优、缺点分析	300
10.4 监督式情感分析：挑选训练数据集是所有人心中的痛	301
10.4.1 TFIDF指标	301
10.4.2 构建语料库	302
10.4.3 随机森林模型	304
10.4.4 算法评估：随机森林应该建多少棵树	308
10.5 一种准确率高达90%的新方法	316
10.5.1 拿来主义的启示	316
10.5.2 情感词典和规则构建	317
10.5.3 朴素贝叶斯情感分析器	329
10.5.4 支持向量机（SVM）、决策树等情感分析器	330
10.5.5 如何选择支持SVM的核函数	339
10.5.6 情感分类器方法评价	343
10.6 谈谈情感分析的下一步思考	344
第11章 话题模型：很多牛人过不去的坎儿	346
11.1 话题模型与文案文本集	346
11.1.1 任务仍然是以处理dirty data 开始	347
11.1.2 数据清洗	348
11.2 话题模型中几个重要的数据处理步骤	350
11.2.1 中文分词	350
11.2.2 数据整型	352
11.2.3 怎样设定“阈值”	353
11.3 上帝有多少个色子：话题数量估计	356
11.3.1 通俗地说一遍话题模型	356
11.3.2 主题数估计与交叉检验	357
11.3.3 如何使用复杂度、对数似然值确定主题数	362
11.4 LDA话题模型竟然能输出这么多关系	368
11.4.1 输出主题——词汇及其概率矩阵	368
11.4.2 输出主题——文档归属及其概率矩阵	369
11.5 话题之间也有社交（衍生）关系吗	370
11.6 话题模型的几个强大衍生品	372
11.6.1 话题模型提取特征词	372
11.6.2 三种方法确定聚类的类数和文本层次聚类	373
11.6.3 漂亮的文本聚类树和批量绘制大类词云图	375
第12章 排名就是简单的推荐系统吗？	378
12.1 全球宜居城市综合实力排行	378
12.1.1 综合实力排行：专家法VS数据驱动法	379
12.1.2 怎么比较两个排名结果	382
12.2 协同过滤推荐系统	383
12.2.1 基于商品的协同过滤系统（ItemCF）	386
12.2.2 基于用户的系统过滤系统（UserCF）	388
12.2.3 推荐系统效果评比	390
第13章 生物信息学中的数据挖掘案例	392
13.1 生物信息学与R语言	392

13.2 生物信息学中常用的软件包	392
13.2.1 软件包简介	392
13.2.2 数据表示方式——对象类 (class)	393
13.2.3 生物信息学R包简介：Bioconductor和CRAN	393
13.2.4 ape包	394
13.2.5 读懂你的对象	404
13.2.6 修改工具包中的函数以适应新情况	407
第14章 产品化：关于内存、速度和自动化	411
14.1 不同终端调用、自动化执行R脚本及参数传递	411
14.2 与速度、内存、并行相关的程序优化	414
· · · · · (收起)	

[探寻数据背后的逻辑：R语言数据挖掘之道](#) [下载链接1](#)

## 标签

R语言

数据分析

22

## 评论

翻过部分章节，不是很适合初学

[探寻数据背后的逻辑：R语言数据挖掘之道](#) [下载链接1](#)

## 书评

[探寻数据背后的逻辑：R语言数据挖掘之道](#) [下载链接1](#)