Scaling Up Machine Learning: Parallel and Distributed Approaches

大规模机器学习
——并行和分布式技术

【美】罗恩·贝克曼（Ron Bekkerman）
【美】米哈伊尔·比伦科（Mikhail Bilenko） 主编
【美】约翰·兰福特（John Langford）
柳 征 王莹桂 张建廷 杨 乐 汝小虎 井沛良 译

国防工业出版社
National Defense Industry Press
CAMBRIDGE

大规模机器学习
——并行和分布式技术

Scaling Up Machine Learning:
Parallel and Distributed Approaches

[美] 罗恩·贝克曼(Ron Bekkerman)
[美] 米哈伊尔·比伦科(Mikhail Bilenko) 主编
[美] 约翰·兰福特(John Langford)

柳 征 王莹桂 张建廷
杨 乐 汝小虎 井沛良 译

国防工业出版社

·北京·

著作权合同登记　图字:军-2020-003 号

ISBN 9780521192248

图书在版编目(CIP)数据

大规模机器学习:并行和分布式技术/(美)罗恩·贝克曼(Ron Bekkerman),(美)米哈伊尔·比伦科(Mikhail Bilenko),(美)约翰·兰福特(John Langford)主编;柳征等译. —北京:国防工业出版社,2021.3
书名原文: Scaling Up Machine Learning: Parallel and Distributed Approaches
ISBN 978-7-118-12289-3

Ⅰ.①大… Ⅱ.①罗… ②米… ③约… ④柳… Ⅲ.①机器学习-并行算法 ②机器学习-分布式算法 Ⅳ.①TP181 ②TP301.6

中国版本图书馆 CIP 数据核字(2021)第 037547 号

※

国防工业出版社出版发行
(北京市海淀区紫竹院南路 23 号　邮政编码 100048)
天津嘉恒印务有限公司印刷
新华书店经售
*
开本 710×1000　1/16　印张 32　字数 574 千字
2021 年 3 月第 1 版第 1 次印刷　印数 1—2000 册　定价 108.00 元

(本书如有印装错误,我社负责调换)

国防书店:(010)88540777　　书店传真:(010)88540776
发行业务:(010)88540717　　发行传真:(010)88540762

译者序

随着科技的发展，大量的业务数据不断涌现，学术界和工业界不遗余力地从大量业务数据中挖掘有价值的信息，这其中起主导作用的就是机器学习算法。针对不同类型的数据，机器学习算法大体分为监督学习、半监督学习和无监督学习。面对海量的数据，为了高效地运用机器学习算法，研究人员开发了一系列并行算法，其充分利用 CPU 和 GPU 资源，使得常规机器学习算法计算效率有数量级的提高。

值得一提的是，近些年不断涌现的大数据框架，从初期的 Hadoop - MapReduce 并行批处理框架，过渡到基于内存计算的 Spark 框架和流式计算框架 Storm，再到目前更加灵活的 Flink 框架。其中：MapReduce 只能进行并行批处理任务；Spark 既可以进行批处理任务也可以进行流式计算任务，但其流式计算也是小型的批处理计算；Storm 虽然是流式计算框架，但其数据吞吐量有限；而 Flink 是真正意义上可以进行批处理和流式处理计算的框架。以上大数据处理框架，除了 Storm，都包含了自己的机器学习库，这也体现了并行机器学习算法在大数据处理中的重要地位。

本书内容涉及一些机器学习算法的并行化，使得大规模机器学习成为可能，内容分为大规模机器学习的框架、监督和非监督学习算法、其他的学习算法及相关应用四大部分。到目前来讲，有些技术虽不是该领域内的最新技术，但可以为相关研究人员提供一些技术参考。由于本书是以论文集的形式组织的，因此内容仅包含了一段时期内大规模并行机器学习算法的发展情况。本书不是一本相关领域的入门读物，而是需要读者具有基本的机器学习和并行/分布计算方面的知识。

参与本书翻译工作的有柳征、王莹桂、张建廷、杨乐、汝小虎和井沛良，感谢大家的共同努力。感谢国防工业出版社崔艳阳老师对本书的翻译提出的宝贵意见。限于译者水平，译文的缺点和错误在所难免，恳请读者批评指正。

译者

2020 年 7 月 27 日

前言

本书试图汇集在并行和分布式机器学习领域的最新研究。我们相信并行化为适应大数据集和复杂方法的规模化机器学习提供了关键的途径。虽然大规模机器学习在工业和学术研究界已经日益流行，但是还没有独特的资源来涵盖最近提出的各种方法。我们尽最大的努力汇集现在最具代表性的研究于一册。各章聚集于不同方法和问题，它和参考文献一起提供了该领域内的全面展示。

相信本书将会对研究人员、工程技术人员和想要掌握机器学习将来发展的读者有所帮助。为了使初学者容易理解，前 5 章提供了关于机器学习算法和并行平台的介绍性材料。尽管本书一些部分涉及较深的技术，但只假定读者有基本的机器学习和并行/分布计算方面的先验知识和大学水平的数学知识。我们希望熟悉分类器并接触过线程、MPI 或者 MapReduce 的工科本科生能够理解本书大部分内容。同时也希望经验丰富的专家会发现本书充满新的、有趣的观点，并进一步激励本领域的将来研究。

我们非常感激所有章节的作者投入大量的时间、智慧和创造力，以及为本册书所做的贡献。感谢剑桥大学出版社编辑们所做的努力：发起该项目的 Heather Bergman，在整个过程中和我们一同工作并指导本书到完成的 Lauren Cowles。感谢给每章作者们提供详细彻底反馈的审阅者们，他们的工作对完成本书是非常宝贵的。他们是 David Andrzejewski, Yoav Artzi, Arthur Asuncion, Hongjie Bai, Sugato Basu, Andrew Bender, Mark Chapman, Wen-Yen Chen, Sulabh Choudhury, Adam Coates, Kamalika Das, Kevin Duh, Igor Durdanovic, Clément Farabet, Dennis Fetterly, Eric Garcia, Joseph Gonzalez, Isaac Greenbaum, Caden Howell, Ferris Jumah, Andrey Kolobov, Jeremy Kubica, Bo Li, Luke McDowell, W. P. McNeill, Frank McSherry, Chris Meek, Xu Miao, Steena Monteiro, Miguel Osorio, Sindhu Vijaya Raghavan, Paul Rodrigues, Martin Scholz, Suhail Shergill, Sameer Singh, Tom Sommerville, Amarnag Subramanya, Narayanan Sundaram, Krysta Svore, Shirish Tatikonda, Amund Tveit, Jean Wu, Evan Xiang, Elad Yom-Tov, Bin Zhang。

Ron Bekkerman 要感谢 Martin Scholz 从本书开始阶段就参与本项目。Ron 也非常感激他的妈妈 Faina、妻子 Anna 和女儿 Naomi 在其做任何事情时给出的无尽的爱和支持。

目 录

第四部分 应 用

第1章

大规模机器学习：引言

Ron Bekkerman, Mikhail Bilenko, John Langford

近几十年来，超大规模数据集的分布式和并行处理在金融界和石油工业界等专门的、高预算的部门得到了应用。随着近些年并行计算平台的可用性、性价比和多样性的极大发展和提高，一系列数据分析和机器学习算法也相应产生。

当前，对大规模机器学习应用的兴趣不断增加，这部分归结于硬件结构和编程框架的演化和改进，也使得许多学习算法可以容易地实现并行化。一些平台可以方便地对数据本身或者其特征进行并行处理。这允许把许多无序的批量数据作为输入，并且对每批数据单独计算后进行合并归总，这种学习算法能够较直接地实现并行化。

由于在许多现代应用中超大数据集的增加，大规模机器学习得到了越来越多的关注。这些数据集常常在分布式存储平台累积，这就促使相应的分布式学习算法的发展。最后，由于基于高维复杂特征表示执行实时推断的感知设备的增多，对以学习为中心的应用提出了并行化处理的要求。这方面的例子包括语音识别和光学目标检测，并且在自主机器人和移动设备中，这些技术的使用变得很普遍。

分布式平台的多样性给机器学习算法拥有更好的性能和处理超大数据集的能力提供了许多选择。这些选择包括可定制集成电路（如域可编程门阵列（Field-Programmable Gate Array，FPGA）），常规的处理单元（如通用图形处理单元（Graphics Processing Unit，GPU）），多处理器和多核并行操作，通过局域网连接的高性能计算（High-Performance Computing，HPC）集群，可从商业云计算提供商租用的以数据为中心的虚拟集群。除了多种平台选择，也存在各种各样的能够把机器学习算法嵌入其中的编程框架。对分布式结构而言，框架的选择也趋