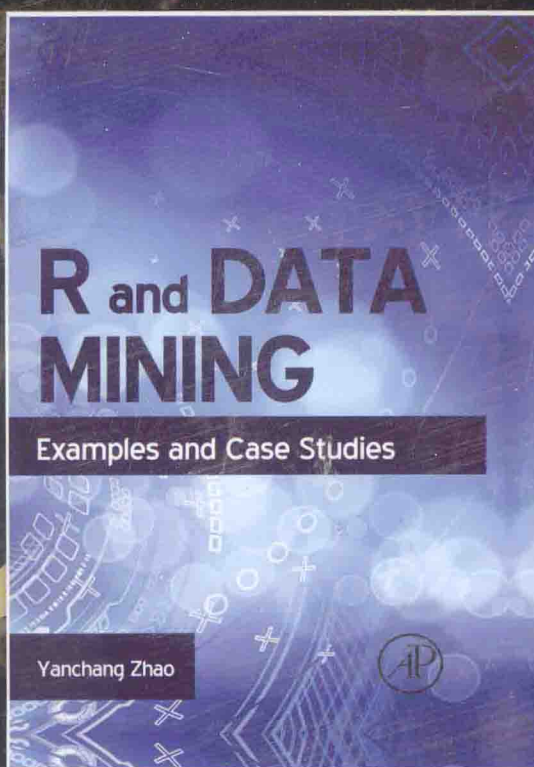


R语言与数据挖掘

最佳实践和经典案例

(澳) Yanchang Zhao 著 陈健 黄琰 译

R and Data Mining
Examples and Case Studies



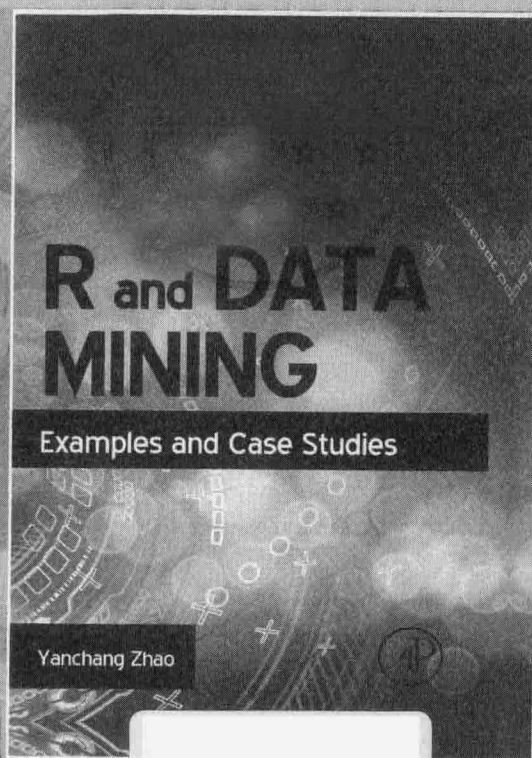
计 算 机 科 学 丛

R语言与数据挖掘

最佳实践和经典案例

(澳) Yanchang Zhao 著 陈健 黄琰 译

R and Data Mining
Examples and Case Studies



机械工业出版社
China Machine Press

图书在版编目 (CIP) 数据

R 语言与数据挖掘最佳实践和经典案例 / (澳) 赵彦昌著; 陈健, 黄琰译. —北京: 机械工业出版社, 2014.8

(计算机科学丛书)

书名原文: R and Data Mining: Examples and Case Studies

ISBN 978-7-111-47541-5

I. R… II. ① 赵… ② 陈… ③ 黄… III. ① 程序语言 – 程序设计 ② 数据采集
IV. ① TP312 ② TP274

中国版本图书馆 CIP 数据核字 (2014) 第 170002 号

本书版权登记号: 图字: 01-2013-6000

R and Data Mining: Examples and Case Studies

Yanchang Zhao

ISBN:978-0-123-96963-7

Copyright © 2013 Yanchang Zhao. Published by Elsevier Inc. All rights reserved.

Authorized Simplified Chinese translation edition published by the Proprietor.

Copyright © 2014 by Elsevier (Singapore) Pte Ltd. All rights reserved.

Printed in China by China Machine Press under special arrangement with Elsevier (Singapore) Pte Ltd. This edition is authorized for sale in China only, excluding Hong Kong SAR, Macau SAR and Taiwan. Unauthorized export of this edition is a violation of the Copyright Act. Violation of this Law is subject to Civil and Criminal Penalties.

本书简体中文版由 Elsevier (Singapore) Pte Ltd. 授权机械工业出版社在中国大陆境内独家出版和发行。本版仅限在中国境内 (不包括香港特别行政区、澳门特别行政区及台湾地区) 出版及标价销售。未经许可之出口, 视为违反著作权法, 将受法律之制裁。

本书封底贴有 Elsevier 防伪标签, 无标签者不得销售。

本书主要介绍如何使用 R 工具进行数据挖掘, 提供了 R 用于数据挖掘应用的案例、代码示例和数据, 涵盖了最常用的数据挖掘技术。

本书讲述简洁、内容实用, 配有现实应用案例和代码示例以及数据, 在线资源丰富, 适合数据挖掘方向的研究生、数据挖掘领域的研究人员, 以及数据挖掘工程师和分析师阅读。

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: 盛思源

责任校对: 殷 虹

印 刷: 北京瑞德印刷有限公司

版 次: 2014 年 9 月第 1 版第 1 次印刷

开 本: 185mm×260mm 1/16

印 张: 13.5

书 号: ISBN 978-7-111-47541-5

定 价: 49.00 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) 88378991 88361066

投稿热线: (010) 88379604

购书热线: (010) 68326294 88379649 68995259

读者信箱: hzjsj@hzbook.com

版权所有·侵权必究

封底无防伪标均为盗版

本书法律顾问: 北京大成律师事务所 韩光 / 邹晓东

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中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, Brian W. Kernighan, Dennis Ritchie, Jim Gray, Alfred 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，我们的联系方式如下：

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



华章科技图书出版中心

译者序

R and Data Mining: Examples and Case Studies

数据挖掘自产生以来就是以分析数据、理解数据的实际需求为推动力的，其研究发展也逐步渗透到工业、农业、医疗卫生和商业的实际需求当中。R 语言是在 S 语言的基础上逐步衍生出来的致力于数据统计分析与制图的语言。目前开源软件 R 也成为了世界上最流行的数据分析和处理工具之一，在学术研究和商业应用中都得到了广大数据分析者的青睐。

本书不是一本入门指导书，没有详细介绍数据挖掘技术的概念和理论，也没有介绍 R 语言的语言环境和语法规则，每一章节都结合具体例子详细介绍了 R 语言在数据挖掘的数据分析中的使用，实用性强。本书的读者最好具备数据挖掘和 R 语言的基础知识，通过结合本书介绍的具体案例分析来进一步了解 R 的对象、函数以及程序包，通过完整的 R 代码进一步理解数据挖掘技术在 R 语言工具上的具体实现，包括聚类、分类、关联规则分析、时间序列分析等。非常适合想要进一步学习 R 语言的人。

本书的代码都是采用 R 语言编写的，所使用的程序包都可以从 CRAN 网站上下载。R 程序包数量众多，所涵盖的领域也越来越广泛，包括生物学、统计学、遗传学、金融学等。相信使用 R 语言进行数据挖掘工作也能让读者尽快领略到这门学科的精妙之处。

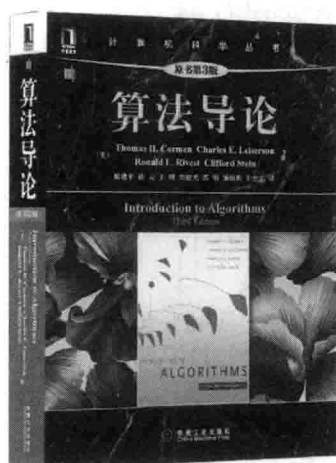
感谢所有在翻译过程中一直给予支持和鼓励的家人和朋友，感谢所有帮助和指导我的人。由于译者水平有限，本书的翻译难免会出现疏漏，还望各位读者不吝提出意见和建议。

译者

2014 年 6 月 9 日于华南理工大学

ARIMA (Autoregressive Integrated Moving Average)	自回归综合移动平均
ARMA (Autoregressive Moving Average)	自回归移动平均
AVF (Attribute Value Frequency)	属性值频率
CLARA (Clustering for Large Application)	大型应用程序的聚类
CRISP-DM (Cross Industry Standard Process for Data Mining)	数据挖掘的跨行业标准过程
DBSCAN (Density-Based Spatial Clustering of Application with Noise)	带噪声应用的基于密度的空间聚类
DTW (Dynamic Time Warping)	动态时间规整
DWT (Discrete Wavelet Transform)	离散小波变换
GLM (Generalized Linear Mode)	广义线性模型
IQR (InterQuartile Range)	四分位差, 即第一四分位数与第 三四分位数的差值
LOF (Local Outlier Factor)	局部离群点因子
PAM (Partitioning Around Medoids)	基于中心点的划分
PCA (Principal Component Analysis)	主成分分析
STL (Seasonal-Trend decomposition based on Loess)	基于局部加权回归的季节性趋势分解
TF-IDF (Term Frequency-Inverse Document Frequency)	词频 - 逆文档频率

推荐阅读



算法导论（原书第3版）

作者：Thomas H. Cormen 等 译者：殷建平 等 ISBN: 978-7-111-40701-0 定价：128.00元



机器学习

作者：Tom Mitchell 译者：曾华军 等 ISBN: 978-7-111-10993-7 定价：35.00元



数据挖掘：概念与技术（原书第3版）

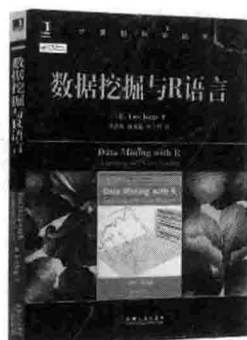
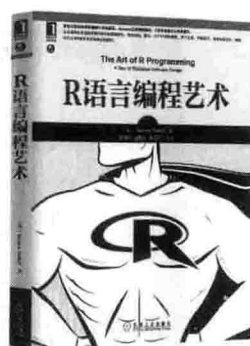
作者：Jiawei Han 等 译者：范明 等 ISBN: 978-7-111-39140-1 定价：79.00元



数据挖掘：实用机器学习工具与技术（原书第3版）

作者：Ian H. Witten 等 译者：李川 等 ISBN: 978-7-111-45381-9 定价：79.00元

推荐阅读



推荐阅读



Storm实时数据处理

作者: Quinton Anderson ISBN: 978-7-111-46663-5 定价: 49.00元



Splunk大数据分析

作者: Peter Zadrozny 等 ISBN: 978-7-111-46429-7 定价: 69.00元



Spark快速数据处理

作者: Holden Karau ISBN: 978-7-111-46311-5 定价: 29.00元



Hadoop应用开发技术详解

作者: 刘刚 ISBN: 978-7-111-45244-7 定价: 79.00元

目 录

R and Data Mining: Examples and Case Studies

出版者的话

译者序

缩写词表

第 1 章 简介 1

1.1 数据挖掘 1

1.2 R 1

1.3 数据集 2

1.3.1 iris 数据集 2

1.3.2 bodyfat 数据集 3

第 2 章 数据的导入与导出 4

2.1 R 数据的保存与加载 4

2.2 .CSV 文件的导入与导出 4

2.3 从 SAS 中导入数据 5

2.4 通过 ODBC 导入与导出数据 6

2.4.1 从数据库中读取数据 7

2.4.2 从 Excel 文件中导入与
导出数据 7

第 3 章 数据探索 8

3.1 查看数据 8

3.2 探索单个变量 10

3.3 探索多个变量 12

3.4 更多探索 15

3.5 将图表保存到文件中 19

第 4 章 决策树与随机森林 21

4.1 使用 party 包构建决策树 21

4.2 使用 rpart 包构建决策树 24

4.3 随机森林 29

第 5 章 回归分析 33

5.1 线性回归 33

5.2 逻辑回归 38

5.3 广义线性回归 38

5.4 非线性回归 40

第 6 章 聚类 41

6.1 k -means 聚类 41

6.2 k -medoids 聚类 43

6.3 层次聚类 45

6.4 基于密度的聚类 46

第 7 章 离群点检测 50

7.1 单变量的离群点检测 50

7.2 局部离群点因子检测 53

7.3 用聚类方法进行离群点检测 56

7.4 时间序列数据的离群点检测 58

7.5 讨论 59

第 8 章 时间序列分析与挖掘 60

8.1 R 中的时间序列数据 60

8.2 时间序列分解 60

8.3 时间序列预测 62

8.4 时间序列聚类 63

8.4.1 动态时间规整 63

8.4.2 合成控制图的时间序列数据 64

8.4.3 基于欧氏距离的层次聚类 65

8.4.4 基于 DTW 距离的层次聚类 66

8.5 时间序列分类 67

8.5.1 基于原始数据的分类	67	11.3 双模式网络	107
8.5.2 基于特征提取的分类	68	11.4 讨论与延伸阅读	110
8.5.3 k -NN 分类	69	第 12 章 案例 I：房价指数的分析	
8.6 讨论	70	与预测	111
8.7 延伸阅读	70	12.1 HPI 数据导入	111
第 9 章 关联规则	71	12.2 HPI 数据探索	112
9.1 关联规则的基本概念	71	12.3 HPI 趋势与季节性成分	118
9.2 Titanic 数据集	71	12.4 HPI 预测	120
9.3 关联规则挖掘	73	12.5 房地产估价	122
9.4 消除冗余	78	12.6 讨论	122
9.5 解释规则	79	第 13 章 案例 II：客户回复预测与效益	
9.6 关联规则的可视化	80	最大化	123
9.7 讨论与延伸阅读	82	13.1 简介	123
第 10 章 文本挖掘	84	13.2 KDD Cup 1998 的数据	123
10.1 Twitter 的文本检索	84	13.3 数据探索	131
10.2 转换文本	85	13.4 训练决策树	137
10.3 提取词干	86	13.5 模型评估	140
10.4 建立词项 - 文档矩阵	88	13.6 选择最优决策树	143
10.5 频繁词项与关联	90	13.7 评分	145
10.6 词云	91	13.8 讨论与总结	148
10.7 词项聚类	92	第 14 章 案例 III：内存受限的大数据	
10.8 推文聚类	94	预测模型	150
10.8.1 基于 k -means 算法的推文		14.1 简介	150
聚类	94	14.2 研究方法	150
10.8.2 基于 k -medoids 算法的推文		14.3 数据与变量	151
聚类	96	14.4 随机森林	152
10.9 程序包、延伸阅读与讨论	98	14.5 内存问题	153
第 11 章 社交网络分析	99	14.6 样本数据的训练模型	154
11.1 词项网络	99	14.7 使用已选变量建立模型	156
11.2 推文网络	102	14.8 评分	162

14.9 输出规则	168	15.7 R 的关联规则挖掘	181
14.9.1 以文本格式输出规则	168	15.8 R 的空间数据分析	181
14.9.2 输出 SAS 规则的得分	172	15.9 R 的文本挖掘	182
14.10 总结与讨论	177	15.10 R 的社交网络分析	182
第 15 章 在线资源	178	15.11 R 的数据清洗与转换	182
15.1 R 参考文档	178	15.12 R 的大数据与并行计算	182
15.2 R	178	R 语言数据挖掘参考文档	184
15.3 数据挖掘	179	参考资料	197
15.4 R 的数据挖掘	180	通用索引	201
15.5 R 的分类与预测	181	包索引	203
15.6 R 的时间序列分析	181	函数索引	204

简介

本书主要介绍使用 R 工具进行数据挖掘, 将给出在 R 中实现数据挖掘功能的许多例子, 以及 3 个真实应用场景的案例。本书适合硕士研究生、研究人员和想要使用 R 工具进行数据挖掘研究和项目的数据挖掘工程师。我们假定读者已经具备基础的数据挖掘知识, 并且接触过 R 语言。我们希望能够通过这本书鼓励更多的人在他们的研究与应用中使用 R 进行数据挖掘工作。

本章主要介绍一些关于数据挖掘的基本概念和技术, 包括数据挖掘的过程和常用的数据挖掘技术。同时, 还将介绍 R 语言, 以及进行数据挖掘时常用的 R 程序包、函数和任务视图。本章的最后还将介绍本书中将要使用到的几个数据集。

1.1 数据挖掘

数据挖掘是从大量的数据中发现有趣知识的过程 [Han and Kamber, 2000]。数据挖掘是一个涉及多个领域的交叉学科, 包括统计学、机器学习、信息检索、模式识别以及生物信息学。数据挖掘已经在许多领域中得到了广泛应用, 例如零售、金融、通信以及社交媒体行业。

数据挖掘的主要技术包括分类与预测、聚类、离群点检测、关联规则、序列分析、时间序列分析和文本挖掘, 同时还包括一些新的技术, 例如社交网络分析和情感分析。想要了解更多关于数据挖掘相关技术的详细介绍可以参考有关数据挖掘的教材 [Han and Kamber, 2000; Hand et al, 2001; Witten and Frank, 2005]。CRISP-DM (Cross Industry Standard Process For Data Mining, 数据挖掘的跨行业标准过程)^①将一个真实应用中的数据挖掘过程划分为 6 个主要阶段: 业务理解、数据理解、数据预处理、建模、评估和部署。本书的重点是建模阶段, 其中也有部分章节介绍数据探索和模型评估。读者想要获取更多关于数据挖掘的信息可以参考第 15 章中的在线资源。

1

1.2 R

R^② [R 核心开发团队, 2012] 是一个自由软件, 主要用于统计计算和统计制图, 它提供了大量的统计和制图工具。R 可以简单地通过程序包的形式进行扩展, 截至 2012 年 8 月 1 日, CRAN^③程序包库中已经收藏了大约 4000 个程序包。更多关于 R 的详细介绍可以参考《An Introduction to R》^④ (Venables et al., 2010) 和《R Language Definition》^⑤ [R 核心开发团队,

① <http://www.crisp-dm.org/>.

② <http://www.r-project.org/>.

③ <http://cran.r-project.org/>.

④ <http://cran.r-project.org/doc/manuals/R-intro.pdf>.

⑤ <http://cran.r-project.org/doc/manuals/R-lang.pdf>.

2010b], 这两本书都可以从 CRAN 网站上获取。R 已经广泛应用于学术界和工业界。

为了帮助用户更快地找到需要使用的 R 程序包, CRAN 提供了任务视图 (Task Views)[⊖], 将所有的 R 程序包按照不同的处理任务组织起来。与数据挖掘相关的几个任务视图是:

- 机器学习和统计学习。
- 聚类分析和有限混合模型。
- 时间序列分析。
- 多元统计。
- 空间数据分析。

另一个关于 R 实现数据挖掘的入门指导是《R Reference Card for Data Mining》, 它提供了一套完整的用于数据挖掘的 R 程序包和函数的索引, 并根据功能进行分类。最新的版本可以从 <http://www.rdatamining.com/docs> 上下载。

读者想要了解更多关于 R 的详细信息, 可以参考第 15 章的在线资源。

1.3 数据集

本节简单地介绍本书中将要使用到的数据集。

1.3.1 iris 数据集

在许多的科研著作中都在 iris 数据集上做分类操作。该数据集由 3 种不同类型的鸢尾花的 50 个样本数据构成 [Frank and Asuncion, 2010]。其中的一个种类与另外两个种类是线性可分离的, 后两个种类是非线性可分离的。这个数据集包含了 5 个属性:

- Sepal.Length (花萼长度), 单位是 cm。
- Sepal.Width (花萼宽度), 单位是 cm。
- Petal.Length (花瓣长度), 单位是 cm。
- Petal.Width (花瓣宽度), 单位是 cm。
- 种类: Iris Setosa (山鸢尾)、Iris Versicolour (杂色鸢尾), 以及 Iris Virginica (维吉尼亚鸢尾)。

```
> str(iris)

'data.frame': 150 obs. of 5 variables:
 $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width: num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width: num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species: Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1
 1 1 1 1 1 ...
```

[⊖] <http://cran.r-project.org/web/views/>.

1.3.2 Bodyfat 数据集

Bodyfat 数据集由 mboost 包 [Hothorn et al., 2012] 提供。该数据集包含 71 行，每一行代表一个客户的信息。数据集包含 10 列数值型数据。

- age: 年龄。
- DEXfat: 以 DXA 计算的体脂重，响应变量。
- waistcirc: 腰围。
- hipcirc: 臀围。
- elbowbreadth: 肘宽。
- kneebreadth: 膝宽。
- anthro3a: 三项人体测量的对数和。
- anthro3b: 三项人体测量的对数和。
- anthro3c: 三项人体测量的对数和。
- anthro4: 三项人体测量的对数和。

DEXfat 的值可通过其他的变量预测得到。

3

```
> data("bodyfat", package = "mboost")
> str(bodyfat)
'data.frame':    71 obs. of 10 variables:
 $ age: num 57 65 59 58 60 61 56 60 58 62 ...
 $ DEXfat: num 41.7 43.3 35.4 22.8 36.4 ...
 $ waistcirc: num 100 99.5 96 72 89.5 83.5 81 89 80 79 ...
 $ hipcirc: num 112 116.5 108.5 96.5 100.5 ...
 $ elbowbreadth: num 7.1 6.5 6.2 6.1 7.1 6.5 6.9 6.2 6.4 7 ...
 $ kneebreadth: num 9.4 8.9 8.9 9.2 10 8.8 8.9 8.5 8.8 8.8 ...
 $ anthro3a: num 4.42 4.63 4.12 4.03 4.24 3.55 4.14 4.04 3.91 3.66
 ...
 $ anthro3b: num 4.95 5.01 4.74 4.48 4.68 4.06 4.52 4.7 4.32 4.21
 ...
 $ anthro3c: num 4.5 4.48 4.6 3.91 4.15 3.64 4.31 4.47 3.47 3.6 ...
 $ anthro4: num 6.13 6.37 5.82 5.66 5.91 5.14 5.69 5.7 5.49 5.25 ...
```

4

数据的导入与导出

本章将介绍如何将外部数据导入 R 中，以及将 R 对象导出到其他格式的文件中。首先，我们将给出将 R 对象保存到 .Rdata 文件并从 .Rdata 文件中加载 R 对象的例子；然后分别演示从 .CSV 文件、SAS 数据库、ODBC 数据库以及 EXCEL 文件中导入与导出数据。读者想要了解更多关于数据导入与导出的详细内容可以参考《R Data Import/Export》[⊖] [R 核心开发团队, 2010a]。

2.1 R 数据的保存与加载

R 中的数据可以通过函数 `save()` 保存为 .Rdata 文件，并且可以通过函数 `load()` 将数据加载到 R 中。在下面的代码中，使用函数 `rm()` 将对象 `a` 从 R 中删除。

```
> a <- 1:10

> save(a, file="./data/dumData.Rdata")

> rm(a)

> load("./data/dumData.Rdata")

> print(a)

[1] 1 2 3 4 5 6 7 8 9 10
```

2.2 CSV 文件的导入与导出

下面的例子创建了一个名为 `df1` 的数据框，并通过函数 `write.csv()` 将其保存为一个 .CSV 文件。然后，使用函数 `read.csv()` 将数据框 `df1` 加载到数据框 `df2` 中。

```
> var1 <- 1:5

> var2 <- (1:5) / 10

> var3 <- c("R", "and", "Data Mining", "Examples", "Case
  Studies")

> df1 <- data.frame(var1, var2, var3)

> names(df1) <- c("VariableInt", "VariableReal", "VariableChar")

> write.csv(df1, "./data/dummyData.csv", row.names = FALSE)

> df2 <- read.csv("./data/dummyData.csv")
```

[⊖] <http://cran.r-project.org/doc/manuals/R-data.pdf>.