

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

Scaling Up Machine Learning

Parallel and Distributed Approaches

This book comprises a collection of representative approaches for scaling up machine learning and data mining methods on parallel and distributed computing platforms. Demand for parallelizing learning algorithms is highly task-specific: in some settings it is driven by the enormous dataset sizes, in others by model complexity or by real-time performance requirements. Making task-appropriate algorithm and platform choices for large-scale machine learning requires understanding the benefits, trade-offs, and constraints of the available options.

Solutions presented in the book cover a range of parallelization platforms from FPGAs and GPUs to multi-core systems and commodity clusters; concurrent programming frameworks that include CUDA, MPI, MapReduce, and DryadLINQ; and various learning settings: supervised, unsupervised, semi-supervised, and online learning. Extensive coverage of parallelization of boosted trees, support vector machines, spectral clustering, belief propagation, and other popular learning algorithms accompanied by deep dives into several applications make the book equally useful for researchers, students, and practitioners.

Dr. Ron Bekkerman is a computer engineer and scientist whose experience spans across disciplines from video processing to business intelligence. Currently a senior research scientist at LinkedIn, he previously worked for a number of major companies including Hewlett-Packard and Motorola. Ron's research interests lie primarily in the area of large-scale unsupervised learning. He is the corresponding author of several publications in top-tier venues, such as ICML, KDD, SIGIR, WWW, IJCAI, CVPR, EMNLP, and JMLR.

Dr. Mikhail Bilenko is a researcher in the Machine Learning Group at Microsoft Research. His research interests center on machine learning and data mining tasks that arise in the context of large behavioral and textual datasets. Mikhail's recent work has focused on learning algorithms that leverage user behavior to improve online advertising. His papers have been published in KDD, ICML, SIGIR, and WWW among other venues, and I have received best paper awards from SIGIR and KDD.

Dr. John Langford is a computer scientist working as a senior researcher at Yahoo! Research. Previously, he was affiliated with the Toyota Technological Institute and IBM T. J. Watson Research Center. John's work has been published in conferences and journals including ICML, COLT, NIPS, UAI, KDD, JMLR, and MLJ. He received the Pat Goldberg Memorial Best Paper Award, as well as best paper awards from ACM EC and WSDM. He is also the author of the popular machine learning weblog, hunch.net.

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

Scaling Up Machine Learning

Parallel and Distributed Approaches

Edited by

Ron Bekkerman

Mikhail Bilenko

John Langford



CAMBRIDGE
UNIVERSITY PRESS

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

CAMBRIDGE UNIVERSITY PRESS

Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press

32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org

Information on this title: www.cambridge.org/9780521192248

© Cambridge University Press 2012

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2012

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication data

Scaling up machine learning : parallel and distributed approaches / [edited by] Ron Bekkerman,
Mikhail Bilenko, John Langford.

p. cm.

Includes index.

ISBN 978-0-521-19224-8 (hardback)

1. Machine learning. 2. Data mining. 3. Parallel algorithms. 4. Parallel programs (Computer
programs) I. Bekkerman, Ron, 1974– II. Bilenko, Mikhail, 1978– III. Langford, John, 1975–
Q325.5.S28 2011

006.3'1–dc23 2011016323

ISBN 978-0-521-19224-8 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for
external or third-party Internet Web sites referred to in this publication and does not guarantee that
any content on such Web sites is, or will remain, accurate or appropriate.

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

Contents

<i>Contributors</i>	xi
<i>Preface</i>	xv
1 Scaling Up Machine Learning: Introduction	1
<i>Ron Bekkerman, Mikhail Bilenko, and John Langford</i>	
1.1 Machine Learning Basics	2
1.2 Reasons for Scaling Up Machine Learning	3
1.3 Key Concepts in Parallel and Distributed Computing	6
1.4 Platform Choices and Trade-Offs	7
1.5 Thinking about Performance	9
1.6 Organization of the Book	10
1.7 Bibliographic Notes	17
References	19
Part One Frameworks for Scaling Up Machine Learning	
2 MapReduce and Its Application to Massively Parallel Learning of Decision Tree Ensembles	23
<i>Biswanath Panda, Joshua S. Herbach, Sugato Basu, and Roberto J. Bayardo</i>	
2.1 Preliminaries	24
2.2 Example of PLANET	30
2.3 Technical Details	33
2.4 Learning Ensembles	38
2.5 Engineering Issues	39
2.6 Experiments	41
2.7 Related Work	44
2.8 Conclusions	46
Acknowledgments	47
References	47

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

vi

CONTENTS

3 Large-Scale Machine Learning Using DryadLINQ	49
<i>Mihai Budiu, Dennis Fetterly, Michael Isard, Frank McSherry, and Yuan Yu</i>	
3.1 Manipulating Datasets with LINQ	49
3.2 k -Means in LINQ	52
3.3 Running LINQ on a Cluster with DryadLINQ	53
3.4 Lessons Learned	65
References	67
4 IBM Parallel Machine Learning Toolbox	69
<i>Edwin Pednault, Elad Yom-Tov, and Amol Ghoting</i>	
4.1 Data-Parallel Associative-Commutative Computation	70
4.2 API and Control Layer	71
4.3 API Extensions for Distributed-State Algorithms	76
4.4 Control Layer Implementation and Optimizations	77
4.5 Parallel Kernel k -Means	79
4.6 Parallel Decision Tree	80
4.7 Parallel Frequent Pattern Mining	83
4.8 Summary	86
References	87
5 Uniformly Fine-Grained Data-Parallel Computing for Machine Learning Algorithms	89
<i>Meichun Hsu, Ren Wu, and Bin Zhang</i>	
5.1 Overview of a GP-GPU	91
5.2 Uniformly Fine-Grained Data-Parallel Computing on a GPU	93
5.3 The k -Means Clustering Algorithm	97
5.4 The k -Means Regression Clustering Algorithm	99
5.5 Implementations and Performance Comparisons	102
5.6 Conclusions	105
References	105
Part Two Supervised and Unsupervised Learning Algorithms	
6 PSVM: Parallel Support Vector Machines with Incomplete Cholesky Factorization	109
<i>Edward Y. Chang, Hongjie Bai, Kaihua Zhu, Hao Wang, Jian Li, and Zhihuan Qiu</i>	
6.1 Interior Point Method with Incomplete Cholesky Factorization	112
6.2 PSVM Algorithm	114
6.3 Experiments	121
6.4 Conclusion	125
Acknowledgments	125
References	125
7 Massive SVM Parallelization Using Hardware Accelerators	127
<i>Igor Durdanovic, Eric Cosatto, Hans Peter Graf, Srihari Cadambi, Venkata Jakkula, Srimat Chakradhar, and Abhinandan Majumdar</i>	
7.1 Problem Formulation	128
7.2 Implementation of the SMO Algorithm	131

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

CONTENTS	vii
7.3 Micro Parallelization: Related Work	132
7.4 Previous Parallelizations on Multicore Systems	133
7.5 Micro Parallelization: Revisited	136
7.6 Massively Parallel Hardware Accelerator	137
7.7 Results	145
7.8 Conclusion	146
References	146
8 Large-Scale Learning to Rank Using Boosted Decision Trees	148
<i>Krysta M. Svore and Christopher J. C. Burges</i>	
8.1 Related Work	149
8.2 LambdaMART	151
8.3 Approaches to Distributing LambdaMART	153
8.4 Experiments	158
8.5 Conclusions and Future Work	168
8.6 Acknowledgments	169
References	169
9 The Transform Regression Algorithm	170
<i>Ramesh Natarajan and Edwin Pednault</i>	
9.1 Classification, Regression, and Loss Functions	171
9.2 Background	172
9.3 Motivation and Algorithm Description	173
9.4 TReg Expansion: Initialization and Termination	177
9.5 Model Accuracy Results	184
9.6 Parallel Performance Results	186
9.7 Summary	188
References	189
10 Parallel Belief Propagation in Factor Graphs	190
<i>Joseph Gonzalez, Yucheng Low, and Carlos Guestrin</i>	
10.1 Belief Propagation in Factor Graphs	191
10.2 Shared Memory Parallel Belief Propagation	195
10.3 Multicore Performance Comparison	209
10.4 Parallel Belief Propagation on Clusters	210
10.5 Conclusion	214
Acknowledgments	214
References	214
11 Distributed Gibbs Sampling for Latent Variable Models	217
<i>Arthur Asuncion, Padhraic Smyth, Max Welling, David Newman, Ian Porteous, and Scott Triglia</i>	
11.1 Latent Variable Models	217
11.2 Distributed Inference Algorithms	220
11.3 Experimental Analysis of Distributed Topic Modeling	224
11.4 Practical Guidelines for Implementation	229
11.5 A Foray into Distributed Inference for Bayesian Networks	231
11.6 Conclusion	236
Acknowledgments	237
References	237

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

viii

CONTENTS

12 Large-Scale Spectral Clustering with MapReduce and MPI	240
<i>Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, and Edward Y. Chang</i>	
12.1 Spectral Clustering	241
12.2 Spectral Clustering Using a Sparse Similarity Matrix	243
12.3 Parallel Spectral Clustering (PSC) Using a Sparse Similarity Matrix	245
12.4 Experiments	251
12.5 Conclusions	258
References	259
13 Parallelizing Information-Theoretic Clustering Methods	262
<i>Ron Bekkerman and Martin Scholz</i>	
13.1 Information-Theoretic Clustering	264
13.2 Parallel Clustering	266
13.3 Sequential Co-clustering	269
13.4 The DataLoom Algorithm	270
13.5 Implementation and Experimentation	274
13.6 Conclusion	277
References	278
Part Three Alternative Learning Settings	
14 Parallel Online Learning	283
<i>Daniel Hsu, Nikos Karampatziakis, John Langford, and Alex J. Smola</i>	
14.1 Limits Due to Bandwidth and Latency	285
14.2 Parallelization Strategies	286
14.3 Delayed Update Analysis	288
14.4 Parallel Learning Algorithms	290
14.5 Global Update Rules	298
14.6 Experiments	302
14.7 Conclusion	303
References	305
15 Parallel Graph-Based Semi-Supervised Learning	307
<i>Jeff Bilmes and Amarnag Subramanya</i>	
15.1 Scaling SSL to Large Datasets	309
15.2 Graph-Based SSL	310
15.3 Dataset: A 120-Million-Node Graph	317
15.4 Large-Scale Parallel Processing	319
15.5 Discussion	327
References	328
16 Distributed Transfer Learning via Cooperative Matrix Factorization	331
<i>Evan Xiang, Nathan Liu, and Qiang Yang</i>	
16.1 Distributed Coalitional Learning	333
16.2 Extension of DisCo to Classification Tasks	343

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

CONTENTS	ix
16.3 Conclusion	350
References	350
17 Parallel Large-Scale Feature Selection	352
<i>Jeremy Kubica, Sameer Singh, and Daria Sorokina</i>	
17.1 Logistic Regression	353
17.2 Feature Selection	354
17.3 Parallelizing Feature Selection Algorithms	358
17.4 Experimental Results	363
17.5 Conclusions	368
References	368
Part Four Applications	
18 Large-Scale Learning for Vision with GPUs	373
<i>Adam Coates, Rajat Raina, and Andrew Y. Ng</i>	
18.1 A Standard Pipeline	374
18.2 Introduction to GPUs	377
18.3 A Standard Approach Scaled Up	380
18.4 Feature Learning with Deep Belief Networks	388
18.5 Conclusion	395
References	395
19 Large-Scale FPGA-Based Convolutional Networks	399
<i>Clément Farabet, Yann LeCun, Koray Kavukcuoglu, Berin Martini, Polina Akselrod, Selcuk Talay, and Eugenio Culurciello</i>	
19.1 Learning Internal Representations	400
19.2 A Dedicated Digital Hardware Architecture	405
19.3 Summary	416
References	417
20 Mining Tree-Structured Data on Multicore Systems	420
<i>Shirish Tatikonda and Srinivasan Parthasarathy</i>	
20.1 The Multicore Challenge	422
20.2 Background	423
20.3 Memory Optimizations	427
20.4 Adaptive Parallelization	431
20.5 Empirical Evaluation	437
20.6 Discussion	442
Acknowledgments	443
References	443
21 Scalable Parallelization of Automatic Speech Recognition	446
<i>Jike Chong, Ekaterina Gonina, Kisun You, and Kurt Keutzer</i>	
21.1 Concurrency Identification	450
21.2 Software Architecture and Implementation Challenges	452
21.3 Multicore and Manycore Parallel Platforms	454
21.4 Multicore Infrastructure and Mapping	455

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

x

CONTENTS

21.5 The Manycore Implementation	459
21.6 Implementation Profiling and Sensitivity Analysis	462
21.7 Application-Level Optimization	464
21.8 Conclusion and Key Lessons	467
References	468
<i>Subject Index</i>	471

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

Contributors

Polina Akselrod

Yale University, New Haven, CT, USA

Arthur Asuncion

University of California, Irvine, CA,
USA

Hongjie Bai

Google Research, Beijing, China

Sugato Basu

Google Research, Mountain View, CA,
USA

Roberto J. Bayardo

Google Research, Mountain View, CA,
USA

Ron Bekkerman

LinkedIn Corporation, Mountain View,
CA, USA

Mikhail Bilenko

Microsoft Research, Redmond, WA,
USA

Jeff Bilmes

University of Washington, Seattle, WA,
USA

Mihai Badiu

Microsoft Research, Mountain View,
CA, USA

Christopher J. C. Burges

Microsoft Research, Redmond, WA,
USA

Srihari Cadambi

NEC Labs America, Princeton, NJ, USA

Srimat Chakradhar

NEC Labs America, Princeton, NJ, USA

Edward Y. Chang

Google Research, Beijing, China

Wen-Yen Chen

University of California, Santa Barbara,
CA, USA

Jike Chong

Parasians LLC, Sunnyvale, CA, USA

Adam Coates

Stanford University, Stanford, CA, USA

Eric Cosatto

NEC Labs America, Princeton, NJ, USA

Eugenio Culurciello

Yale University, New Haven, CT, USA

Igor Durdanovic

NEC Labs America, Princeton, NJ, USA

Clément Farabet

New York University, New York, NY,
USA

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

xii

CONTRIBUTORS

Dennis FetterlyMicrosoft Research, Mountain View,
CA, USA**Amol Ghoting**IBM Research, Yorktown Heights, NY,
USA**Ekaterina Gonina**University of California, Berkeley, CA,
USA**Joseph Gonzalez**Carnegie Mellon University, Pittsburgh,
PA, USA**Hans Peter Graf**

NEC Labs America, Princeton, NJ, USA

Carlos GuestrinCarnegie Mellon University, Pittsburgh,
PA, USA**Joshua S. Herbach**

Google Inc., Mountain View, CA, USA

Daniel HsuRutgers University, Piscataway, NJ, USA
and University of Pennsylvania,
Philadelphia, PA, USA**Meichun Hsu**

HP Labs, Palo Alto, CA, USA

Michael IsardMicrosoft Research, Mountain View,
CA, USA**Venkata Jakkula**

NEC Labs America, Princeton, NJ, USA

Nikos Karampatziakis

Cornell University, Ithaca, NY, USA

Koray Kavukcuoglu

NEC Labs America, Princeton, NJ, USA

Kurt KeutzerUniversity of California, Berkeley, CA,
USA**Jeremy Kubica**

Google Inc., Pittsburgh, PA, USA

John Langford

Yahoo! Research, New York, NY, USA

Yann LeCunNew York University, New York, NY,
USA**Jian Li**

Google Research, Beijing, China

Chih-Jen LinNational Taiwan University, Taipei,
Taiwan**Nathan Liu**Hong Kong University of Science and
Technology, Kowloon, Hong Kong**Yucheng Low**Carnegie Mellon University, Pittsburgh,
PA, USA**Abhinandan Majumdar**

NEC Labs America, Princeton, NJ, USA

Berin Martini

Yale University, New Haven, CT, USA

Frank McSherryMicrosoft Research, Mountain View,
CA, USA**Ramesh Natarajan**IBM Research, Yorktown Heights, NY,
USA**David Newman**University of California, Irvine, CA,
USA**Andrew Y. Ng**

Stanford University, Stanford, CA, USA

Biswanath Panda

Google Inc., Mountain View, CA, USA

Srinivasan ParthasarathyOhio State University, Columbus, OH,
USA**Edwin Pednault**IBM Research, Yorktown Heights, NY,
USA

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

CONTRIBUTORS

xiii

Ian Porteous

Google Inc., Kirkland, WA, USA

Zhihuan Qiu

Google Research, Beijing, China

Rajat Raina

Facebook Inc., Palo Alto, CA, USA

Martin Scholz

HP Labs, Palo Alto, CA, USA

Sameer Singh

University of Massachusetts, Amherst,
MA, USA

Alex J. Smola

Yahoo! Research, Santa Clara, NY, USA

Padhraic Smyth

University of California, Irvine, CA,
USA

Yangqiu Song

Tsinghua University, Beijing, China

Daria Sorokina

Yandex Labs, Palo Alto, CA, USA

Amarnag Subramanya

Google Research, Mountain View, CA,
USA

Krysta M. Svore

Microsoft Research, Redmond, WA,
USA

Selcuk Talay

Yale University, New Haven, CT, USA

Shirish Tatikonda

IBM Research, San Jose, CA, USA

Scott Triglia

University of California, Irvine, CA,
USA

Hao Wang

Google Research, Beijing, China

Max Welling

University of California, Irvine, CA,
USA

Ren Wu

HP Labs, Palo Alto, CA, USA

Evan Xiang

Hong Kong University of Science and
Technology, Kowloon, Hong Kong

Qiang Yang

Hong Kong University of Science and
Technology, Kowloon, Hong Kong

Elad Yom-Tov

Yahoo! Research, New York, NY, USA

Kisun You

Seoul National University, Seoul, Korea

Yuan Yu

Microsoft Research, Mountain View,
CA, USA

Bin Zhang

HP Labs, Palo Alto, CA, USA

Kaihua Zhu

Google Research, Beijing, China

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

Preface

This book attempts to aggregate state-of-the-art research in parallel and distributed machine learning. We believe that parallelization provides a key pathway for scaling up machine learning to large datasets and complex methods. Although large-scale machine learning has been increasingly popular in both industrial and academic research communities, there has been no singular resource covering the variety of approaches recently proposed. We did our best to assemble the most representative contemporary studies in one volume. While each contributed chapter concentrates on a distinct approach and problem, together with their references they provide a comprehensive view of the field.

We believe that the book will be useful to the broad audience of researchers, practitioners, and anyone who wants to grasp the future of machine learning. To smooth the ramp-up for beginners, the first five chapters provide introductory material on machine learning algorithms and parallel computing platforms. Although the book gets deeply technical in some parts, the reader is assumed to have only basic prior knowledge of machine learning and parallel/distributed computing, along with college-level mathematical maturity. We hope that an engineering undergraduate who is familiar with the notion of a classifier and had some exposure to threads, MPI, or MapReduce will be able to understand the majority of the book's content. We also hope that a seasoned expert will find this book full of new, interesting ideas to inspire future research in the area.

We are deeply thankful to all chapter authors for significant investments of their time, talent, and creativity in preparing their contributions to this volume. We appreciate the efforts of our editors at Cambridge University Press: Heather Bergman, who initiated this project, and Lauren Cowles, who worked with us throughout the process, guiding the book to completion. We thank chapter reviewers who provided detailed, thoughtful feedback to chapter authors that was invaluable in shaping the book: 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

Cambridge University Press

978-0-521-19224-8 - Scaling Up Machine Learning: Parallel and Distributed Approaches

Edited by Ron Bekkerman, Mikhail Bilenko and John Langford

Frontmatter

[More information](#)

xvi

PREFACE

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, and Bin Zhang.

Ron Bekkerman would like to thank Martin Scholz for his personal involvement in this project since its initial stage. Ron is deeply grateful to his mother Faina, wife Anna, and daughter Naomi, for their endless love and support throughout all his ventures.