HZ BOOKS
华章IT

[PACKT]
PUBLISHING

IBM首席数据科学家亲笔撰写，全方位讲解在Spark上应用机器学习的各种实用技术

提供9个实际案例分析，涵盖整体视图、欺诈检测、风险评分、流失预测、产品推荐、教育分析、城市分析和开放数据建模等方面

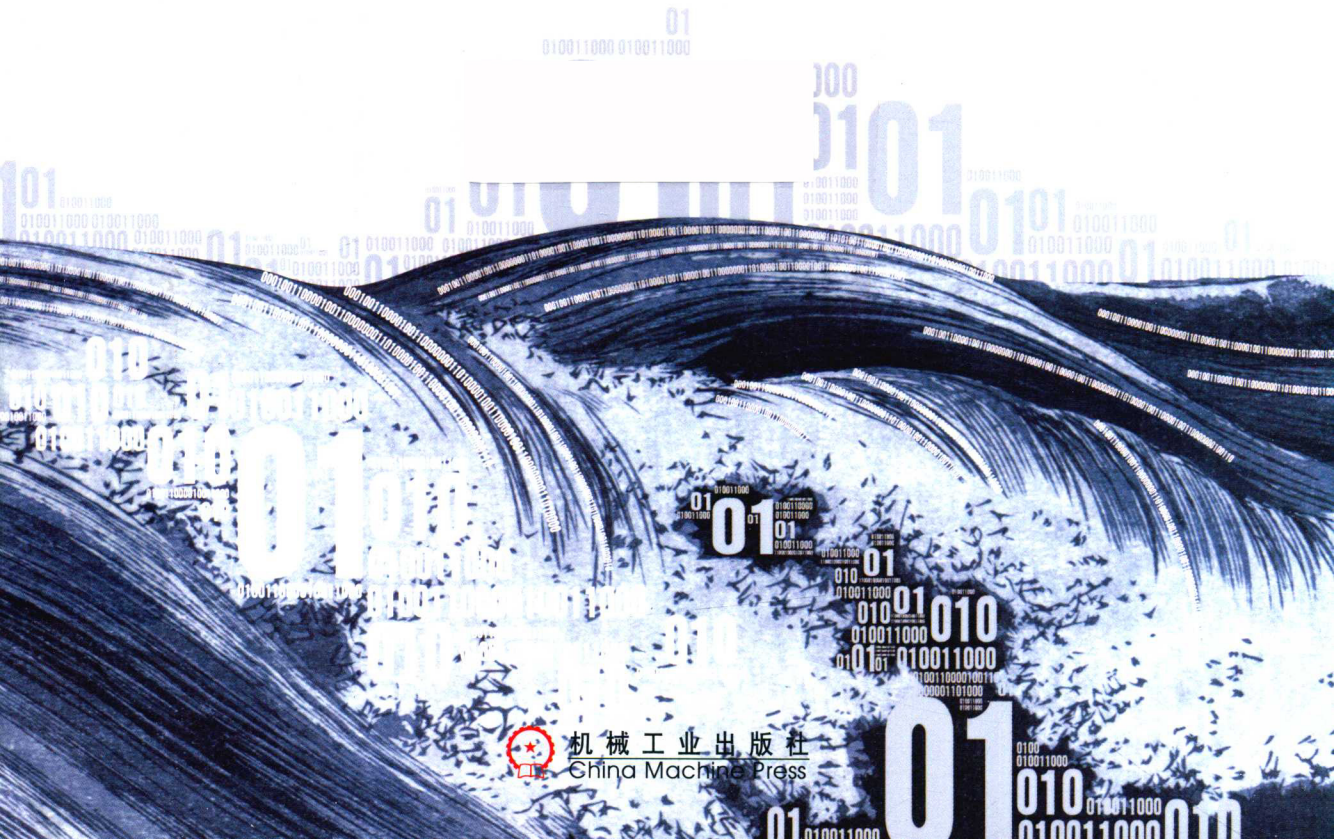
大数据技术丛书

Apache Spark Machine Learning Blueprints

Apache Spark 机器学习

[美] 刘永川（Alex Liu） 著

闫龙川 高德荃 李君婷 译



机械工业出版社
China Machine Press

技术丛书

Apache Spark Machine Learning Blueprints

Apache Spark 机器学习

[美] 刘永川（Alex Liu） 著

闫龙川 高德荃 李君婷 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Apache Spark 机器学习 /（美）刘永川（Alex Liu）著；闫龙川，高德荃，李君婷译 . —北京：机械工业出版社，2017.3
（大数据技术丛书）
书名原文：Apache Spark Machine Learning Blueprints

ISBN 978-7-111-56255-9

I. A… II. ①刘… ②闫… ③高… ④李… III. 数据处理软件 - 机器学习 IV. TP274

中国版本图书馆 CIP 数据核字（2017）第 043165 号

本书版权登记号：图字：01-2016-8649

Apache Spark 机器学习

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：缪　杰　　责任校对：李秋荣
印　　刷：三河市宏图印务有限公司　　版　　次：2017 年 3 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：13.75
书　　号：ISBN 978-7-111-56255-9　　定　　价：59.00 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

The Translator's Words 译 者 序

近年来，大数据发展迅猛，如雨后春笋般出现在各行各业，企业收集和存储的数据成倍增长，数据分析成为企业核心竞争力的关键因素。大数据的核心是发现和利用数据的价值，而驾驭大数据的核心就是数据分析能力。面向大数据分析，数据科学家和专业的统计分析人员都需要简单、快捷的工具，将大数据与机器学习有机地结合，从而开展高效的统计分析和数据挖掘。

为了解决大数据的分析与挖掘问题，国内外陆续出现了很多计算框架与平台，其中，Apache Spark 以其卓越的性能和丰富的功能备受关注，其相应的机器学习部分更是让人激动不已。本书的作者 Alex Liu 先生密切结合实际，以清晰的思路和精心的选题，详细阐述了 Spark 机器学习的典型案例，为我们的大数据分析挖掘实践绘制了精美蓝图。

本书首先介绍了 Apache Spark 概况和机器学习基本框架 RM4E，其中包括 Spark 计算架构和一些最重要的机器学习组件，把 Spark 和机器学习有机地联系在一起，帮助开展机器学习有关项目的读者做好充分准备。接着，作者介绍了 Spark 机器学习数据准备工作，包括数据加载、数据清洗、一致性匹配、数据重组、数据连接、特征提取以及数据准备工作流和自动化等内容。完成了数据准备工作后，我们就跟随作者进入到本书的核心部分，实际案例分析。作者围绕 Spark 机器学习先后介绍了 9 个案例，内容涵盖整体视图、欺诈检测、风险评分、流失预测、产品推荐、教育分析、城市分析和开放数据建模等方面，囊括了大数据分析挖掘的主要应用场景。在每个案例中，作者对所使用的机器学习算法、数据与特征准备、模型评价方法、结果的解释都进行了详细的阐述，并给出了 Scala、R 语言、SPSS 等环境下的关键代码，使得本书具有非常强的实用性和可操作性。

无论读者是数据科学家、数据分析师、R 语言或者 SPSS 用户，通过阅读本书，一定能

够对 Spark 机器学习有更加深入的理解和掌握，能够将所学内容应用到大数据分析挖掘的具体工作中，并在学习和实践中不断加深对 Spark 大数据机器学习的理解和认识。

大数据时代最鲜明的特征就是变化，大数据技术也在日新月异的变化之中，同时，Spark 自身和机器学习领域都在快速地进行迭代演进，让我们共同努力，一起进入这绚丽多彩的大数据时代！

最后，我们要感谢本书的作者 Alex Liu 先生，感谢他奉献出引领大数据时代发展潮流和新技术应用的重要作品。感谢机械工业出版社华章公司的编辑们，是她们的远见和鼓励使得本书能与读者很快见面。感谢家人的支持和理解。尽管我们努力准确、简洁地表达作者的思想，但仍难免有词不达意之处。译文中的错误和不当之处，敬请读者朋友不吝指正，请将相关意见发往 yanlongchuan@iie.ac.cn，我们将不胜感激。

闫龙川　高德荃　李君婷

2016 年 10 月

Preface 前　言

作为数据科学家和机器学习专业人员，我们的工作是建立模型进行欺诈检测、预测客户流失，或者在广泛的领域将数据转换为洞见。为此，我们有时需要处理大量的数据和复杂的计算。因此，我们一直对新的计算工具满怀期待，例如 Spark，我们花费了很多时间来学习新工具。有很多可用的资料来学习这些新的工具，但这些资料大多都由计算机科学家编写，更多的是从计算角度来描述。

作为 Spark 用户，数据科学家和机器学习专业人员更关心新的系统如何帮助我们建立准确度更高的预测模型，如何使数据处理和编程更加简单。这是本书的写作目的，也是由数据科学家来执笔本书的主要原因。

与此同时，数据科学家和机器学习专业人员已经开发了工作框架、处理过程，使用了一些较好的建模工具，例如 R 语言和 SPSS。我们了解到一些新的工具，例如 Spark 的 MLlib，可以用它们来取代一些旧的工具，但不能全部取代。因此，作为 Spark 的用户，将 Spark 与一些已有的工具共同使用对我们十分关键，这也成为本书主要的关注点之一，是本书不同于其他 Spark 书籍的一个关键因素。

整体而言，本书是一本由数据科学家写给数据科学家和机器学习专业人员的 Spark 参考书，目的是让我们更加容易地在 Spark 上使用机器学习。

主要内容

第 1 章，从机器学习的角度介绍 Apache Spark。我们将讨论 Spark DataFrame 和 R 语言、Spark pipeline、RM4E 数据科学框架，以及 Spark notebook 和模型的实现。

第 2 章，主要介绍使用 Apache Spark 上的工具进行机器学习数据准备，例如 Spark SQL。我们将讨论数据清洗、一致性匹配、数据合并以及特征开发。

第 3 章，通过实际例子清晰地解释 RM4E 机器学习框架和处理过程，同时展示使用 Spark 轻松获得整体商业视图的优势。

第 4 章，讨论如何通过机器学习简单快速地进行欺诈检测。同时，我们会一步一步地说明从大数据中获得欺诈洞见的过程。

第 5 章，介绍一个风险评估项目的机器学习方法和处理过程，在 DataScientist-Workbench 环境下，使用 Spark 上的 R notebook 实现它们。该章我们主要关注 notebook。

第 6 章，通过开发客户流失预测系统提高客户留存度，进一步说明我们在 Spark 上使用 MLlib 进行机器学习的详细步骤。

第 7 章，描述如何使用 Spark 上的 SPSS 开发推荐系统，用 Spark 处理大数据。

第 8 章，将应用范围拓展到教育机构，如大学和培训机构，这里我们给出机器学习提升教育分析的一个真实的例子，预测学生的流失。

第 9 章，以一个基于 Spark 的服务请求预测的实际例子，帮助读者更好地理解 Spark 在商业和公共服务领域服务城市的应用。

第 10 章，进一步拓展前面章节学习的内容，让读者将所学的动态机器学习和 Spark 上的海量电信数据结合起来。

第 11 章，通过 Spark 上的开放数据介绍动态机器学习，用户可以采取数据驱动的方法，并使用所有可用的技术来优化结果。该章是第 9 章和第 10 章的扩展，同时也是前面章节所有实际例子的一个良好回顾。

预备知识

在本书中，我们假设读者有一些 Scala 或 Python 的编程基础，有一些建模工具（例如 R 语言或 SPSS）的使用经验，并且了解一些机器学习和数据科学的基础知识。

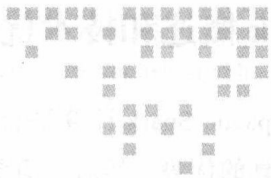
读者对象

本书主要面向需要处理大数据的分析师、数据科学家、研究人员和机器学习专业人员，但不要求相关人员熟悉 Spark。

下载彩图

我们以 PDF 文件的形式提供本书中屏幕截图和图标的彩色图片。这些彩色图片会有助于你更好地理解输出的变化。可以在以下网址下载该文件：http://www.packtpub.com/sites/default/files/downloads/ApacheSparkMachineLearningBlueprints_ColorImages.pdf。

目　录 *Contents*



第 1 章 Chapter 1

Spark 机器学习简介

本章从机器学习和数据分析视角介绍 Apache Spark，并讨论 Spark 中的机器学习计算处理技术。本章首先概括介绍 Apache Spark，通过与 MapReduce 等计算平台进行比较，展示 Spark 在数据分析中的技术优势和特点。接着，讨论如下五个方面的内容：

- 机器学习算法与程序库
- Spark RDD 和 DataFrame
- 机器学习框架
- Spark pipeline 技术
- Spark notebook 技术

以上是数据科学家或机器学习专业人员必须掌握的五项最重要的技术内容，以便于充分运用 Spark 处理计算优势。同时，本章将涵盖以下六个主题：

- Spark 概述和技术优势
- 机器学习算法和 Spark 机器学习库
- Spark RDD 和 Dataframe
- 机器学习框架、RM4E 和 Spark 计算
- 机器学习工作流和 Spark pipeline 技术
- Spark notebook 技术简介

1.1 Spark 概述和技术优势

本节对 Apache Spark 计算平台作总体介绍，通过与 MapReduce 等计算平台对比，总结 Spark 计算的优势。然后，简要介绍 Spark 计算如何适用于现代机器学习和大数据分析。

通过本节学习，读者将对 Spark 计算有一个基本了解，同时掌握一些基于 Spark 计算开展机器学习的技术优点。

1.1.1 Spark 概述

Apache Spark 是面向大数据快速处理的计算框架，该框架包含一个分布式计算引擎和一个专门设计的编程模型。2009 年，Spark 起源于美国加州大学伯克利分校 AMPLab 实验室的一个研究项目，然后在 2010 年成为 Apache 软件基金完全开源项目。之后，Apache Spark 经历了指数级增长，目前 Spark 是大数据领域最活跃的开源项目。

Spark 计算利用了内存分布式计算方法，该方法使得 Spark 计算成为最快的计算方式之一，尤其是对于反复迭代计算。根据多次测试表明，它的运行速度比 Hadoop MapReduce 快 100 倍以上。

Apache Spark 是一个统一的平台，平台由 Spark 核心引擎和四个库组成：SparkSQL、Spark Streaming、MLlib 和 GraphX。这四个库都有 Python、Java 和 Scala 的编程 API。

除了上面提到的四个内置库，Apache Spark 还有数十个由第三方提供的程序包，这些程序包可用于处理数据源、机器学习，以及其他任务。

Apache Spark 产品版本更新周期为 3 个月，Spark 1.6.0 版本更新于 2016 年 1 月 4 日。Apache Spark 1.3 版本包含有 DataFrames API 和 ML Pipelines API。自 Apache Spark 1.4 版本开始，程序已默认包含 R 界面（SparkR）。

读者可以通过链接 http://spark.apache.org/downloads.html 下载 Apache Spark。想要安装和运行 Apache Spark，可以到链接 http://spark.apache.org/docs/latest/ 下载最新说明文档。

1.1.2　Spark 优势

相对于 MapReduce 等其他大数据处理平台，Apache Spark 拥有诸多优势。其中，比较突出的两项优势是快速运行和快速写入能力。

Apache Spark 保留了诸如可扩展性和容错能力等一些 MapReduce 最重要的优势，并且利用新技术对其保留的优势进行了大幅提升。

与 MapReduce 相比，Apache Spark 的引擎可以为用户执行更为常见的**有向无环图**（DAG）。因此，使用 Apache Spark 来执行 MapReduce 风格的图计算，用户可以获得比在 Hadoop 平台上更好的批处理性能。

Apache Spark 拥有内存处理能力，并且使用了新的数据提取方法，即**弹性分布式数据集**（RDD），使得 Apache Spark 能够进行高度迭代计算和响应型编程，并且扩展了容错能力。

同时，Apache Spark 只需要几行简短的代码就可以使复杂的 pipeline 展现变得更为容易。最为人所熟知的是，它可以轻松创建算法，捕捉复杂甚至是混乱数据的真谛，并帮助用户得到实时处理结果。

Apache Spark 团队为 Spark 总结的功能包括：

- 机器学习中的迭代算法
- 交互式数据挖掘和数据处理
- 兼容 Hive 数据仓库并可提升百倍运行速度
- 流处理
- 传感器数据处理

对于在实际应用中需要处理上述问题的数据科学家，Apache Spark 在处理以下问题时可以轻而易举地显现出其优势：

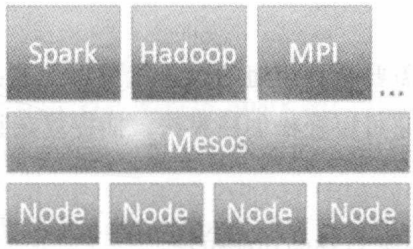
- 并行计算
- 交互式分析
- 复杂计算

大部分用户对于 Apache Spark 在速度和性能上的优势都很满意，但是部分人也注意到 Apache Spark 产品仍在不断完善。

http://svds.com/user-cases-for-apache-spark/ 提供了一些展现 Spark 优势的实例。

1.2 在机器学习中应用 Spark 计算

基于 RDD 和内存处理的创新功能，Apache Spark 真正使得分布式计算对于数据科学家和机器学习专业人员来说简便易用。Apache Spark 团队表示：Apache Spark 基于 Mesos 集群管理器运行，使其可以与 Hadoop 以及其他应用共享资源。因此，Apache Spark 可以从任何 Hadoop 输入源（如 HDFS）中读取数据。



Apache Spark 计算模型非常适合机器学习中的分布式计算。特别是在快速交互式机器学习、并行计算和大型复杂模型情境下，Apache Spark 无疑可以发挥其卓越效能。

Spark 开发团队表示，Spark 的哲学是使数据科学家和机器学习专业人员的生活更加轻松和高效。因此，Apache Spark 拥有以下特点：

- 拥有详细说明文档，表达清晰的 API
- 强大的专业领域库
- 易于与存储系统集成