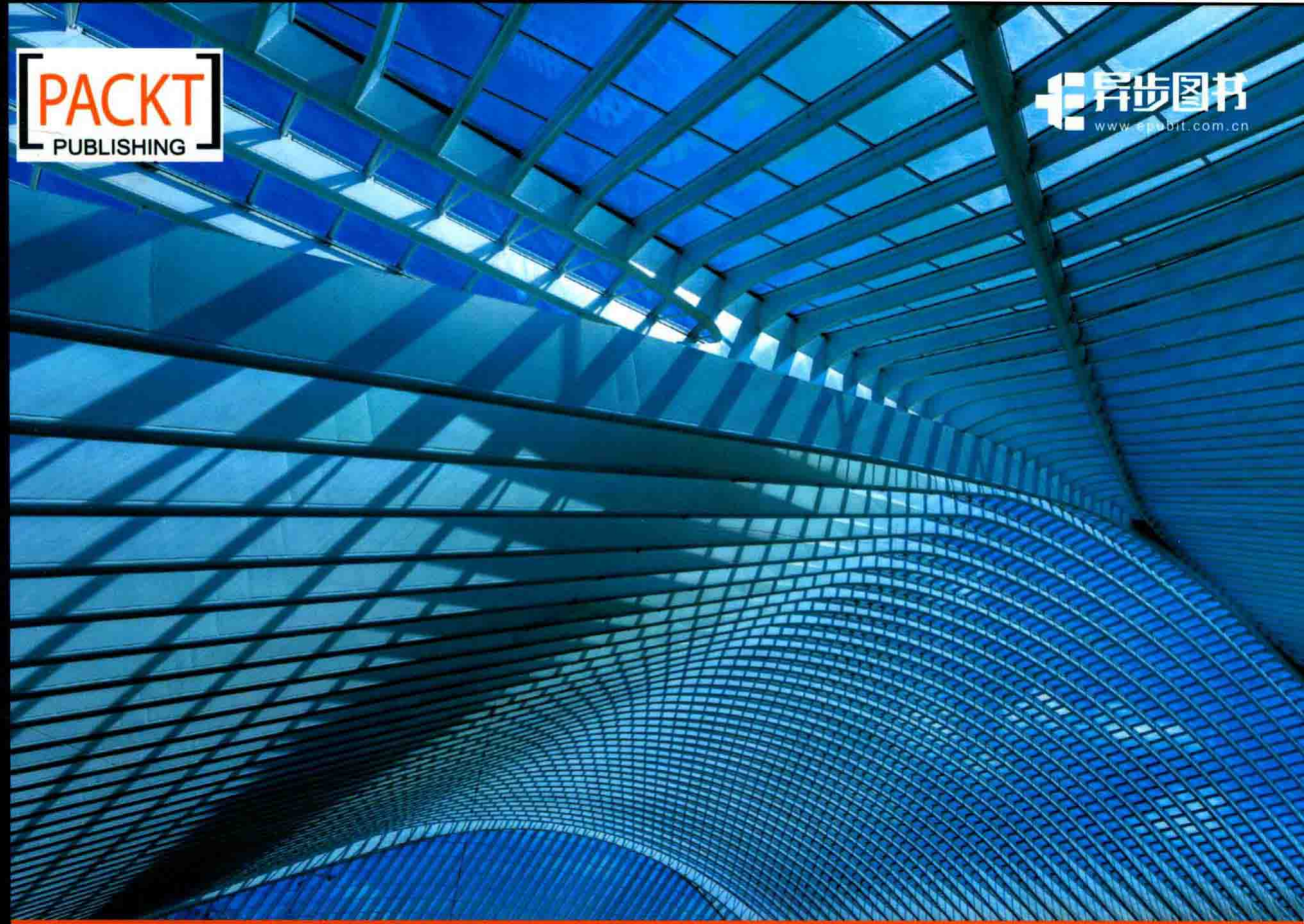


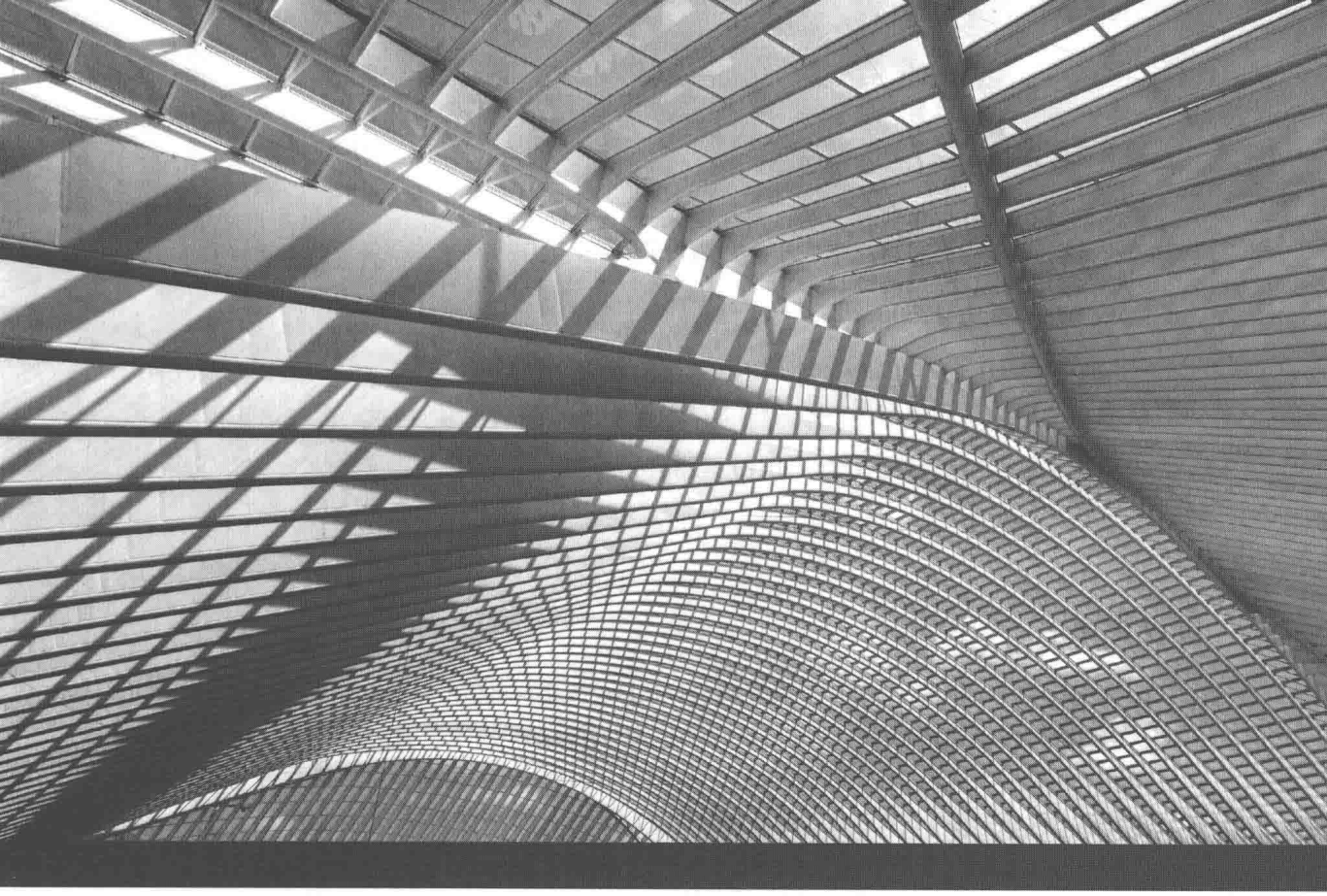
利用R语言学习、探索数据科学的基本原理

数据科学R语言实战

R for Data Science

[美] Dan Toomey 著
刘丽君 李成华 卢青峰 译

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS



数据科学
R语言实战

[美] Dan Toomey 著
刘丽君 李成华 卢青峰 译

人民邮电出版社
北京

图书在版编目（CIP）数据

数据科学 : R语言实战 / (美) 丹·图米(Dan Toomey) 著 ; 刘丽君, 李成华, 卢青峰译. -- 北京 : 人民邮电出版社, 2016.11
ISBN 978-7-115-43590-3

Ⅰ. ①数… Ⅱ. ①丹… ②刘… ③李… ④卢… Ⅲ. ①程序语言－程序设计②数据采集 Ⅳ. ①TP312②TP274

中国版本图书馆CIP数据核字(2016)第233185号

版权声明

Copyright © Packt Publishing 2014. First published in the English language under the title R for Data Science，ISBN 978-1-78439-086-0. All rights reserved.

本书中文简体字版由 Packt Publishing 公司授权人民邮电出版社出版。未经出版者书面许可，对本书的任何部分不得以任何方式或任何手段复制和传播。

版权所有，侵权必究。

◆ 著　[美] Dan Toomey
译　刘丽君　李成华　卢青峰
责任编辑　王峰松
责任印制　焦志炜

◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路 11 号
邮编　100164　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
三河市海波印务有限公司印刷

◆ 开本：800×1000　1/16
印张：21.5
字数：426 千字　2016 年 11 月第 1 版
印数：1－3 000 册　2016 年 11 月河北第 1 次印刷

著作权合同登记号　图字：01-2015-4183 号

定价：69.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315

内容提要

本书讲述的是R语言在数据科学中的应用，目标读者是从事不同行业的数据分析师、数据挖掘工程师、机器学习工程师、自然语言处理工程师、数据科学家，以及从事大数据和人工智能领域的工作者、学生、老师等。

本书最大的优点在于其通俗易懂、容易上手，每一个实例都有现成的数据和源代码，读者不仅能理解整个案例的来龙去脉，还可以直接编译本书提供的所有源代码，从而了解怎么从实际问题转变成可实现的代码，感受R语言的魅力，让数据产生价值。这种学习和实践相结合的方式非常适合初学者和有一定经验的数据分析师。

本书的内容涵盖了基于数据挖掘的常用模型，包括分类、聚类、关联分析、预测、异常检测等，还包括机器学习的常用算法和自然语言处理、数据可视化等内容。本书内容全面，做到了易读、易用、易理解、易实现、易上手，是不可多得的R语言书籍。

作者简介

Dan Toomey 具有 20 多年开发应用程序方面的经验，曾在多个行业及公司担任不同的职位，包括唯一投稿人、副总裁及首席技术官。近 10 年，Dan 一直在美国马萨诸塞州东部地区的公司工作。Dan 以 Dan Toomey 软件公司的名义，成为这些领域的开发承包商。

试读结束：需要全本请在线购买： www.ertongbook.com

译者简介

刘丽君，韩国国立全北大学博士，加拿大圣西维尔大学博士后，一直从事物联网、工业大数据等方面的数据分析、市场分析等工作，目前任武汉泰迪智慧科技有限公司 CEO，对数据敏感，并对数据怎么转变成价值、数据与商业的关系有独到见解。

李成华，数据挖掘与机器学习方向博士，约克大学博士后，麻省理工学院访问科学家，曾任海信集团数据挖掘专家，京东深度神经网络实验室首席科学家，长期从事数据挖掘、机器学习、深度学习和自然语言处理等方面的研究和工作，擅长自动问答以及基于自然语言的人机交互。

卢青峰，硕士毕业于美国威斯康辛州立大学，毕业后从事数据分析、挖掘等相关工作至今。曾先后在敦煌网、百度、京东等行业领先的公司从事数据挖掘、用户行为分析、推荐等工作。

关于中文版审稿人

在本书翻译的过程中，来自“统计之都”的核心成员，包括九峰医疗的数据研究总监李舰、雪晴数据网的创始人陈堰平、京东搜索与大数据平台部的刘思喆等，对本书的中文译稿进行了仔细的审阅，并提出了很多有价值的意见和建议。在此对他们的付出和努力表示感谢。

英文版审核人简介

Amar Gondaliya 是数据科学家，任职于一家著名的医疗机构。他是大数据及数据挖掘的爱好者，也是 Pingax（www.pingax.com）的顾问，专注于使用数据挖掘技术构建预测模型。他热衷于研究大数据技术，并提出大数据技术的解决方案。

他在数据科学领域工作了 2 年多，是 Pingax 的投稿人，撰写关于机器学习及其在 R 中实现的文章。他一直致力于机器学习新技术及大数据分析的研究。

Mohammad Rafi 是一名软件工程师，热衷数据分析、编程以及修修补补。他致力于 R、Python、Hadoop 及 JavaScript 等技术的研究。白天，他是工程师；晚上，他是游戏发烧友。自撰写本书起，他热衷于 Raspberry Pi 超级小电脑。

他在应用开发、数据处理、搜索专家及网络数据分析方面有 6 年以上的专业经验。最初，他在一家名为 Position2 的网络营销公司工作。此后，先后在《印度斯坦时报》、谷歌及 InMobi 等公司工作。

Tengfei Yin 获得了中国南开大学生命科学与技术学院的理学学士学位，并且在美国爱荷华州立大学完成了分子、细胞和发育生物学（MCDB）的学习，重点研究计算生物学和生物信息学。他的研究领域包括信息可视化、高通量生物数据分析、数据挖掘、机器学习及应用统计遗传学。他开发并且维护了 R 及 Bioconductor 中的多种软件包。

前言

R 是为数据操作及统计计算提供语言及环境的软件包，同样也能够用图表表示产生的统计数据。

R 具有以下特性：

- 语法简洁，可对数据执行操作；
- 附带的工具可通过本地和互联网以多种格式加载和存储数据；
- 语言一致，可对内存中的数据集进行操作；
- 具有用于数据分析的内置和开源工具；
- 采用生成实时图形和将图示存储到磁盘的方法。

本书内容

第 1 章：模式的数据挖掘，包括 R 中的数据挖掘。在此示例中，我们会寻找数据集中的模式。本章探讨通过使用多种工具进行聚类分析的示例，同样也包括了异常检测和关联规则的使用。

第 2 章：序列的数据挖掘，探讨了 R 中可以让读者发现数据中序列的方法。多种可用的 R 功能包能够帮助读者确定序列并通过图形描绘做进一步分析。

第 3 章：文本挖掘，介绍了多种挖掘 R 中文本的方法。我们会着眼于能够帮读者处理并分析源中文本或文字的工具，同样也会研究 XML 处理能力。

第 4 章：数据分析——回归分析，探讨了对数据进行回归分析的不同方法。本章运用多种方法来运行简易回归和多元回归以及后续的显示。

第 5 章：数据分析——相关性，探讨了多种相关功能包。本章不仅运用皮尔森、多分格、四分、异构及部分相关性，还运用基本的相关性及协方差对数据进行分析。

第 6 章：数据分析——聚类，探讨了各种聚类分析的参考文献。本章包括 K-means、PAM 和若干其他聚类技术。R 程序员可利用所有技术。

第 7 章：数据可视化——R 图形，探讨了各种使数据可视化的方法。我们会着眼于数据的全部范围，包括典型的类显示、第三方工具的相互作用和地图的使用。

第 8 章：数据可视化——绘图，探讨了绘制 R 中数据的不同方法。本章不仅提供了可用于绘制数据的自定义显示的示例，还提供了带有标准化显示的简易绘图的示例。

第 9 章：数据可视化——三维，作为直接从 R 创建数据三维显示的指南，我们同样也会考虑使用更大数据集的三维显示。

第 10 章：机器学习实战，探讨了如何使用 R 进行机器学习。本章包括把数据集分成训练数据及测试数据，从训练数据中构建模型并对模型进行试验，与试验数据进行比较。

第 11 章：用机器学习预测事件，使用了时间序列数据集。本章包括将数据转化成 R 时间序列，然后将季节性、趋势及不规则分量分离出来。其目标在于模拟或预测未来事件。

第 12 章：监督学习和无监督学习，阐释了如何使用监督和无监督学习来构建模型。本章包括多种有关监督和无监督学习的方法。

必备工具

关于本书，读者需要将 R 安装在机器（或将要运行脚本的机器）上。许多平台都可使用 R。本书并非局限于 R 的现有特定版本。

为了更好地使用本书，读者需要开发 R 项目的交互工具。主要工具是 R Studio，是许多平台均可使用的完全交互式且自包含程序，可以让读者输入脚本、显示数据并显示图形结果。各种 R 的安装总会有 R 命令行工具。

目标读者

本书的目标读者是牢固掌握先进数据分析技术的数据分析师。同样也要求基本了解 R 语言及某些数据科学话题。本书假设读者可以使用 R 环境，并且对涉及的统计有所了解。

读者反馈

欢迎来自读者的反馈，以让我们了解读者对本书的看法——读者喜欢或不喜欢的内容。读者反馈是重要的部分，能够帮助我们更好地阐述本书的主题，以便读者以后能够更好地掌握其内容。

将一般反馈发送至邮箱：feedback@packtpub.com，需在邮件主题中标明本书的标题。

如果读者擅长某一话题，并且对写书或撰稿感兴趣，登录 www.packtpub.com/authors，可查看作者指南。

用户支持

读者购买了 Packt 出版的书籍，我们会帮助读者从中获得最大收获。

下载示例代码

凡购买过 Packt 出版的书籍的用户，均可访问 http://www. packtpub. com，登录账户，并下载书籍的示例代码文件。如果读者是在别处购买的本书，可登录 http://www.packtpub.com/support，注册账户，以便通过电子邮箱直接获取文件。

下载本书的彩色图像

我们同样也为读者提供了含有本书中截屏/图表的彩色图像 PDF 文件。彩色图像会帮助读者更好地理解输出中产生的变化。登录 https://www.packtpub.com/sites/default/ files/downloads/0860OS_ColoredImages.pdf 即可下载本文件。

勘误

尽管我们已格外小心以确保内容的准确性，仍然无法避免书中会出现疏漏。如果读者发现在我们出版的书籍中存在错误——可能是文字或代码错误——请告知我们，我们将感激不尽。疏漏改正后，可以避免影响其他读者的阅读体验，并且有助于提高本书的后续版本。如果读者发现任何勘误，请按照以下方法告知：登录 http://www.packtpub.com/submit-errata，选择书籍，单击“**勘误提交表格**”链接，并输入勘误内容。一旦勘误通过核实，将会接收提交，并且勘误会上传至我们的网站或添加至此标题下“勘误”一节中现有的勘误表。

登录 https://www.packtpub.com/books/content/support 并在“搜索栏”输入书名，可查看以前提交的勘误。“**勘误**”一节中会提供必要信息。

版权保护

非法翻印网上有版权的材料是所有媒体中存在已久的问题。在 Packt，我们很重视对自身版权及授权的保护。如果读者在网上看见以任何形式盗版我们著作的行为，请立即向我们提供位置地址或网站名称，以便我们采取补救措施。

请将带有涉嫌盗版内容链接的邮件发送至 copyright@packtpub.com 与我们联系。

感谢读者对保护作者及我们能为读者提供有价值的内容而提供的帮助。

疑问

如果读者对本书留有疑问，可发送邮件至 questions@packtpub.com 与我们联系，我们会竭尽所能解答读者的疑问。

试读结束：需要全本请在线购买： www.ertongbook.com

目录

第 1 章　模式的数据挖掘 …… 1

1.1　聚类分析 …… 2

1.1.1　K-means 聚类 …… 3

1.1.2　K-medoids 聚类 …… 7

1.1.3　分层聚类 …… 12

1.1.4　期望最大化 …… 15

1.1.5　密度估计 …… 21

1.2　异常检测 …… 24

1.2.1　显示异常值 …… 25

1.2.2　计算异常 …… 28

1.3　关联规则 …… 30

1.4　问题 …… 33

1.5　总结 …… 34

第 2 章　序列的数据挖掘 …… 35

2.1　模式 …… 35

2.1.1　Eclat …… 36

2.1.2　arulesNBMiner …… 40

2.1.3　Apriori …… 43

2.1.4　用 TraMineR 确定序列 …… 47

2.1.5　序列相似点 …… 54

2.2　问题 …… 57

2.3　总结 …… 57

第3章 文本挖掘 ····· 59

3.1 功能包 ····· 60

3.1.1 文本处理 ····· 60

3.1.2 文本集群 ····· 69

3.2 问题 ····· 80

3.3 总结 ····· 80

第4章 数据分析——回归分析 ····· 81

4.1 功能包 ····· 81

4.1.1 简单回归 ····· 81

4.1.2 多次回归 ····· 88

4.1.3 多变量回归分析 ····· 94

4.1.4 稳健回归 ····· 100

4.2 问题 ····· 106

4.3 总结 ····· 106

第5章 数据分析——相关性 ····· 107

5.1 功能包 ····· 107

5.1.1 基本相关性 ····· 108

5.1.2 可视化相关性 ····· 112

5.1.3 协方差 ····· 114

5.1.4 皮尔森相关性 ····· 117

5.1.5 多分格相关性 ····· 118

5.1.6 四分相关性 ····· 122

5.1.7 异构相关矩阵 ····· 126

5.1.8 部分相关性 ····· 128

5.2 问题 ····· 129

5.3 总结 ····· 129

第6章 数据分析——聚类 ····· 131

6.1 功能包 ····· 131

6.2 K-means 聚类 ····· 132

6.2.1 示例 ····· 132

6.2.2 Medoids 集群 ····· 140

6.2.3 cascadeKM 函数 ····· 142

6.2.4 基于贝叶斯定理信息选取集群 …… 144
6.2.5 仿射传播聚类 …… 146
6.2.6 用于估测集群数量的间隙统计量 …… 149
6.2.7 分级聚类 …… 151
6.3 问题 …… 153
6.4 总结 …… 154
第 7 章 数据可视化——R 图形 …… 155
7.1 功能包 …… 155
7.1.1 交互式图形 …… 156
7.1.2 latticist 功能包 …… 160
7.1.3 ggplot2 功能包 …… 169
7.2 问题 …… 180
7.3 总结 …… 181
第 8 章 数据可视化——绘图 …… 183
8.1 功能包 …… 183
8.2 散点图 …… 183
8.2.1 回归线 …… 187
8.2.2 lowess 线条 …… 188
8.2.3 scatterplot 函数 …… 189
8.2.4 Scatterplot 矩阵 …… 192
8.2.5 密度散点图 …… 197
8.3 直方图和条形图 …… 200
8.3.1 条形图 …… 200
8.3.2 直方图 …… 203
8.3.3 ggplot2 …… 203
8.3.4 词云 …… 204
8.4 问题 …… 206
8.5 总结 …… 206
第 9 章 数据可视化——三维 …… 207
9.1 功能包 …… 207
9.2 生成三维图形 …… 208
9.2.1 Lattice Cloud——三维散点图 …… 212
9.2.2 scatterplot3d …… 215

9.2.3 scatter3d ····· 216
9.2.4 cloud3d ····· 218
9.2.5 RgoogleMaps ····· 220
9.2.6 vrmlgenbar3D ····· 221
9.2.7 大数据 ····· 223
9.2.8 研究方向 ····· 228
9.3 问题 ····· 234
9.4 总结 ····· 234
第 10 章 机器学习实战 ····· 235
10.1 功能包 ····· 235
10.2 数据集 ····· 236
10.2.1 数据划分 ····· 240
10.2.2 模型 ····· 241
10.2.3 train 方法 ····· 254
10.3 问题 ····· 264
10.4 总结 ····· 264
第 11 章 用机器学习预测事件 ····· 265
11.1 自动预测功能包 ····· 265
11.1.1 时间序列 ····· 266
11.1.2 SMA 函数 ····· 272
11.1.3 分解函数 ····· 273
11.1.4 指数平滑法 ····· 274
11.1.5 预测 ····· 277
11.1.6 霍尔特指数平滑法 ····· 281
11.2 问题 ····· 293
11.3 总结 ····· 293
第 12 章 监督学习和无监督学习 ····· 295
12.1 功能包 ····· 296
12.1.1 监督学习 ····· 296
12.1.2 无监督学习 ····· 316
12.2 问题 ····· 327
12.3 总结 ····· 327

第 1 章
模式的数据挖掘

数据挖掘常用于检测数据中的模式或规则。

兴趣点在于仅能够通过使用大数据集进行检测的不明显模式。一段时间内可以检测更简易的模式，如用于购买关联或时间选择的购物篮分析。我们对 R 编程的兴趣在于检测意外的关联，这能够带来新的机会。

某些模式本质上是有序的，例如，基于以往结果预测系统中的故障，通过使用大数据集，以往结果会更加明确。下一章会探讨相关内容。

本章探讨使用 R 来发现数据集不同方法中的模式。

- **聚类分析**：这是检测数据和建立多组相似数据点的过程。通过运用多种算法可以进行聚类分析。不同的算法着重使用数据分布的不同属性，如点间的距离、密度或统计范围。
- **异常检测**：这是着眼于表面相似但某些属性存在差异或异常的数据的过程。异常检测经常用于执法机关、欺诈检测及保险索赔等领域。
- **关联规则**：这是一组可以根据数据做出的决策。我们会寻找具体的步骤，这样，如果我们找到数据点，就可以运用规则确定是否可能存在另一个数据点。此规则经常用于购物篮方法。在数据挖掘中，我们寻找存在于数据中更深层次的不明显规则。

1.1 聚类分析

通过运用各种算法（如下表列举的一些算法）可进行聚类分析。

模 型 类 别	模型如何工作
连通性	此模型计算了点间的距离，并且基于距离远近对点进行组织
分割	此模型将数据分割成集群，并且将每个数据点与一个集群结合起来，最主要的集群是 K-means
分布模型	此模型使用了统计分布以便确定集群
密度	此模型确定了数据点的紧密度以便到达分布的密集区。DBSCAN 常用于密集的分布，OPTICS 常用于更加稀疏的分布

在算法中，同样也有更高级别的粒度，其中包括硬聚类或软聚类等。

- **硬聚类或软聚类**：用于界定数据点是否可以成为一个以上集群的一部分。
- **分割规则**：用于确定如何向不同分区分配数据点，规则如下。
  - **严密**：此规则检查分区是否包括不紧密的数据点。
  - **重叠**：此规则检查分区是否以任何方式重叠。
  - **分层**：此规则检查分区是否分层。

在 R 编程中，聚类工具用于：

- K-means 聚类；
- K-medoids 聚类；
- 分层聚类；
- 期望最大化；
- 密度估计。

试读结束：需要全本请在线购买： www.ertongbook.com