Ansgar Steland
Kwok-Leung Tsui *Editors*

Artificial Intelligence, Big Data and Data Science in Statistics

Challenges and Solutions in Environmetrics, the Natural Sciences and Technology

Springer

Artificial Intelligence, Big Data and Data Science in Statistics

Ansgar Steland • Kwok-Leung Tsui
Editors

Artificial Intelligence, Big Data and Data Science in Statistics

Challenges and Solutions in Environmetrics, the Natural Sciences and Technology

Springer

Editors

Ansgar Steland
Institute of Statistics and AI Center
RWTH Aachen University
Aachen, Germany

Kwok-Leung Tsui
Grado Department of Industrial
and Systems Engineering
Virginia Polytechnic Institute and State
University
Blacksburg, VA, USA

ISBN 978-3-031-07154-6 ISBN 978-3-031-07155-3 (eBook)
https://doi.org/10.1007/978-3-031-07155-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The change to data-centrism in many fields, the need to extract information and knowledge from big data, and the increasing success of machine learning (ML) and artificial intelligence (AI) have created both opportunities and challenges to the field of statistics. These developments have, to some extent, led to the creation of data science, partially regarded as a new discipline, related to statistics and computer science. The intersections among ML/AI, data science, and statistics are much larger than people expect, particularly on theory, models, practical methods, and problems under investigation. All communities can learn a lot from each other.

The impressive successes of ML and AI methods, especially deep learners and convolutional networks, in many practical problems might seem to devalue statistical approaches. Quite a few researchers as well as practitioners regard machine learning as being more focused on problem solving and benchmark data sets than statistics. But, on the other hand, ML solutions are often tailored to a specific problem and thus can be difficult to generalize and implement for a wide range of applications.

Further, there is wide range of problems related to data for which statistics provides more appropriate or even optimal solutions and allows specific interpretable models. Stochastic models often provide mathematical descriptions of physical processes rather than relying on black boxes. Indeed, lack of model interpretability, potential bias, causality, and stability, and why and when deep learners may work are common questions for the ML approaches. Statistical thinking and approaches are good alternatives to rectify these problems, in terms of both theories, models, and practical methods. A further issue where statistics is indispensable is the question whether a given data set satisfies proper sampling designs, as studied by statistical sampling theory, and the sound statistical preprocessing, handling, and cleaning of data. Both topics are important to evaluate given data, to ensure high data quality, and to clarify what can be learnt from a certain data set. On the other hand, the flexibility of many ML and AI methods may yield superior results when reliable first-class data from well-selected variables are not available and one has to rely on noisy and surrogate data.

Focusing on environmental science, natural science, and technology, this book contributes to the discussions of various issues and general interplay among statistics, data science, machine learning, and artificial intelligence. The chapters cover theoretical studies of machine learning methods, expositions of general methodologies for sound statistical analyses of data, as well as novel approaches for modeling and analyzing data in specific areas and problems. In terms of applications, the chapters deal with data as arising in industrial quality control, autonomous driving, transportation and traffic, chip manufacturing, photovoltaics, football, transmission of infectious diseases, Covid-19, and public health.

The idea for this volume came from the meetings of the Section on Environmetrics, Natural Science and Technology of Deutsche Statistische Gesellschaft of the last few years, and most authors have presented research at the annual conferences Statistische Woche. All chapters of this volume have been peer reviewed, and the editors are grateful to those colleagues who helped in the evaluation process as anonymous reviewers. Nevertheless, the authors of each chapters are solely responsible for their work.

Aachen, Germany
Blacksburg, VA, USA
November 2021

Ansgar Steland
Kwok-Leung Tsui

Contents

Part II Challenges and Solutions in Applications