R Data Analysis Projects

R语言
数据分析项目开发实战

[印] 戈皮 · 萨博拉曼尼 著　　杨崇珉 译

清华大学出版社

R语言数据分析项目开发实战

[印] 戈皮·萨博拉曼尼 著
杨崇珉 译

清华大学出版社
北 京

内 容 简 介

本书详细阐述了与数据分析相关的基本解决方案，主要包括关联规则挖掘、基于内容的模糊逻辑推荐系统、协同过滤机制、基于深度神经网络的时序数据、Twitter 文本情感分类、记录链接——随机和机器学习方案、流式数据聚类分析、分析并理解网络等内容。此外，本书还提供了相应的示例、代码，以帮助读者进一步理解相关方案的实现过程。

本书既可作为高等院校计算机及相关专业的教材和教学参考书，也可作为相关开发人员的自学教材和参考手册。

北京市版权局著作权合同登记号　图字：01-2018-7421

图书在版编目（CIP）数据

R 语言数据分析项目开发实战 /（印）戈皮 • 萨博拉曼尼（Gopi Subramanian）著；杨崇珉译. —北京：清华大学出版社，2019

书名原文：R Data Analysis Projects

ISBN 978-7-302-53364-1

I. ①R…　II. ①戈…　②杨…　III. ①程序语言-程序设计　IV. ①TP312

中国版本图书馆 CIP 数据核字（2019）第 168237 号

责任编辑：贾小红
封面设计：刘　超
版式设计：文森时代
责任校对：马军令
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市君旺印务有限公司
经　　销：全国新华书店
开　　本：185mm×230mm　　**印　　张：**20　　**字　　数：**408 千字
版　　次：2019 年 9 月第 1 版　　**印　　次：**2019 年 9 月第 1 次印刷
定　　价：109.00 元

产品编号：081936-01

译 者 序

近几年来，随着计算机和新一代信息技术的蓬勃发展，商业大数据也呈爆炸性地增长。在商业领域，各个行业、企业或组织都遇到了前所未有的全球化、区域化或细分市场多元化的挑战和机遇，他们在激烈竞争中对生存和成长的需求也推动了探索和研究大数据的发展。如何有效地处理、分析和应用这些大数据，已成为当今各个商业领域的迫切需求，相应地也使数据分析师变得炙手可热。

本书向读者介绍了如何利用 R 数据包处理数据分析等问题，其中包含了针对各类数据分析的不同 R 数据包的功能，并帮助读者使用正确的数据包实现相关任务。其中，每章将从头开始构建一个完整项目，进而帮助读者更好地理解如何构建端到端的预测分析解决方案，包括利用 R 语言构建深度学习网络、流数据分析、情绪分类以及推荐系统。

在本书的翻译过程中，除杨崇珉外，刘璋、刘晓雪、张博、刘祎、张华臻等人也参与了部分翻译工作，在此一并表示感谢。

由于译者水平有限，难免有疏漏和不妥之处，恳请广大读者批评指正。

译 者

前　言

本书向读者介绍了如何利用 R 数据包处理数据分析等问题，其中包含了针对各类数据分析的不同 R 数据包的功能，并帮助读者使用正确的数据包实现相关任务。其中，每章将从头开始构建一个完整项目，进而帮助读者更好地理解如何构建端到端的预测分析解决方案。本书涵盖了不同的主题，包括利用 R 语言构建深度学习网络、流数据分析、情绪分类以及推荐系统。

本书内容

第 1 章：关联规则挖掘。通过事务处理数据构建推荐系统，并对关联销售和促销行为进行识别。

第 2 章：基于内容的模糊逻辑推荐系统。将处理推荐系统中的“冷启动”问题，并尝试通过模糊集方案解决包含多相似度的排名问题。

第 3 章：协同过滤机制。将对推荐系统的协同过滤机制引入不同的方案。

第 4 章：基于深度神经网络的时序数据。讨论 MXNet R，即 R 语言中的深度学习数据包。本章将利用 MXNet 构建一个深连接网络，进而预测股票收盘价格。

第 5 章：Twitter 文本情感分类。将考查 R 语言中 Twitter 数据的处理能力，并引入了一种全新的情绪分类方式，即 Delta Tfidf。除此之外，还将利用基于朴素贝叶斯算法的核密度估计对情绪进行分类。

第 6 章：记录链接——随机和机器学习方案。主要讨论数据管理方面的问题，以及如何利用 recordLinkage 数据包在 R 语言中对此加以处理。

第 7 章：流式数据聚类分析。将解决 R 语言中的流数据处理、流数据集群、在线/离线集群模型等问题。

第 8 章：分析并理解网络。通过 igraph 数据包执行 R 中的图分析，同时利用图算法解决产品网络分析等问题。

软件环境

首先需要安装 R。此外，本书代码采用 R version 3.3.1（single candle 版本）并在 Mac

OS 达尔文 15.6.0 环境下编写，同时兼容于 Linux 和 Windows 操作系统。R 代码的编写和编译则在 RStudio version 0.99.491 编辑器中完成。

适用读者

本书将引领读者利用 R 语言以及高级、高效的数据分析方法解决实际应用问题。因此，读者应了解一些与 R 语言和数据分析相关的一些基本概念。

本书约定

本书通过不同的文本风格区分相应的信息类型。下面通过一些示例对此类风格以及具体含义的解释予以展示。

命令行输入或输出如下所示：

```
zero.matrix.gup <- mx.nd.zeros(c(3,3), mx.gpu(0))
```

图标则表示较为重要的说明事项。

图标则表示提示信息和操作技巧。

读者反馈和客户支持

欢迎读者对本书的建议或意见予以反馈。

对此，读者可向 feedback@packtpub.com 发送邮件，并以书名作为邮件标题。若读者对本书有任何疑问，均可发送邮件至 questions@packtpub.com，我们将竭诚为您服务。

若读者针对某项技术具有专家级的见解，抑或计划撰写书籍或完善某部著作的出版工作，则可访问 www.packtpub.com/authors。

资源下载

读者可访问 http://www.packtpub.com 并通过个人账户下载示例代码文件。另外，在 http://www.packtpub.com/support 中注册成功后，我们将以电子邮件的方式将相关文件发与读者。

读者可根据下列步骤下载代码文件：

- 利用电子邮件地址和密码登录或注册我们的网站。

- 单击 SUPPORT 选项卡。
- 单击 Code Downloads & Errata。
- 在 Serach 文本框中输入书名。
- 搜索需要下载代码文件的书名。
- 从下拉菜单中选择本书的购买方式。
- 单击 Code Download 按钮。

当文件下载完毕后，确保使用下列最新版本软件解压文件夹：

- Windows 系统下的 WinRAR/7-Zip。
- Mac 系统下的 Zipeg/iZip/UnRarX。
- Linux 系统下的 7-Zip/PeaZip。

另外，读者还可访问 GitHub 获取本书的代码包，对应网址为 https://github.com/PacktPublishing/R-Data-Analysis-Projects。此外，读者还可访问 https://github.com/PacktPublishing/ 以了解丰富的代码和视频资源。

勘误表

尽管我们在最大程度上做到尽善尽美，但错误依然在所难免。如果读者发现谬误之处，无论是文字错误抑或是代码错误，还望不吝赐教。对此，读者可访问 http://www.packtpub.com/submit-errata，选取对应书籍，单击 Errata Submission Form 超链接，并输入相关问题的详细内容。

版权须知

一直以来，互联网上的版权问题从未间断，Packt 出版社对此类问题异常重视。若读者在互联网上发现本书任意形式的副本，请告知网络地址或网站名称，我们将对此予以处理。关于盗版问题，读者可发送邮件至 copyright@packtpub.com。

问题解答

若读者对本书有任何疑问，均可发送邮件至 questions@packtpub.com，我们将竭诚为您服务。

目　录

第 1 章　关联规则挖掘

本章所讨论的内容其复杂度呈现逐渐上升之势。首先，我们将介绍一些用例，并针对零售商设计关联销售营销方案。随后，将对这一营销方案定义相关目标和标准。接下来，本章将考查第一个推荐算法，即关联规则挖掘算法。关联规则挖掘也称作购物篮（market basket）分析，是一个用于分析事务数据并提取产品关联的方法。

后续各小节将探讨关联规则的普通版本，并引入一些与兴趣度相关的内容。包括设置最小支持度、置信度阈值、两种主要的兴趣度，以及关联规则挖掘算法的参数。最后，本章还将介绍其他一些兴趣度问题，例如提升度和置信度，并以此针对零售商的关联营销方案生成推荐系统。

除此之外，本章还将讨论关联挖掘算法的变化版本，即加权关联规则挖掘算法，并以加权事务的形式整合一些零售商的输入数据。其中，事务的盈利能力视作一个权值。除了事务中的商品之外，事务的盈利能力也将被予以记录。随后，我们将得到一种更为智能的算法，进而生成更具盈利能力的产品关联结果。

接下来，本章还将介绍 HITS 算法。如果无法获取零售商的加权输入信息，也就是说，当不存在与事务重要性相关的明显信息时，HITS 提供了一种方法，并可针对事务生成加权（重要性）结果。

最后，本章将讨论关联规则挖掘的一个变化版本，即负关联规则挖掘，该算法可用于获取事务数据库中的反模式。当需要从分析结果中排除特定条目时（例如较低的库存量或其他约束条件），负关联规则挖掘可视为一种最佳方案。在讨论结束后，我们将引入一个 arulesViz 数据包，这是一个包含了图表、图像的 R 数据包，并实现了关联规则的可视化效果。此外，我们还将编写一个小型 Web 应用程序，并通过 R 语言中的 RShiny 数据包生成相应的分析结果。

本章主要涉及以下内容：

- 理解推荐系统。
- 零售商用例和数据。
- 关联规则挖掘。
- 关联市场营销。
- 加权关联规则挖掘。
- 基于超链接分析的主题搜索（HITS）。
- 负关联规则。

- 规则的可视化。
- 封装操作。
- 进一步阅读。

1.1　理解推荐系统

推荐系统或推荐引擎在机器学习算法中较为常见，并广泛地应用于在线零售系统中。根据与用户和商品交互相关的历史数据，推荐系统可产生与用户及其商品偏好相关的有效信息。

在过去的几十年中，无论是在线零售商还是实体店，推荐系统均取得了极大的成功。推荐系统使得零售商摒弃了过去那种团体营销行为，即一组客户接收单一报价。该技术可视为一种革命性的市场营销行为。如今，零售商可针对每位客户提供定制的建议方案，进而显著地增加了客户与零售商之间的粘性关系。

采取这种营销模式，零售商可实现追加销售（向上销售）和关联销售。这里，追加销售是指零售商可向其客户推送高价值的商品；关联销售则是指向客户推荐附加产品。相应地，推荐系统会提供某种经验方法，并针对零售商的追加销售和关联销售有效方案生成推荐结果。

根据坚实的统计和数学计算，零售商可制定相应的量化决策，从而改善其业务行为。今天，推荐系统得到了广泛的应用，并在一些顶级的公司中饰演了重要的角色，其中包括亚马逊、YouTube、Netflix、LinkedIn、Facebook、TripAdvisor 和 IMDb。

基于数据的类型和容量，可构建不同复杂类型和准确度的推荐系统。如前所述，我们将用户及其所购商品的交互结果定义为历史数据。下面将采用这一定义方式在推荐系统环境中展示不同的数据类型。

1.1.1　事务

事务（交易）表示为客户和零售商之间某次访问过程中生成的购买行为。通常情况下，事务数据可涵盖所购买的商品、购买数量、价格、折扣（如果存在）以及时间戳。其间，单一事务可包含多件商品。在某些情况下，可注册与交易用户相关的信息。其中，通过积分奖励机制，零售商可存储某些用户信息。

二元矩阵则是较为简单的事务数据视图，在视图中，矩阵中的各行对应于唯一的事务标识符，这里可将其称作事务 ID。相应地，每列对应于唯一的商品标识符，即商品 ID。另外，二元矩阵中的单元值表示为 0 或 1，代表商品被排除或纳入当前事务中。

表 1.1 显示了包含 n 项事务和 m 件商品的二元矩阵。

表　1.1

Txn/Product	P1	P2	P3	…	Pm
T1	0	1	1	…	0
T2	1	1	1	…	1
…	…	…	…	…	…
Tn	0	1	1	…	1

1.1.2　加权事务

这表示为添加至事务中的附加信息以表示其重要性，例如事务的整体盈利能力，或者事务中单件商品的盈利能力。在前述二元矩阵示例中，加入了权重列进而存储事务的重要程度。

本章将向读者展示如何使用事务数据支持关联销售营销行为，并查看客户的购买偏好，或者是源自客户商品交互行为（事务/加权事务）的推荐内容对成功的营销行为所产生的影响。除此之外，本章还将进一步理解、实现基于此类数据的相关算法，并在某些特定的用例上展开我们的工作，进而生成推荐系统，以对零售商的关联销售营销行为提供支持。

1.1.3　Web 应用程序

图 1.1 展示了本章 Web 应用程序的运行效果。

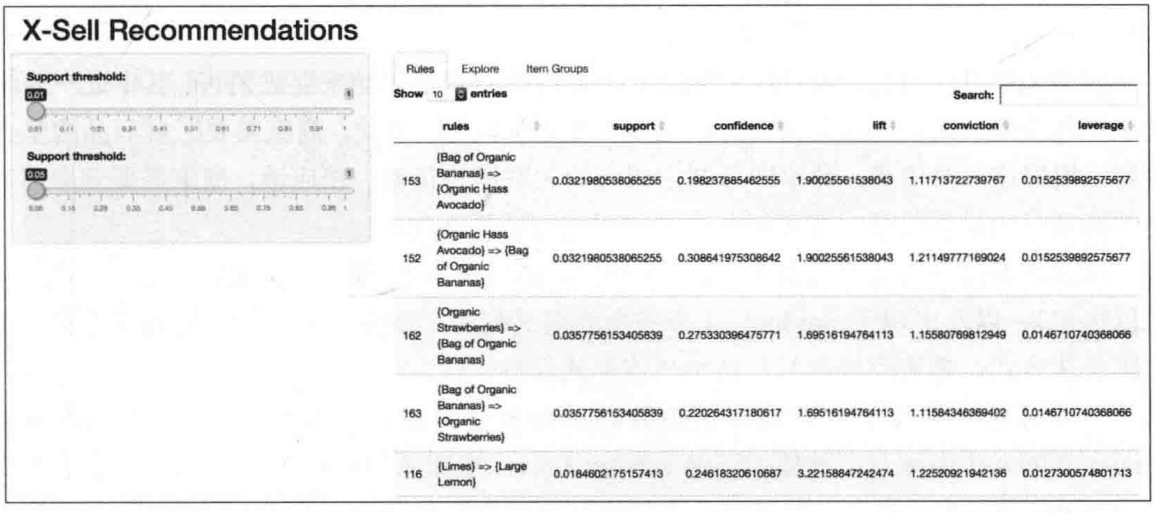

图　1.1

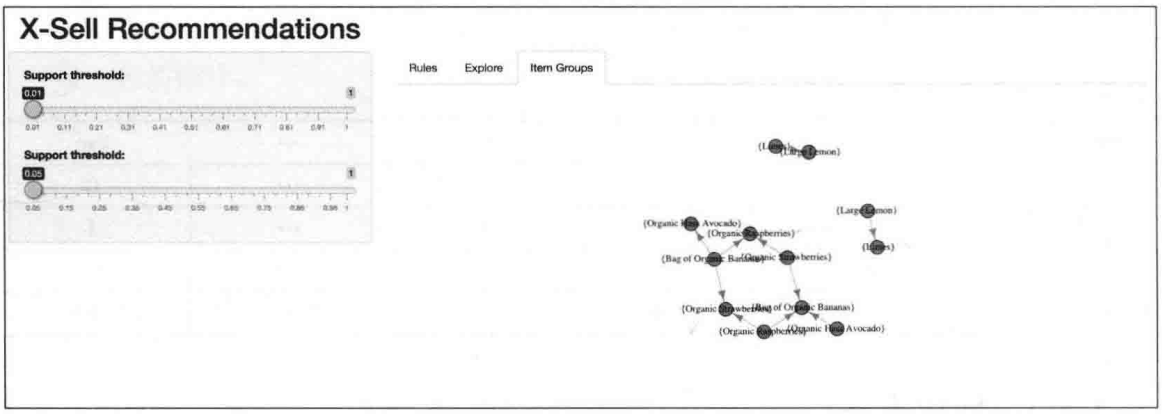

图　1.1（续）

本章的目标是理解关联规则挖掘以及相关主题的各种概念，并利用关联规则挖掘解决给定的关联销售营销问题。此外，我们还将学习如何利用关联规则挖掘算法解决关联销售中的各种问题，以及在 R 语言中的实现方式，最终构建一个 Web 应用程序并显示相应的分析结果。本书的特点是首先引入一个真实的问题，随后将引入处理该问题的相关算法/技术；最后将展示具体的实现代码。具体来说，我们将对相关算法进行简要的描述，随后介绍基于 R 数据包的实现算法、编写 R 代码并以算法期望的方式准备数据。考虑到将调用 R 中的算法并对结果进行分析，因而我们需要了解算法所涉及的具体细节。另外，我们还将向读者提供相关的参考读物以供进一步学习。

1.2　零售商用例和数据

在未来的几个月内，商家计划通过一个大规模的营销活动来促进销售，其中之一便是关联销售策略。关联销售是指向客户推荐一些附加商品，对此，商家需要了解商品的整合内容。根据这一类信息，商家即可制定相应的关联销售策略。相应地，商家需要获取前 N 件关联商品的推荐信息，并可以从中选取以进行营销活动。

商家需要提供历史交易数据，其中涉及以往的交易事务（每项事务通过唯一的 order_id 予以标识），以及出现于 product_id 事务中的商品列表。回忆一下之前讨论的事务的二元矩阵表达方式，零售商提供的数据集与该形式保持一致。

下面开始着手读取零售商提供的数据。需要说明的是，本章中的代码在 RStudio version 0.99.491 中编写，并使用了 R version 3.3.1。当对示例加以考查时，将引入 R 语言中的 arules 数据包。在当前描述中，我们将交替使用订单/事务、用户/客户、条目/商品等

术语。另外，本章并不打算详细介绍 R 数据包的安装过程，假设读者已对此有所了解并已经安装了相关数据包。

数据的下载过程如下所示：

```
data.path = '../../data/data.csv'
  data = read.csv(data.path)
  head(data)
order_id product_id
  1 837080 Unsweetened Almondmilk
  2 837080 Fat Free Milk dairy
  3 837080 Turkey
  4 837080 Caramel Corn Rice Cakes
  5 837080 Guacamole Singles
  6 837080 HUMMUS 10OZ WHITE BEAN EAT WELL
```

上述给定数据以表格格式呈现。其中，每一行表示为 order_id（事务）、product_id（事务中的包含的商品）和 department_id（商品隶属的部门）构成的元组。这体现了一种二元数据表达方式，进而支持经典的关联规则挖掘算法。该算法也称作购物篮分析，也就是说，我们将对客户的购物篮（即事务）进行分析。当设置了客户事务的大型数据库后，其中，每项事务由客户访问过程中所购买的商品构成，关联规则挖掘算法将在数据库的商品之间生成有效的关联规则。

什么是关联规则？下面的示例来自一家零售店的交易事务，关联规则是指形如{peanut butter, jelly} => { bread }的推荐结果。具体而言，根据当前事务，bread 很可能也会出现于 peanut butter 和 jelly 这一项交易事务中。作为向零售商的一类推荐结果，数据库中有足够的证据表明，购买了花生酱和果冻的顾客很可能会购买面包。

下面快速浏览一下当前数据，并统计事务和商品的数量，如下所示：

```
library(dplyr)
data %>%
  group_by('order_id') %>%
  summarize(order.count = n_distinct(order_id))

data %>%
  group_by('product_id') %>%
  summarize(product.count = n_distinct(product_id))
# A tibble: 1 <U+00D7> 2
  `"order_id"` order.count
```

```
  <chr> <int>
  1 order_id 6988
# A tibble: 1 <U+00D7> 2
  `"product_id"` product.count
  <chr> <int>
  1 product_id 16793
```

其中包含了 6988 项事务和 16793 件商品，但并不存在与事务中所购买商品数量相关的信息。此处使用了 dplyr 库执行此类聚合计算，该库通常用于执行数据帧（data frame）上高效的数据整理工作。

dplyr 是 tidyverse 中的部分内容，同时也是一个围绕相同理念而设计的 R 数据包集合。dplyr 定义为一类数据操控语法，提供了一致的方法集，进而可解决常见的数据操控问题。关于 dplyr，读者可访问 http://tidyverse.org/和 http://dplyr.tidyverse.org/以了解更多内容。

1.3 节将讨论关联规则挖掘算法，以及如何利用该算法生成前 *N* 件商品推荐结果，以满足零售商的关联销售营销活动。

1.3　关联规则挖掘

关联规则挖掘包含多种算法实现。其中，较为重要的是 Rakesh Agrawal 和 Ramakrishnan Srikanth 在其论文（*Fast Algorithms for Mining Association Rules*）中提出的 Apriori 算法。在后续章节中，我们将交替使用 Apriori 算法和关联规则挖掘算法。

Apriori 是一种参数化算法，需要使用到源自用户的名为 support 和 confidence 的两个参数。这里，support 用于生成频繁项目集；而 confidence 参数将从频繁项目集中过滤诱导规则。support 和 confidence 一般称作兴趣度。除了 support 和 confidence 之外，还存在很多其他形式的兴趣度。

当编写 R 代码时，还将进一步解释关联规则挖掘，以及算法中的兴趣度效果；同时深入理解代码的工作方式，包括项目集等算法技术，以及如何利用兴趣度以支持关联销售营销活动。

此处将使用 arules 数据包（版本为 1.5-0），并在数据集上执行关联挖掘操作，如下所示：