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.

Scaling Up Machine Learning

Parallel and Distributed Approaches

Edited by
Ron Bekkerman
Mikhail Bilenko
John Langford

CAMBRIDGE
UNIVERSITY PRESS

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

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

Contents

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 Budiu
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

Dennis Fetterly
Microsoft 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 Guestrin
Carnegie Mellon University, Pittsburgh, PA, USA

Joshua S. Herbach
Google Inc., Mountain View, CA, USA

Daniel Hsu
Rutgers University, Piscataway, NJ, USA and University of Pennsylvania, Philadelphia, PA, USA

Meichun Hsu
HP Labs, Palo Alto, CA, USA

Michael Isard
Microsoft 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 Keutzer
University of California, Berkeley, CA, USA

Jeremy Kubica
Google Inc., Pittsburgh, PA, USA

John Langford
Yahoo! Research, New York, NY, USA

Yann LeCun
New York University, New York, NY, USA

Jian Li
Google Research, Beijing, China

Chih-Jen Lin
National 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 McSherry
Microsoft 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 Parthasarathy
Ohio State University, Columbus, OH, USA

Edwin Pednault
IBM Research, Yorktown Heights, NY, USA

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

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

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.