Cognitive Technologies

Pavel Brazdil
Jan N. van Rijn
Carlos Soares
Joaquin Vanschoren

Metalearning

Applications to Automated Machine Learning and Data Mining

Second Edition

OPEN ACCESS

Springer

Cognitive Technologies

Titles in this series now included in the Thomson Reuters Book Citation Index and Scopus!

The Cognitive Technologies (CT) series is committed to the timely publishing of high-quality manuscripts that promote the development of cognitive technologies and systems on the basis of artificial intelligence, image processing and understanding, natural language processing, machine learning and human-computer interaction.

It brings together the latest developments in all areas of this multidisciplinary topic, ranging from theories and algorithms to various important applications. The intended readership includes research students and researchers in computer science, computer engineering, cognitive science, electrical engineering, data science and related fields seeking a convenient way to track the latest findings on the foundations, methodologies and key applications of cognitive technologies.

The series provides a publishing and communication platform for all cognitive technologies topics, including but not limited to these most recent examples:

- Interactive machine learning, interactive deep learning, machine teaching
- Explainability (XAI), transparency, robustness of AI and trustworthy AI
- Knowledge representation, automated reasoning, multiagent systems
- Common sense modelling, context-based interpretation, hybrid cognitive technologies
- Human-centered design, socio-technical systems, human-robot interaction, cognitive robotics
- Learning with small datasets, never-ending learning, metacognition and introspection
- Intelligent decision support systems, prediction systems and warning systems
- Special transfer topics such as CT for computational sustainability, CT in business applications and CT in mobile robotic systems

The series includes monographs, introductory and advanced textbooks, state-of-the-art collections, and handbooks. In addition, it supports publishing in Open Access mode.

More information about this series at https://link.springer.com/bookseries/5216

Pavel Brazdil · Jan N. van Rijn ·
Carlos Soares · Joaquin Vanschoren

Metalearning

Applications to Automated Machine Learning and Data Mining

Second Edition

Springer

Pavel Brazdil
INESC TEC - LIAAD and FEP
University of Porto
Porto, Portugal

Carlos Soares
Fraunhofer AICOS Portugal
and Laboratory for Artificial
Intelligence and Computer Science
Faculdade de Engenharia
Universidade do Porto
Porto, Portugal

Jan N. van Rijn
Leiden Institute of Advanced
Computer Science
Leiden University
Leiden, The Netherlands

Joaquin Vanschoren
Department of Mathematics
and Computer Science
Technische Universiteit Eindhoven
Eindhoven, The Netherlands

ISSN 1611-2482 ISSN 2197-6635 (electronic)
Cognitive Technologies
ISBN 978-3-030-67023-8 ISBN 978-3-030-67024-5 (eBook)
https://doi.org/10.1007/978-3-030-67024-5

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

To my lifelong companion Fátima, and to Oliver and Jakub

Pavel

To Nico and Leuntje van Rijn, for teaching me
what is important in life

Jan

To my parents and also to Manuela, Quica, Manel, and Artur

Carlos

To Ada, Elias, Kobe, and Veerle, for reminding me
how wonder-full the world is

Joaquin

Preface

The first edition of this book was published in 2009, that is at the moment of writing already more than 10 years ago. As this area has progressed substantially, we decided to prepare the second edition. Our aim was to incorporate the most important advances, so that the new version would present an up-to-date account of the area and be useful to researchers, postgraduate students, and practitioners active in this area.

What are the major changes? First, if we just compare the number of chapters of the two editions, we note that it has doubled. So did roughly the number of pages.

We note that, at the time the first edition was written, the term AutoML was not yet around. So obviously we had to cover it in the new edition and also clarify its relationship to metalearning. Also, the automation of the design methods of chains of operations – nowadays referred to as pipelines or workflows – was in its infancy. So obviously we felt the need to update the existing material to keep up with this development.

In recent years the research areas of AutoML and metalearning have attracted a great deal of attention from not only researchers, but also many artificial intelligence companies, including, for instance, Google and IBM. The issue of how one can exploit metalearning to improve AutoML systems is one of the crucial questions that many researchers are trying to answer nowadays.

This book looks also into the future. As is usually the case, better understanding of some areas allows us to pose new research questions. We took care to include some in the respective chapters.

The authors of the first edition were Pavel Brazdil, Christophe Giraud-Carrier, Carlos Soares, and Ricardo Vilalta. With the extensive developments in the area, we wanted to strengthen the team by inviting Joaquin Vanschoren and Jan N. van Rijn to join the project. Unfortunately, in the end, Christophe and Ricardo were unavailable to work on the new edition. Nevertheless, the authors of the second edition are very grateful for their contributions at the onset of the project.

How This Book Was Put Together

This book comprises two parts. Part I (chapters 2–7) discusses the basic concepts and architecture of metalearning and AutoML systems, while Part II (chapters 8–15) discusses various extensions. Part III (chapters 16–18) discusses metadata organization and management (e.g., metadata repositories) and ends with concluding remarks.

Part I – Basic Concepts and Architecture

Chapter 1 starts by explaining the basic concepts used in this book, such as machine learning, metalearning, automatic machine learning, among others. It then continues with an overview of the basic architecture of a metalearning system and serves as an introduction to the rest of the book. All co-authors of the book collaborated on this chapter.

Chapter 2 focuses on ranking approaches that exploit metadata, as these are relatively easy to construct, but can still be very useful in practical applications. This chapter was written by P. Brazdil and J. N. van Rijn.[1] Chapter 3, written by P. Brazdil and J. N. van Rijn, is dedicated to the topic of evaluation of metalearning and AutoML systems. Chapter 4 discusses different dataset measures that play an important role as *metafeatures* in metalearning systems. This chapter was written by P. Brazdil and J. N. van Rijn. Chapter 5, written by P. Brazdil and J. N. van Rijn, can be seen as a continuation of Chapter 2. It discusses various metalearning approaches including, for instance, pairwise comparisons that were proposed in the past. Chapter 6 discusses hyperparameter optimization. It covers both basic search methods and also more advanced ones introduced in the area of automated machine learning (AutoML). This chapter was written by P. Brazdil, J. N. van Rijn, and J. Vanschoren. Chapter 7 discusses the problem of automating the construction of workflows or pipelines, representing sequences of operations. This chapter was written by P. Brazdil, but it reused some material from the first edition which was prepared by C. Giraud-Carrier.

Part II – Advanced Techniques and Methods

Part 2 (chapters 8–15) continues with the topics in Part I, but covers different extensions of the basic methodology. Chapter 8, written by P. Brazdil and J. N. van Rijn, is dedicated to the topic of the design of configuration spaces and how to plan experiments. The two subsequent chapters discuss the specific topic of ensembles. Chapter 9, written by C. Giraud-Carrier, represents an invited chapter in this book. It describes different ways of organizing a set of base-level algorithms into ensembles. The authors of the second edition did not see any need to change this material, and so it is kept as it appeared in the first edition.

[1] Parts of Chapters 2 and 3 of the first edition, written by C. Soares and P. Brazdil, were reused and adapted for this chapter.

Chapter 10 continues with the topic of ensembles and shows how metalearning can be exploited in the construction of ensembles (ensemble learning). This chapter was written by C. Soares and P. Brazdil. The subsequent chapters are dedicated to rather specific topics. Chapter 11, written by J. N. van Rijn, describes how one can use metalearning to provide algorithm recommendations in data stream settings. Chapter 12, written by R. Vilalta and M. Meskhi, covers the transfer of meta-models and represents the second invited chapter of this book. It represents a substantial update of the similar chapter in the first edition, which was written by R. Vilalta. Chapter 13, written by M. Huisman, J. N. van Rijn, and A. Plaat, discusses metalearning in deep neural networks and represents the third invited chapter of this book. Chapter 14 is dedicated to the relatively new topic of automating Data Science. This chapter was drafted by P. Brazdil and incorporates various contributions and suggestions of his co-authors. The aim is to discuss various operations normally carried out within Data Science and to consider whether automation is possible and whether meta-knowledge can be exploited in this process. The aim of Chapter 15 is also to look into the future and consider whether it is possible to automate the design of more complex solutions. This chapter was written by P. Brazdil. These may involve not only pipelines of operations, but also more complex control structures (e.g., iteration), and automatic changes in the underlying representation.

Part III – Organizing and Exploiting Metadata

Part III covers some practical issues and includes the final three chapters (16–18). Chapter 16, written by J. Vanschoren and J. N. van Rijn, discusses repositories of metadata, and in particular the repository known under the name *OpenML*. This repository includes machine-usable data on many machine learning experiments carried in the past and the corresponding results. Chapter 17, written by J. N. van Rijn and J. Vanschoren, shows how the metadata can be explored to obtain further insights in machine learning and metalearning research and thereby obtain new effective practical systems. Chapter 18 ends the book with brief concluding remarks about the role of metaknowledge and also presents some future challenges. The first version was elaborated by P. Brazdil, but includes various contributions of other co-authors, in particular of J. N. van Rijn and C. Soares.

Acknowledgements

We wish to express our gratitude to all those who have helped in bringing this project to fruition.

We acknowledge the support of grant 612.001.206 from the Dutch Research Council (NWO) for granting the 'Open Access Books' funding, making the book publicly available.

P. Brazdil is grateful to the University of Porto, Faculty of Economics, R&D institution INESC TEC, and one of its centres, namely the *Laboratory of Artificial*

Intelligence and Decision Support (LIAAD), for their continued support. The work carried out was partly supported by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

J. N. van Rijn is grateful to the University of Freiburg, the Data Science Institute (DSI) at Columbia University, and Leiden Institute of Advanced Computer Science (LIACS) at Leiden University, for their support throughout this project.

C. Soares expresses his gratitude to the University of Porto and to the Faculty of Engineering for their support.

J. Vanschoren is grateful to the Eindhoven University of Technology and to the Data Mining Group for their support.

We are greatly indebted to many of our colleagues for may useful discussions and suggestions that are reflected in this book: Salisu M. Abdulrahman, Bernd Bischl, Hendrik Blockeel, Isabelle Guyon, Holger Hoos, Geoffrey Holmes, Frank Hutter, João Gama, Rui Leite, Andreas Mueller, Bernhard Pfahringer, Ashwin Srinivasan, and Martin Wistuba.

We also acknowledge the influence of many other researchers resulting from various encounters and discussions, in person or by email: Herke van Hoof, Peter Flach, Pavel Kordík, Tom Mitchell, Katharina Morik, Sergey Muravyov, Aske Plaat, Ricardo Prudêncio, Luc De Raedt, Michele Sebag, and Kate Smith-Miles.

We wish to thank also many others with whom we have collaborated: Pedro Abreu, Mitra Baratchi, Bilge Celik, Vítor Cerqueira, André Correia, Afonso Costa, Tiago Cunha, Katharina Eggensperger, Matthias Feurer, Pieter Gijsbers, Carlos Gomes, Taciana Gomes, Hendrik Jan Hoogeboom, Mike Huisman, Matthias König, Lars Kotthoff, Jorge Kanda, Aaron Klein, Walter Kosters, Marius Lindauer, Marta Mercier, Péricles Miranda, Felix Mohr, Mohammad Nozari, Sílvia Nunes, Catarina Oliveira, Marcos L. de Paula Bueno, Florian Pfisterer, Fábio Pinto, Peter van der Putten, Sahi Ravi, Adriano Rivolli, André Rossi, Cláudio Sá, Prabhant Singh, Arthur Sousa, Bruno Souza, Frank W. Takes, and Jonathan K. Vis. We are grateful also to the OpenML community, for their efforts to make machine learning research reproducible.

We are also grateful to our editor, Ronan Nugent from Springer, for his patience and encouragement throughout this project.

We wish to express our thanks to Manuel Caramelo for his careful proofreading of the draft of the whole book and for suggesting many corrections.

Porto, Eindhoven, Leiden — *Pavel Brazdil*, *Jan N. van Rijn*, *Carlos Soares*, *Joaquin Vanschoren*

March 2021

Contents