

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

This is a translation of the following title published by Cambridge University Press: Scaling up Machine Learning: Parallel and Distributed Approaches.

ISBN 9780521192248

This translation for the People's Republic of China (excluding Hong Kong, Macau and Taiwan) is published by arrangement with the Press Syndicate of the University of Cambridge, Cambridge, United Kingdom.

© Cambridge University Press and National Defense Industry Press 2021

This translation is authorized for sale in the People's Republic of China (excluding Hong Kong, Macau and Taiwan) only. Unauthorized export of this translation is a violation of the Copyright Act. No part of this publication may be reproduced or distributed by any means, or stored in a database or retrieval system, without the prior written permission of Cambridge University Press and National Defense Industry Press.

Copies of this book sold without a Cambridge University Press sticker on the cover are unauthorized and illegal.

本书封面贴有 Cambridge University Press 防伪标签,无标签者不得销售。

### 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 Faret, 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 Tatkonda, 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。

第 1 章 大规模机器学习:引言 .....	1
1.1 机器学习基础 .....	2
1.2 大规模机器学习的缘由 .....	4
1.2.1 大量的数据实例 .....	4
1.2.2 高输入维数 .....	5
1.2.3 模型和算法的复杂性 .....	5
1.2.4 对推断时间的约束 .....	5
1.2.5 预测串 .....	6
1.2.6 模型选择和参数搜索 .....	6
1.3 在并行分布式计算中的关键概念 .....	6
1.3.1 数据并行化 .....	6
1.3.2 任务并行化 .....	7
1.4 平台的选择和折中 .....	8
1.5 性能方面的考虑 .....	9
1.6 本书的组织结构 .....	11
1.6.1 第一部分:大规模机器学习的框架 .....	12
1.6.2 第二部分:监督和非监督学习算法 .....	13
1.6.3 第三部分:可替代的学习环境 .....	15
1.6.4 第四部分:应用 .....	16
1.7 文献注解 .....	18
参考文献 .....	19

---

## 第一部分 大规模机器学习的框架

---

第 2 章 MapReduce 及其在决策树集的大规模并行学习中的应用 .....	22
---	----

2.1	序言 .....	23
2.1.1	MapReduce .....	24
2.1.2	树模型 .....	26
2.1.3	树模型的学习 .....	27
2.1.4	回归树 .....	29
2.2	PLANET 的例子 .....	30
2.2.1	组成元素 .....	30
2.2.2	继续讨论本例子 .....	31
2.3	技术细节 .....	32
2.3.1	MR_Expand 节点:扩展单一节点 .....	32
2.3.2	MR_InMemory:内存中的树归纳 .....	36
2.3.3	控制器的设计 .....	37
2.4	集成学习 .....	38
2.5	工程方面的问题 .....	40
2.5.1	提前调度 .....	40
2.5.2	指纹法 .....	41
2.5.3	可靠性 .....	41
2.6	试验 .....	42
2.6.1	设置 .....	42
2.6.2	结果 .....	43
2.7	相关工作 .....	46
2.8	结论 .....	48
	致谢 .....	48
	参考文献 .....	48
<b>第3章</b>	<b>使用 DryadLINQ 的大规模机器学习 .....</b>	<b>52</b>
3.1	使用 LINQ 操作数据集 .....	52
3.2	用 LINQ 实现 $k$ -均值 .....	55
3.3	使用 DryadLINQ 在集群上运行 LINQ .....	56
3.3.1	Dryad .....	57
3.3.2	DryadLINQ .....	57
3.3.3	MapReduce 与 DryadLINQ .....	60
3.3.4	$k$ -均值聚类的 DryadLIN 实现 .....	61
3.3.5	DryadLINQ 实现决策树生成 .....	63
3.3.6	应用举例:奇异值分解 .....	66
3.4	应用经验总结 .....	70

3.4.1	DryadLINQ 的优势 .....	70
3.4.2	DryadLINQ 的缺点 .....	70
3.4.3	应用实例 .....	71
3.4.4	获取代码 .....	72
	参考文献 .....	72
<b>第 4 章</b>	<b>IBM 并行机器学习工具箱 .....</b>	<b>74</b>
4.1	数据并行的结合-交换计算 .....	75
4.2	API 和控制层 .....	76
4.3	分布状态算法的 API 扩展 .....	81
4.4	控制层的实现和优化 .....	82
4.5	并行核 $k$ -均值 .....	83
4.6	并行决策树 .....	86
4.7	并行频率模式挖掘 .....	88
4.8	总结 .....	91
	参考文献 .....	92
<b>第 5 章</b>	<b>机器学习算法的一致性细粒度数据并行计算 .....</b>	<b>95</b>
5.1	GP-GPU 概述 .....	97
5.2	GPU 上的一致性细粒度数据并行计算 .....	99
5.2.1	数据并行计算 .....	99
5.2.2	一致性细粒度数据并行设计 .....	101
5.3	$k$ -均值聚类算法 .....	102
5.3.1	$k$ -均值算法的一致性细粒度数据并行化 .....	103
5.4	$k$ -均值回归聚类算法 .....	105
5.5	运行和性能比较 .....	108
5.5.1	$k$ -均值的 CPU 运行 .....	108
5.5.2	GPU 加速的 $k$ -均值算法 .....	109
5.5.3	GPU 加速的 $k$ -均值 RC 算法 .....	110
5.5.4	处理实际数据时牵涉的问题 .....	111
5.6	结论 .....	111
	参考文献 .....	112

**第二部分 监督和非监督学习算法**

<b>第 6 章</b>	<b>使用不完全 Cholesky 分解的并行支持向量机 .....</b>	<b>115</b>
6.1	使用不完全 Cholesky 分解的内点法 .....	119

6.2	PSVM 算法	121
6.2.1	并行 ICF	121
6.2.2	并行 IPM	127
6.2.3	$b$ 的计算与回写	128
6.3	实验	129
6.3.1	分类-预测精度	129
6.3.2	可扩展性	130
6.3.3	开销	131
6.4	结论	132
	致谢	133
	参考文献	133
<b>第 7 章</b>	<b>使用硬件加速器的大规模并行支持向量机</b>	<b>135</b>
7.1	问题描述	136
7.1.1	求解二次优化问题	138
7.1.2	SMO 算法推导	138
7.1.3	工作集的选取	139
7.2	SMO 算法实现	139
7.3	微并行(Micro Parallelization) 相关文献综述	142
7.4	现有多核系统并行方案	142
7.5	微并行回顾	145
7.6	大规模并行硬件加速器	147
7.6.1	实验数据集	148
7.6.2	数值精度的影响	150
7.6.3	HOST 加速器时序	151
7.6.4	其他并行方案	154
7.6.5	加速其他算法	155
7.7	实验结果	155
7.8	结论	156
	参考文献	157
<b>第 8 章</b>	<b>基于提升决策树的大规模排序学习</b>	<b>159</b>
8.1	相关工作	160
8.2	LambdaMART	162
8.3	LambdaMART 的分布式实现方法	164
8.3.1	基于特征分布的同步方法	165

8.3.2	基于数据分布的同步方法 .....	167
8.3.3	加入随机化 .....	169
8.4	实验 .....	169
8.4.1	数据 .....	170
8.4.2	评估措施 .....	170
8.4.3	时间复杂度比较 .....	171
8.4.4	准确度比较 .....	174
8.4.5	关于数据分布式 LambdaMART 的附加讨论 .....	179
8.5	结论及未来工作 .....	180
8.6	致谢 .....	181
	参考文献 .....	181
<b>第 9 章</b>	<b>变换回归算法 .....</b>	<b>183</b>
9.1	分类、回归和损失函数 .....	184
9.2	背景 .....	185
9.3	动机和算法描述 .....	187
9.4	TReg 扩展:初始化和终止 .....	191
9.4.1	子基函数计算细节 .....	192
9.4.2	阶段基函数计算细节 .....	195
9.4.3	线搜索优化细节 .....	196
9.5	功能评估 .....	197
9.6	并行性能结果 .....	200
9.6.1	可扩展性分析 .....	200
9.6.2	PML 性能优化 .....	200
9.6.3	并行可扩展性结果 .....	201
9.7	总结 .....	202
	参考文献 .....	203
<b>第 10 章</b>	<b>因子图中的并行信度传递算法 .....</b>	<b>204</b>
10.1	因子图中的信度传递 .....	205
10.1.1	信度传递 .....	206
10.1.2	信度传递并行实现的前提条件 .....	208
10.2	共享内存并行信度传递 .....	209
10.2.1	同步(映射化简)信度传递 .....	209
10.2.2	轮询调度信度传递 .....	213
10.2.3	野火信度传递 .....	215

10.2.4	残差信度传递 .....	216
10.2.5	Splash 信度传递 .....	217
10.3	多核性能对比 .....	225
10.4	聚类上的并行信度传递 .....	226
10.4.1	因子图和消息的分割 .....	227
10.4.2	分布式设置下算法对比 .....	228
10.5	结论 .....	229
	致谢 .....	230
	参考文献 .....	230
<b>第 11 章</b>	<b>隐含变量模型的分布式吉布斯采样 .....</b>	<b>233</b>
11.1	隐含变量模型 .....	233
11.1.1	隐含狄利克雷分布 .....	234
11.1.2	分层狄利克雷过程 .....	236
11.2	分布式推理算法 .....	237
11.2.1	近似分布 LDA 和 HDP .....	237
11.2.2	异步分布式学习技术 .....	239
11.3	分布式主题模型实验分析 .....	241
11.3.1	分布式算法的准确性 .....	241
11.3.2	现实世界数据集的可扩展性 .....	245
11.4	实际实现指导 .....	246
11.4.1	分布式并行硬件 .....	246
11.4.2	补充加速技术 .....	247
11.5	延伸到贝叶斯网络的分布式推理 .....	249
11.5.1	贝叶斯网络 .....	249
11.5.2	案例:隐马尔可夫模型 .....	252
11.5.3	贝叶斯网络的分布式推理 .....	253
11.6	结论 .....	255
	参考文献 .....	255
<b>第 12 章</b>	<b>基于 MapReduce 和 MPI 的大规模谱聚类 .....</b>	<b>259</b>
12.1	谱聚类算法 .....	261
12.2	基于稀疏相似度矩阵的谱聚类 .....	262
12.3	基于稀疏相似度矩阵的并行谱聚类 .....	265
12.3.1	MPI 和 MapReduce .....	265
12.3.2	相似度矩阵和最近邻 .....	266

12.3.3	并行特征值分解 .....	268
12.3.4	并行 $k$ -均值算法 .....	270
12.4	实验 .....	271
12.4.1	基于稀疏相似度矩阵的聚类质量 .....	272
12.4.2	分布式环境下速度提升及伸缩性 .....	274
12.5	结论 .....	280
	参考文献 .....	281
<b>第 13 章</b>	<b>并行化信息论聚类方法 .....</b>	<b>284</b>
13.1	信息论聚类 .....	286
13.2	并行聚类 .....	288
13.2.1	并行 IT-CC .....	289
13.3	序贯联合聚类 .....	291
13.4	DataLoom 算法 .....	292
13.5	执行与实验 .....	297
13.5.1	与序贯联合聚类的比较 .....	298
13.5.2	RCV1 数据集 .....	299
13.5.3	Netflix 数据集 .....	300
13.6	结论 .....	301
	参考文献 .....	301

**第三部分 其他的学习算法**

<b>第 14 章</b>	<b>并行在线学习 .....</b>	<b>305</b>
14.1	带宽和延迟带来的限制 .....	307
14.2	并行策略 .....	308
14.3	延迟更新分析 .....	310
14.3.1	约定 .....	311
14.4	并行学习算法 .....	313
14.4.1	多核特征分片 .....	313
14.4.2	多节点特征分片 .....	314
14.4.3	实验 .....	318
14.5	全局更新规则 .....	320
14.5.1	延迟全局更新 .....	321
14.5.2	纠正更新 .....	321

14.5.3	延迟的反向传播 .....	321
14.5.4	Minibatch 梯度下降 .....	322
14.5.5	Minibatch 共轭梯度 .....	323
14.5.6	确定更新 .....	324
14.6	实验 .....	325
14.7	结论 .....	327
	参考文献 .....	327
<b>第 15 章</b>	<b>基于图的半监督学习并行化 .....</b>	<b>329</b>
15.1	将 SSL 扩展到大规模数据集 .....	331
15.2	基于图的 SSL .....	332
15.2.1	图构造 .....	334
15.2.2	图正则化 .....	335
15.3	数据集:1.2 亿节点的图 .....	340
15.3.1	大规模问题的图构造 .....	341
15.4	大规模并行处理 .....	342
15.4.1	共享存储器对称多处理器上的推断 .....	342
15.4.2	SMP 的图重排算法 .....	343
15.4.3	分布式计算机环境中的推断 .....	347
15.5	讨论 .....	351
	参考文献 .....	351
<b>第 16 章</b>	<b>基于协同矩阵分解的分布式迁移学习 .....</b>	<b>356</b>
16.1	分布式联合学习 .....	358
16.1.1	协同矩阵分解 .....	358
16.1.2	CoMF 的分布式学习 .....	360
16.1.3	CoMF 实现知识迁移 .....	366
16.1.4	总结 .....	367
16.2	DisCo 扩展到分类任务 .....	368
16.2.1	监督协同矩阵分解 .....	368
16.2.2	监督 CoMF 的分布式学习 .....	370
16.2.3	监督 CoMF 实现知识迁移 .....	374
16.2.4	总结 .....	375
16.3	结论 .....	375
	参考文献 .....	375
<b>第 17 章</b>	<b>并行大规模特征选择 .....</b>	<b>377</b>

17.1	逻辑回归 .....	378
17.2	特征选择 .....	379
17.2.1	前向特征选择 .....	380
17.2.2	单特征优化 .....	380
17.2.3	移植 .....	381
17.2.4	多类预测问题 .....	382
17.3	并行化特征选择算法 .....	382
17.3.1	并行完全前向特征选择 .....	383
17.3.2	并行 SFO .....	383
17.3.3	并行移植 .....	386
17.3.4	其他相关算法 .....	387
17.4	实验结果 .....	388
17.4.1	UCI 互联网广告 (Internet Ads) 数据集 .....	389
17.4.2	RCV1 数据集 .....	390
17.4.3	时间测定结果 .....	392
17.5	结论 .....	394
	参考文献 .....	394

## 第四部分 应用

<b>第 18 章</b>	<b>基于 GPU 的计算机视觉大规模学习 .....</b>	<b>398</b>
18.1	标准管线 .....	399
18.2	GPU 介绍 .....	401
18.2.1	数据并行编程 .....	402
18.2.2	CUDA 编程模型 .....	402
18.2.3	GPU 上的卷积例子 .....	403
18.2.4	GPU 的结论 .....	405
18.3	标准方法扩展 .....	405
18.3.1	合成训练数据 .....	405
18.3.2	支持 GPU 的特征 .....	406
18.3.3	滑动窗口物体检测和卷积 .....	407
18.3.4	分类 .....	408
18.3.5	实验结果 .....	410
18.4	使用深度信念网络进行特征学习 .....	413

18.4.1	深度信念网络 .....	414
18.4.2	GPU 上的 DBN 学习 .....	415
18.4.3	实验结果 .....	417
18.5	结论 .....	419
	参考文献 .....	419
<b>第 19 章</b>	<b>基于 FPGA 的大规模卷积网络</b> .....	<b>423</b>
19.1	学习内部表示 .....	424
19.1.1	卷积网络 .....	424
19.1.2	发展和应用 .....	426
19.1.3	卷积网络的非监督学习 .....	427
19.2	专用数字硬件架构 .....	429
19.2.1	数据流方法 .....	431
19.2.2	基于 FPGA 的卷积网络处理器 .....	434
19.2.3	卷积网络处理器的卷积网络编译 .....	437
19.2.4	性能表现 .....	438
19.3	总结 .....	440
	参考文献 .....	440
<b>第 20 章</b>	<b>多核系统上的树状结构数据挖掘</b> .....	<b>444</b>
20.1	多核的挑战 .....	445
20.2	背景 .....	446
20.2.1	问题定义 .....	447
20.2.2	研究现状 .....	448
20.2.3	Trips 算法 .....	449
20.3	内存优化 .....	452
20.3.1	即时嵌入列表生成 (NOEM) .....	452
20.3.2	树匹配优化 .....	454
20.3.3	计算块 (CHUNK) .....	455
20.4	自适应并行化 .....	456
20.4.1	任务并行模式 .....	458
20.4.2	数据并行模式 .....	459
20.4.3	块并行模式 .....	460
20.4.4	成本分析 .....	460
20.4.5	调度服务 .....	461
20.5	实例评估 .....	462

---

20.5.1	性能结果 .....	463
20.5.2	CMP 架构的特征研究 .....	465
20.5.3	并行性能 .....	466
20.6	结论 .....	468
	致谢 .....	469
	参考文献 .....	469
<b>第 21 章</b>	<b>自动语音识别的可伸缩并行化 .....</b>	<b>472</b>
21.1	并发识别 .....	476
21.2	软件架构和实施挑战 .....	478
21.3	多核和众核并行平台 .....	480
21.4	多核基础结构和映射 .....	481
21.4.1	数据注意事项 .....	481
21.4.2	任务注意事项 .....	482
21.4.3	运行时注意事项 .....	483
21.4.4	总结 .....	484
21.5	众核实现 .....	486
21.5.1	任务注意事项 .....	487
21.5.2	运行时注意事项 .....	488
21.5.3	总结 .....	488
21.6	实现分析和灵敏度分析 .....	488
21.6.1	语言模型和测试集 .....	488
21.6.2	总体性能 .....	489
21.6.3	灵敏度分析 .....	490
21.7	应用级别优化 .....	491
21.7.1	言语模式的选择 .....	491
21.7.2	替代模型评估 .....	492
21.8	结论和主要经验教训 .....	493
21.8.1	并行化过程 .....	493
21.8.2	使用框架实现高效的并行应用开发 .....	493
	参考文献 .....	494

# 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