Integrated Series in Information Systems 36
Series Editors: Ramesh Sharda · Stefan Voß

Shan Suthaharan

Machine Learning Models and Algorithms for Big Data Classification

Thinking with Examples for Effective Learning

Springer

Integrated Series in Information Systems

Volume 36

Series Editors

Ramesh Sharda
Oklahoma State University, Stillwater, OK, USA

Stefan Voß
University of Hamburg, Hamburg, Germany

More information about this series at http://www.springer.com/series/6157

Shan Suthaharan

Machine Learning Models and Algorithms for Big Data Classification

Thinking with Examples for Effective Learning

Springer

Shan Suthaharan
Department of Computer Science
UNC Greensboro
Greensboro, NC, USA

ISSN 1571-0270 ISSN 2197-7968 (electronic)
Integrated Series in Information Systems
ISBN 978-1-4899-7640-6 ISBN 978-1-4899-7641-3 (eBook)
DOI 10.1007/978-1-4899-7641-3

Library of Congress Control Number: 2015950063

Springer New York Heidelberg Dordrecht London
© Springer Science+Business Media New York 2016
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

It is the quality of our work which will please God and not the quantity – Mahatma Gandhi

If you can't explain it simply, you don't understand it well enough – Albert Einstein

Preface

The interest in writing this book began at the IEEE International Conference on Intelligence and Security Informatics held in Washington, DC (June 11–14, 2012), where Mr. Matthew Amboy, the editor of *Business and Economics: OR and MS*, published by Springer Science+Business Media, expressed the need for a book on this topic, mainly focusing on a topic in data science field. The interest went even deeper when I attended the workshop conducted by Professor Bin Yu (Department of Statistics, University of California, Berkeley) and Professor David Madigan (Department of Statistics, Columbia University) at the Institute for Mathematics and its Applications, University of Minnesota on June 16–29, 2013.

Data science is one of the emerging fields in the twenty-first century. This field has been created to address the big data problems encountered in the day-to-day operations of many industries, including financial sectors, academic institutions, information technology divisions, health care companies, and government organizations. One of the important big data problems that needs immediate attention is in big data classifications. The network intrusion detection, public space intruder detection, fraud detection, spam filtering, and forensic linguistics are some of the practical examples of big data classification problems that require immediate attention.

We need significant collaboration between the experts in many disciplines, including mathematics, statistics, computer science, engineering, biology, and chemistry to find solutions to this challenging problem. Educational resources, like books and software, are also needed to train students to be the next generation of research leaders in this emerging research field. One of the current fields that brings the interdisciplinary experts, educational resources, and modern technologies under one roof is machine learning, which is a subfield of artificial intelligence.

Many models and algorithms for standard classification problems are available in the machine learning literature. However, a few of them are suitable for big data classification. Big data classification is dependent not only on the mathematical and software techniques but also on the computer technologies that help store, retrieve, and process the data with efficient scalability, accessibility, and computability features. One such recent technology is the distributed file system. A particular system

that has become popular and provides these features is the Hadoop distributed file system, which uses the modern techniques called MapReduce programming model (or a framework) with Mapper and Reducer functions that adopt the concept called the (key, value) pairs. The machine learning techniques such as the decision tree (a hierarchical approach), random forest (an ensemble hierarchical approach), and deep learning (a layered approach) are highly suitable for the system that addresses big data classification problems. Therefore, the goal of this book is to present some of the machine learning models and algorithms, and discuss them with examples.

The general objective of this book is to help readers, especially students and newcomers to the field of big data and machine learning, to gain a quick understanding of the techniques and technologies; therefore, the theory, examples, and programs (Matlab and R) presented in this book have been simplified, hardcoded, repeated, or spaced for improvements. They provide vehicles to test and understand the complicated concepts of various topics in the field. It is expected that the readers adopt these programs to experiment with the examples, and then modify or write their own programs toward advancing their knowledge for solving more complex and challenging problems.

The presentation format of this book focuses on simplicity, readability, and dependability so that both undergraduate and graduate students as well as new researchers, developers, and practitioners in this field can easily trust and grasp the concepts, and learn them effectively. The goal of the writing style is to reduce the mathematical complexity and help the vast majority of readers to understand the topics and get interested in the field. This book consists of four parts, with a total of 14 chapters. Part I mainly focuses on the topics that are needed to help analyze and understand big data. Part II covers the topics that can explain the systems required for processing big data. Part III presents the topics required to understand and select machine learning techniques to classify big data. Finally, Part IV concentrates on the topics that explain the scaling-up machine learning, an important solution for modern big data problems.

Greensboro, NC, USA Shan Suthaharan

Acknowledgements

The journey of writing this book would not have been possible without the support of many people, including my collaborators, colleagues, students, and family. I would like to thank all of them for their support and contributions toward the successful development of this book. First, I would like to thank Mr. Matthew Amboy (Editor, *Business and Economics: OR and MS*, Springer Science+Business Media) for giving me an opportunity to write this book. I would also like to thank both Ms. Christine Crigler (Assistant Editor) and Mr. Amboy for helping me throughout the publication process.

I am grateful to Professors Ratnasingham Shivaji (Head of the Department of Mathematics and Statistics at the University of North Carolina at Greensboro) and Fadil Santosa (Director of the Institute for Mathematics and its Applications at University of Minnesota) for the opportunities that they gave me to attend a machine learning workshop at the institute. Professors Bin Yu (Department of Statistics, University of California, Berkeley) and David Madigan (Department of Statistics, Columbia University) delivered an excellent short course on applied statistics and machine learning at the institute, and the topics covered in this course motivated me and equipped me with techniques and tools to write various topics in this book. My sincere thanks go to them. I would also like to thank Jinzhu Jia, Adams Bloniaz, and Antony Joseph, the members of Professor Bin Yu's research group at the Department of Statistics, University of California, Berkeley, for their valuable discussions in many machine learning topics.

My appreciation goes out to University of California, Berkeley, and University of North Carolina at Greensboro for their financial support and the research assignment award in 2013 to attend University of California, Berkeley as a Visiting scholar—this visit helped me better understand the deep learning techniques. I would also like to show my appreciation to Mr. Brent Ladd (Director of Education, Center for the Science of Information, Purdue University) and Mr. Robert Brown (Managing Director, Center for the Science of Information, Purdue University) for their support to develop a course on big data analytics and machine learning at University of North Carolina at Greensboro through a sub-award approved by the National Science Foundation. I am also thankful to Professor Richard Smith, Director of the

Statistical and Applied Mathematical Sciences Institute at North Carolina, for the opportunity to attend the workshops on low-dimensional structure in high-dimensional systems and to conduct research at the institute as a visiting research fellow during spring 2014. I greatly appreciate the resources that he provided during this visiting appointment. I also greatly appreciate the support and resources that the University of North Carolina at Greensboro provided during the development of this book.

The research work conducted with Professor Vaithilingam Jeyakumar and Dr. Guoyin Li at the University of New South Wales (Australia) helped me simplify the explanation of support vector machines. The technical report written by Michelle Dunbar under Professor Jeyakumar's supervision also contributed to the enhancement of the chapter on support vector machines. I would also like to express my gratitude to Professors Sat Gupta, Scott Richter, and Edward Hellen for sharing their knowledge of some of the statistical and mathematical techniques. Professor Steve Tate's support and encouragement, as the department head and as a colleague, helped me engage in this challenging book project for the last three semesters. My sincere gratitude also goes out to Professor Jing Deng for his support and engagement in some of my research activities.

My sincere thanks also go to the following students who recently contributed directly or indirectly to my research and knowledge that helped me develop some of the topics presented in this book: Piyush Agarwal, Mokhaled Abd Allah, Michelle Bayait, Swarna Bonam, Chris Cain, Tejo Sindhu Chennupati, Andrei Craddock, Luning Deng, Anudeep Katangoori, Sweta Keshpagu, Kiranmayi Kotipalli, Varnika Mittal, Chitra Reddy Musku, Meghana Narasimhan, Archana Polisetti, Chadwik Rabe, Naga Padmaja Tirumal Reddy, Tyler Wendell, and Sumanth Reddy Yanala.

Finally, I would like to thank my wife, Manimehala Suthaharan, and my lovely children, Lovepriya Suthaharan, Praveen Suthaharan, and Prattheeba Suthaharan, for their understanding, encouragement, and support which helped me accomplish this project. This project would not have been completed successfully without their support.

Greensboro, NC, USA Shan Suthaharan
June 2015

About the Author

Shan Suthaharan is a Professor of Computer Science at the University of North Carolina at Greensboro (UNCG), North Carolina, USA. He also serves as the Director of Undergraduate Studies at the Department of Computer Science at UNCG. He has more than 25 years of university teaching and administrative experience and has taught both undergraduate and graduate courses. His aspiration is to educate and train students so that they can prosper in the computer field by understanding current real-world and complex problems, and develop efficient techniques and technologies. His current teaching interests include big data analytics and machine learning, cryptography and network security, and computer networking and analysis. He earned his doctorate in Computer Science from Monash University, Australia. Since then, he has been actively working on disseminating his knowledge and experience through teaching, advising, seminars, research, and publications.

Dr. Suthaharan enjoys investigating real-world, complex problems, and developing and implementing algorithms to solve those problems using modern technologies. The main theme of his current research is the signature discovery and event detection for a secure and reliable environment. The ultimate goal of his research is to build a secure and reliable environment using modern and emerging technologies. His current research primarily focuses on the characterization and detection of environmental events, the exploration of machine learning techniques, and the development of advanced statistical and computational techniques to discover key signatures and detect emerging events from structured and unstructured big data.

Dr. Suthaharan has authored or co-authored more than 75 research papers in the areas of computer science, and published them in international journals and refereed conference proceedings. He also invented a key management and encryption technology, which has been patented in Australia, Japan, and Singapore. He also received visiting scholar awards from and served as a visiting researcher at the University of Sydney, Australia; the University of Melbourne, Australia; and the University of California, Berkeley, USA. He was a senior member of the Institute of Electrical and Electronics Engineers, and volunteered as an elected chair of the Central North Carolina Section twice. He is a member of Sigma Xi, the Scientific Research Society and a Fellow of the Institution of Engineering and Technology.

Contents