第 1 章 简　介

R[⊖]是一种用于统计计算的编程语言和环境（R Core Team，2015b），它与由 AT&T 贝尔实验室的 Rick Becker、John Chambers 和 Allan Wilks 开发的 S 语言很相似。根据操作系统的不同，R 语言有 UNIX 版本、Windows 版本和 Mac OS 版本。R 可以在不同体系结构的计算机系统上运行，例如 Intel、PowerPC、Alpha 系统以及 Sparc 系统。其最早由新西兰奥克兰大学的 Ihaka 和 Gentleman 于 1996 年开发。而现在 R 的开发由一个几十人组成的核心团队来负责，核心团队的成员来自世界各地的不同机构和单位，并且得到了 R 基金会的支持。由于 R 的开源性，R 软件的开发采用社区合作的方式。R 的所有源代码都可以免费获取以便进行检验或者利用，这样你就可以检查和测试你应用的 R 软件的可靠性，并且这种能力在许多关键应用领域可能至关重要。人们对开源软件模式有诸多指责，其中最多的是认为开源软件缺少支持是其主要缺陷之一。但是对 R 软件来说，该缺陷却不存在！有许多很好的文档、书籍和网站提供了免费的 R 资料。另外，R 帮助邮件列表是获取免费 R 帮助信息和建议的极好来源。R 有一个可搜索的邮件列表文档。在邮件列表中提问前，可以先查找这个可搜索的邮件列表文档（也应该这样做！）。可以在 R 网站的“Mailing Lists”（邮件列表）部分得到有关 R 邮件列表的更多信息。

数据挖掘是应用统计学、机器学习和模式识别等学科的知识，从数据中发现有用的、有效的、未知的并且可以理解的信息的一项技术。数据挖掘的一项重要特征是数据的维度。随着计算机技术和信息系统的广泛应用，需要探索的数据呈指数增长。这给传统的数据分析学科带来了挑战：必须考虑计算的效率、内存资源的限制、数据库接口等。其他关键的特征是数据挖掘项目中经常遇到的数据源的多样性，以及数据类型（文本、声音和视频等）的多样性。这些使得数据挖掘成为一门高度交叉的学科，它不仅涉及传统数据分析，还包括数据库工作和高维数据可视化等。

由于 R 的所有计算是在计算机的内存中进行的，所以 R 在处理大数据集上有限制。但是这并不意味着我们不能处理这些问题，利用 R 高度灵活的数据库接口，我们可以对大型问题进行数据挖掘。此外，R 社区对数据集大小不断增加的认识导致许多新的 R 添加包的开发，其设计是用于处理大数据，或者是向其他基础设施提供更适合于繁重计算任务的接口。有关此项工作的更多信息，请参见 R 任务视图中的高性能和并行计算[⊖]。

⊖ http://www.r-project.org。

⊖ http://cran.at.r-project.org/web/views/HighPerformanceComputing.html。

总之，我们希望在读完本书之后，你能够相信不用花钱就可以进行大型问题的数据挖掘。这一切都归功于开发出 R 这样优秀软件的人们的慷慨贡献。

1.1 如何阅读本书

本书基于以下宗旨：

做中学

本书的第一部分提供了一些关于 R 和数据挖掘的基本信息，而第二部分则是由一系列案例研究组成的。本书描述了所有得到“解决方案”的必要步骤。通过本书提供的网站⊖，可以获取本书有关的 R 添加包（DMwR2），所有的 R 代码和案例研究数据都在相关的文档中。这能方便你自己进行实验。理论上，你应该在电脑上阅读这些文档并且尝试这些文档中演示的每一个步骤。在这本书里，R 代码是用以下字体来表示的：

```
> citation()


To cite R in publications use:

  R Core Team (2016). R: A language and environment for
  statistical computing. R Foundation for Statistical Computing,
  Vienna, Austria. URL https://www.R-project.org/.

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {R: A Language and Environment for Statistical Computing},
    author = {{R Core Team}},
    organization = {R Foundation for Statistical Computing},
    address = {Vienna, Austria},
    year = {2016},
    url = {https://www.R-project.org/},
  }

We have invested a lot of time and effort in creating R, please
cite it when using it for data analysis. See also
'citation("pkgname")' for citing R packages.
```

在 R 的命令提示符“>”之后输入 R 的指令。当看到这个提示符时，你可以理解为 R 在等待输入指令。在命令提示符后键入命令，然后按下 Enter 键让 R 来执行它们。这可能会产生某种形式的输出（即 R 命令的结果），然后出现一个新的提示符。在提示符处，你可以使用箭头键来浏览和编辑以前输入过的命令。若以前输入了类似的命令，通过编辑以前的命令可以方便地得到你需要的命令，这就避免了重复输入。

2

可以复制、粘贴本书网站提供的代码到 R 编译器或者 R 控制台，这样就可以避免你自己键入书中的那些代码，这绝对会使你的学习更加便捷，并加深你对书中知识的理解。

1.2 重现性

本书的主要目标之一是为你提供说明如何使用 R 提供的工具处理多个数据挖掘任务的

⊖ http://ltorgo.github.io/DMwR2。

示例。为了使这成为可能，我们努力确保书中描述的所有情况都可以使读者在自己的电脑上重现。这意味着如果你遵循我们在书中描述的所有步骤，那么应该得到与我们的描述相同的结果。

这种重现性目标有两个重要组成部分：所使用的 R 代码；案例研究的数据。随附这本书，我们提供了另外两种访问代码和数据的方法：本书网站；本书 R 包。与本书中包含的描述一起，网站和添加包允许你轻松地复制我们所描述的内容，并重新使用或使其适应你自己的应用领域。

本书网站[⊖]以复制 / 粘贴的友好方式提供对本书中使用的所有代码的访问，以便你可以轻松地将其从浏览器复制到 R 中。代码按章节顺序组成，以便于查找任务。

网站还包含其他有用的信息，如我们使用的添加包的列表或数据集，以及其他包含本书中创建的一些对象的文件，特别是对比平均台式机需要花费更长的时间来计算的情况。

R 是一个非常动态的“生态系统”。这意味着当你阅读本书时，最常见的一些我们使用的添加包（甚至是 R 本身）已经有了新的版本。尽管这样做很可能不会造成任何问题，但我们无法确定这些代码是否仍然可以与新版本一起使用。如果由于新版本而导致某些任务停止工作，我们将尝试在本书网站的“ Errata”（勘误）部分快速发布解决方案。本书和其中的 R 代码在以下 R 版本中创建和测试：

```
> R.version

               _
platform       x86_64-apple-darwin13.4.0
arch           x86_64
os             darwin13.4.0
system         x86_64, darwin13.4.0
status
major          3
minor          3.1
year           2016
month          06
day            21
svn rev        70800
language       R
version.string R version 3.3.1 (2016-06-21)
nickname       Bug in Your Hair
```

3

在本书网站上，你还可以在代码执行时找到 R 中所有使用软件包版本的信息。

本书的 R 添加包是实现重现性的另一个关键因素。这个添加包包含了我们在本书中描述和使用的几个函数，以及案例研究中的数据集（我们上面提到的数据集也可以在本书网站中找到）。该添加包可从常规来源（即 R 中央存储库（ R central repository，CRAN））安装。如果在我们提供的代码中发现任何错误，则可能会将包演化为新版本。根据 CRAN 政策的建议，这些更正将趋于缓慢。在这方面，对于更新的添加包版本，可能包括尚未经过良好测试的解决方案（因此需谨慎使用），你可能希望从 https://github.com/ltorgo/DMwR2 下载并安装添加包的开发版本。

4

⊖ http://ltorgo.github.io/DMwR2。