HZ BOOKS
华章科技

10余位数据挖掘领域资深专家和科研人员，10余年大数据挖掘咨询与实施经验结晶。本书注重易用性和实践性，旨在让读者快速掌握运用Python语言进行数据分析与挖掘的方法，从应用层面讲解初学者最急切需要了解的功能，深入浅出地介绍了数据挖掘中常用的建模实现函数

大数据技术丛书

Python and Data Minning

Python与数据挖掘

张良均 杨海宏 何子健 杨征◎等著

机械工业出版社
China Machine Press

大数据技术丛书

Python and Data Minning

Python与数据挖掘

张良均 杨海宏 何子健 杨坦 杨征 陈婷婷 陈玉辉 施兴◎等著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Python 与数据挖掘 / 张良均等著 . —北京：机械工业出版社，2016.11
（大数据技术丛书）

ISBN 978-7-111-55261-1

I. P… II. 张… III. 软件工具 – 程序设计 IV. TP311.56

中国版本图书馆 CIP 数据核字（2016）第 260795 号

Python 与数据挖掘

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：李 艺　　责任校对：殷 虹

印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2016 年 11 月第 1 版第 1 次印刷

开　　本：186mm × 240mm 1/16　　印　　张：11.75

书　　号：ISBN 978-7-111-55261-1　　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）88379426 88361066　　投稿热线：（010）88379604

购书热线：（010）68326294 88379649 68995259　　读者信箱：hzit@hzbook.com

Preface 前　言

为什么要写本书?

Python 是什么?

Python 是一种带有动态语义的、解释性的、面向对象的高级编程语言。其高级内置数据结构，结合动态类型和动态绑定，使其对于敏捷软件开发非常具有吸引力。同时，Python 作为脚本型（胶水）语言连接现有的组件也十分高效。Python 语法简洁，可读性强，从而能降低程序的维护成本。不仅如此，Python 支持模块和包，鼓励程序模块化和代码重用。

Python 语言的解释性使其语法更接近人类的表达和思维过程，开发程序的效率极高。习惯使用 Python 者，总习惯在介绍 Python 时强调一句话："人生苦短，我用 Python。"由于没有编译步骤，"写代码—测试—调试"的流程能被快速地反复执行。

作为一款用途广泛的语言，Python 在数据分析与机器学习领域的表现，称得上"一任群芳妒"。2016 年 3 月，国外知名技术问答社区 StackOverflow 发布了《2016 年开发者调查报告》。此调查号称是有史以来最为全面的开发者调查。其中，数据科学家的十大技术栈中，有 7 个包含 Python。具体来说，数据科学家中有 63% 正在使用 Python，44% 正在使用 R 语言。而且，27% 的人同时使用这两种语言。Python 还在"最多人使用的技术""最受欢迎技术""需求度最高技术"等榜单中名列前十。

Python 的明显优势：

- Python 作为一款优雅、简洁的开源编程语言，吸引了世界各地顶尖的编程爱好者的注意力。每天都有数量众多的开源项目更新自己的功能，作为第三方模块为其他开发者提供更加高效、便利的支持。
- Python 提供了丰富的 API 和工具，以便程序员能够轻松地使用 C、C++、Cython 来编写扩充模块，从而集成多种语言的代码，协同工作。一些算法在底层用 C 实现后，封装在 Python 模块中，性能非常高效。
- Python 受到世界各地开发者的一致喜爱，在世界范围内被广泛使用。这意味着读者可以通过查看代码范例，快速学习和掌握相关内容。
- Python 语言简单易学，语法清晰。Python 开发者的哲学是“用一种方法，最好是只有一种方法来做一件事”。通常，相较其他语言，Python 的源代码被认为具有更好的可读性。

2004 年，Python 已在 Google 内部使用，他们的宗旨是：Python where we can，C++ where we must，即在操控硬件的场合使用 C++，在快速开发时使用 Python。

总的来说，Python 是一款用于数据统计、分析、可视化等任务，以及机器学习、人工智能等领域的高效开发语言。它能满足几乎所有数据挖掘下所需的数据处理、统计模型和图表绘制等功能需求。大量的第三方模块所支持的内容涵盖了从统计计算到机器学习，从金融分析到生物信息，从社会网络分析到自然语言处理，从各种数据库各种语言接口到高性能计算模型等领域。随着大数据时代的来临，数据挖掘将更加广泛地渗透到各行各业中去，而 Python 作为数据挖掘里的热门工具，将会有更多不同行业的人加入到 Python 爱好者的行列中来。完全面向对象的 Python 的教学工作也将成为高校中数学与统计学专业的重点发展对象，这是大数据时代下的必然趋势。

本书特色

笔者从实际应用出发，结合实际例子及应用场景，深入浅出地介绍 Python 开发环境的搭建、Python 基础入门、函数、面向对象编程、实用模块和图表绘制及常用的建模算法在 Python 中的实现方式。本书的编排以 Python 语言的函数应用为主，先介绍了函数

的应用场景及使用格式，再给出函数的实际使用示例，最后对函数的运行结果做出了解释，将掌握函数应用的所需知识点按照实际使用的流程展示出来。

为方便读者理解 Python 语言中相关函数的使用，本书配套提供了书中使用的示例的代码及所用的数据，读者可以从“泰迪杯”全国数据挖掘挑战赛网站（http://www.tipdm.org/ts/755.jhtml）上免费下载。读者也可通过热线电话（40068-40020）、企业 QQ（40068-40020）或以下微信公众号咨询获取。

TipDM

张良均〈大数据挖掘产品与服务〉

本书适用对象

❑ 开设有数据挖掘课程的高校教师和学生。

目前国内不少高校将数据挖掘引入本科教学中，在数学、计算机、自动化、电子信息、金融等专业开设了数据挖掘技术相关的课程，但目前这一课程的教学使用的工具仍然为 SPSS、SAS 等传统统计工具，并没有使用 Python 作为教学工具。本书提供了有关 Python 语言的从安装到使用的一系列知识，将能有效指导高校教师和学生使用 Python。

❑ 数据挖掘开发人员。

这类人员可以在理解数据挖掘应用需求和设计方案的基础上，结合本书提供的 Python 的使用方法快速入门并完成数据挖掘应用的编程实现。

❑ 进行数据挖掘应用研究的科研人员。

许多科研院所为了更好地对科研工作进行管理，纷纷开发了适应自身特点的科研业务管理系统，并在使用过程中积累了大量的科研信息数据。Python 可以提供一个优异的环境对这些数据进行挖掘分析应用。

❑ 关注高级数据分析的人员。

Python 作为一个广泛用于数据挖掘领域的编程语言，能为数据分析人员提供快速的、可靠的分析依据。

如何阅读本书

本书主要分为两大部分，基础篇和建模应用篇。基础篇介绍了有关 Python 开发环境的搭建、Python 基础入门、函数、面向对象编程、实用模块和图表绘制等基础知识。建模应用篇主要介绍了目前数据挖掘中常用的建模方法在 Python 中的实现函数，并对输出结果进行了解释，有助于读者快速掌握应用 Python 进行分析挖掘建模的方法。读者可结合本书提供的示例代码及数据进行上机实验，快速掌握书中所介绍的 Python 的使用方法。

第一部分是基础篇（1 ～ 6 章）。第 1 章旨在让读者从全局把握数据挖掘、常用工具对比、Python 开发环境的搭建以及本书的写作习惯；第 2 章正式开始讲解 Python 的基础知识，包括操作符、变量类型、流程控制、数据结构等内容；第 3、4 章主要对 Python 面向对象的特性进行介绍，包括函数、类与对象等基本概念；第 5 章介绍主流的数据分析与挖掘的模块，以及其中具体的方法及对应的功能，旨在让读者对各个模块建立强大的直觉；第 6 章继续拓展了模块的相关内容，提及图表绘制的专用模块（Matplotlib 和 Bokeh），深入浅出地展示如何方便地绘制点、线、图等。

第二部分是建模应用篇（7 ～ 11 章）。本部分主要对数据挖掘中的常用算法进行介绍，强调在 Python 中对应函数的使用方法及其结果的解释说明。内容涵盖五大主流的数据挖掘算法，包括分类与预测、聚类分析建模、关联规则分析、智能推荐和时间序列分析。按照模型建立至模型评价的架构进行介绍，使读者能熟练掌握从建模到对模型评价的完整建模过程。

勘误和支持

除封面署名外，参加本书编写工作的还有杨坦、刘名军、陈婷婷、陈玉辉、施兴、

黄博、王路、黄东鑫等。由于水平有限，编写时间仓促，书中难免会出现一些错误或者不准确的地方，恳请读者批评指正。本书内容的更新将及时在“泰迪杯”全国数据挖掘挑战赛网站（www.tipdm.org）上发布。读者可通过作者微信公众号 TipDM（微信号：TipDataMining）、TipDM 官网（www.tipdm.com）反馈有关问题。也可通过热线电话（40068-40020）或企业 QQ（40068-40020）进行咨询。

如果您有更多的宝贵意见，欢迎发送邮件至邮箱 13560356095@qq.com，期待能够得到您的真挚反馈。

致谢

本书编写过程中得到了广大高校师生的大力支持！在此谨向华南农业大学、华南师范大学、广东工业大学、广东技术师范学院、华南理工大学、韩山师范学院、中山大学、贵州师范学院等单位给予支持的领导及师生致以深深的谢意。

在本书的编辑和出版过程中还得到了参与“泰迪杯”全国数据挖掘挑战赛（http://www.tipdm.org）的众多师生及机械工业出版社杨福川老师无私的帮助与支持，在此一并表示感谢。

张良均

目录 *Contents*

Part 1

第一部分

基 础 篇

Chapter 1 第 1 章

数据挖掘概述

广义的数据挖掘是指针对收集的大规模数据，应用整套科学工具和挖掘技术（如数据、计算、可视化、分析、统计、实验、问题定义、建模与验证等），从数据之中发现隐含的、对决策有参考意义的信息、价值和趋势。因此，数据挖掘是一个横跨多学科的计算机科学分支。强调它隶属计算机科学范畴，是希望读者认识到这个领域的核心需求，尽早摆脱对编程实现的恐惧，避免陷入“数据挖掘只需将模型或算法套用于数据集之上”的误区。这也是本书的写作目的之一。

1.1 数据挖掘简介

随着计算机技术的全面发展，企业生产、收集、存储和处理数据的能力大大提高，数据量与日俱增。数据的积累实质上是企业的经验和业务的沉淀。越来越多的企业引入“数据思维”——不只是依赖于数据的统计分析，更强调对数据进行挖掘，期待从这一“未来世界的石油”中发现潜在的价值。这一迫切的“开采”需求在世界范围内酝酿了一次“大数据”变革。

数据挖掘的确是 21 世纪最具话题性的技术之一，包含数据预处理、算法应用、模型评价、结果检验等多个部分，并依靠其丰富的内涵向外延伸出数据分析、数据 ETL、机

器学习等多个领域。

1.2 工具简介

数据挖掘软件的历史并不长，甚至连“数据挖掘”这个术语也是在19世纪90年代中期才正式被提出。如今，商用数据挖掘软件和开源工具都已经非常成熟，不仅提供易用的可视化界面，还集成了数据处理、建模、评估等一整套功能。

部分开源的数据挖掘软件，采用可视化编程的设计思路。之所以这么做，是因为它能足够灵活和易用，更适合缺乏计算机科学知识的用户，如WEKA和RapidMiner。

当用户拥有较多特定的分析需求，或正在自行实现一个改进的机器学习算法时，脚本型语言如Python和R将更符合需要。同时，脚本型语言兼具运行效率和开发效率，支持敏捷型的迭代更新。

1.2.1 WEKA

用Java编写的WEKA是一款知名的数据挖掘工作平台，它因解决数据挖掘任务的实际需求而生，集成了大量能处理数据挖掘任务的机器学习算法，这些算法能被用户直接应用于数据集之上。同时，WEKA允许开发者使用Java语言，调用其分析组件，基于WEKA的架构进行二次开发，融入更多的数据挖掘算法，并嵌入到软件或者应用之中，自动完成数据挖掘任务，开发新的机器学习框架。

WEKA支持多种标准数据挖掘任务，包括数据预处理，分类、回归分析、聚类、关联规则等算法的应用，以及特征工程和可视化。其欢迎界面如图1-1所示。

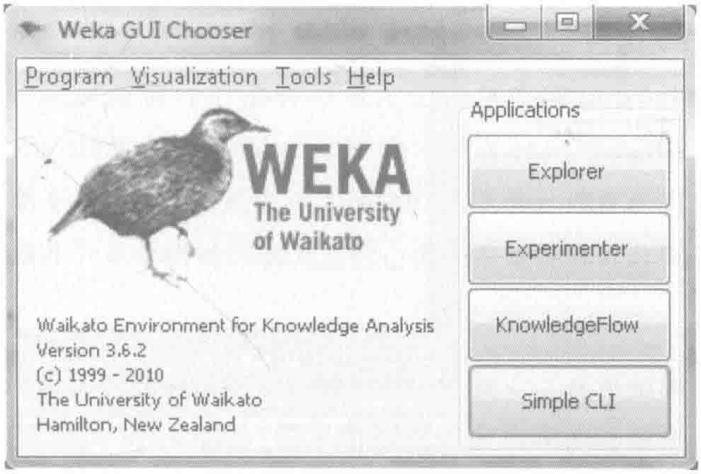

图1-1 WEKA欢迎界面

1.2.2 RapidMiner

RapidMiner 的目标是："成为一个能将数据变成宝贵的战略资产的现代平台"，已被广泛使用于商业应用、学术研究、教育、敏捷开发等领域。

RapidMiner 是一个支持数据挖掘、文本挖掘、机器学习、商业分析等任务的集成环境，如图 1-2 所示。其图形化界面采用了类似 Windows 资源管理器中的树状结构来组织分析组件，提供 500 多种分析组件作为计算单元（Operator），服务于数据挖掘的各个环节，如数据预处理、变换、探索、建模、评估及结果可视化。这些计算单元有详细的 XML 文件记录。

RapidMiner 是基于 WEKA 二次开发的应用，这意味着它可以调用 WEKA 中的各种分析组件。

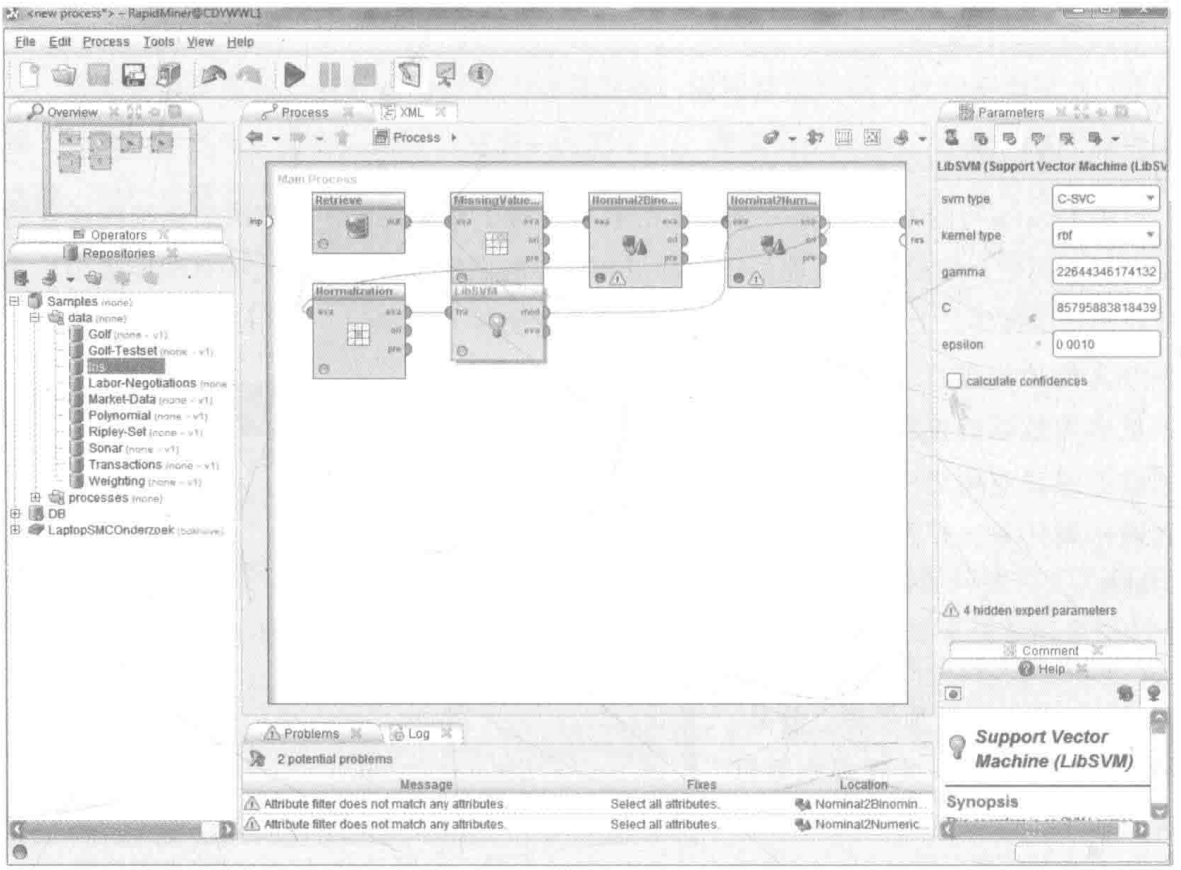

图 1-2 RapidMiner Studio 工作界面

1.2.3 Python

Python是一门编程语言。随着NumPy、SciPy、Matplotlib和Pandas等众多程序库的开发，Python在科学计算和数据分析领域占据着越来越重要的地位。在大多数数据任务上，Python的运行效率已经可以媲美C/C++语言。2016年2月11日，科学家宣布：人类在去年9月首次直接探测到了引力波！引力波高峰只持续了四分之一秒，同时仪器接收了大量干扰噪声，需要处理的数据量以TB计，如图1-3所示。其中，Python的GWPY模块提供专业的数据分析支持。

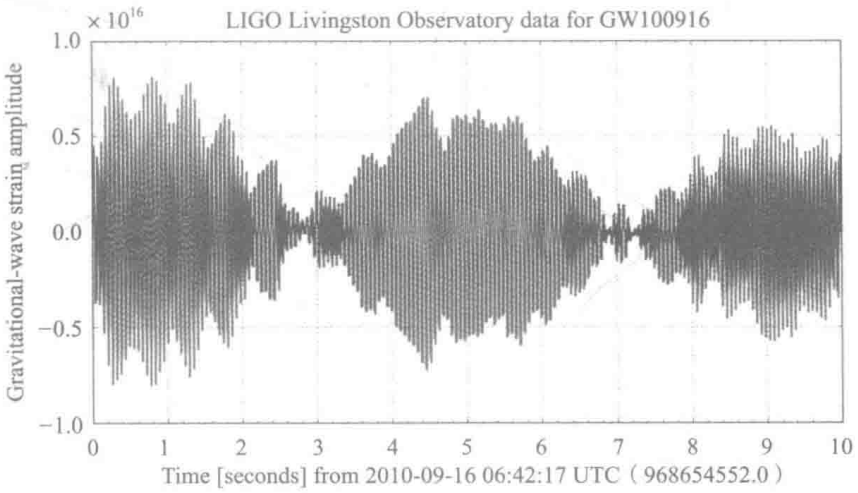

图1-3 利用公开引力波数据绘制波形图

1.2.4 R

R语言是一种为统计计算和图形显示而设计的语言环境，是贝尔实验室（Bell Laboratory）的Rick Becker、John Chambers和Allan Wilks开发的S语言的一种实现，包含一系列统计与图形显示工具，如图1-4所示。它是由一个庞大且活跃的全球性研究型社区维护，主要包括核心的标准包和各个专业领域的第三方包，提供丰富的统计分析和数据挖掘功能。

R语言至少拥有以下优势：①方便地从各种类型的数据源中获取数据；②高可拓展性；③出色的统计计算功能；④顶尖水准的制图功能；⑤不断贡献强大功能的开源社区。它与Python同属数据挖掘主流编程语言，而从功能与代码风格的角度来评价，R与MATLAB是最像的。

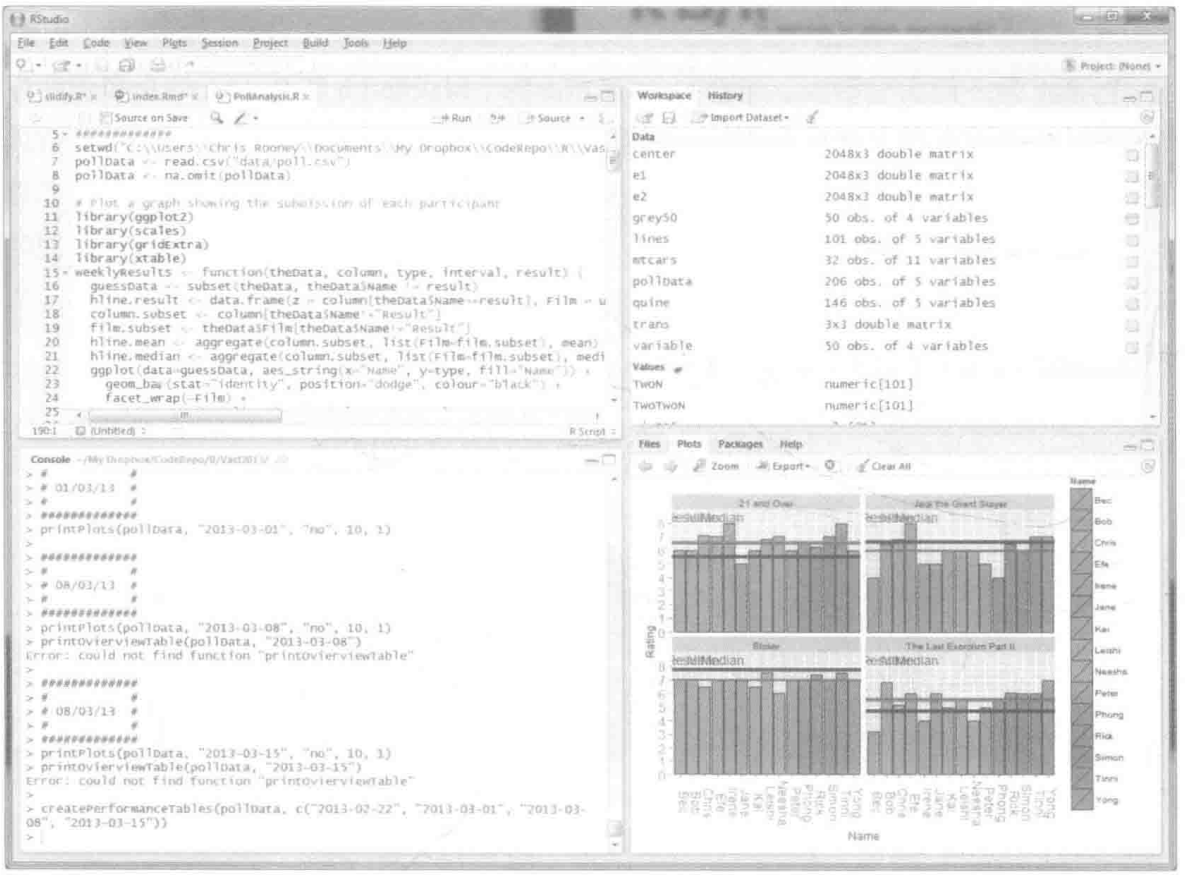

图 1-4 R-Studio 工作界面

1.3 Python 开发环境的搭建

所谓编程语言，意指“与计算机交流时使用的语言”。它是一种被标准化的交流技巧，用于连接程序员的思维和计算机的操作。学习编程语言的第一关，就是安装和环境配置。我们必须与计算机约定如何理解代码、指令和语法，才能够顺利地与计算机交流，赋予它复杂的功能。Python 便是其中的一种“方言”。

本节将向大家详细介绍，如何在不同的操作系统上快捷地使用 Python 进行编程实现。

1.3.1 Python 安装

对于新手，Python 及其第三方模块在安装环节有许多已知的难题。比如源码编译的