Intelligent Systems Reference Library

Volume 147

The aim of this series is to publish a Reference Library, including novel advances and developments in all aspects of Intelligent Systems in an easily accessible and well structured form. The series includes reference works, handbooks, compendia, textbooks, well-structured monographs, dictionaries, and encyclopedias. It contains well integrated knowledge and current information in the field of Intelligent Systems. The series covers the theory, applications, and design methods of Intelligent Systems. Virtually all disciplines such as engineering, computer science, avionics, business, e-commerce, environment, healthcare, physics and life science are included. The list of topics spans all the areas of modern intelligent systems such as: Ambient intelligence, Computational intelligence, Social intelligence, Computational neuroscience, Artificial life, Virtual society, Cognitive systems, DNA and immunity-based systems, e-Learning and teaching, Human-centred computing and Machine ethics, Intelligent control, Intelligent data analysis, Knowledge-based paradigms, Knowledge management, Intelligent agents, Intelligent decision making, Intelligent network security, Interactive entertainment, Learning paradigms, Recommender systems, Robotics and Mechatronics including human-machine teaming, Self-organizing and adaptive systems, Soft computing including Neural systems, Fuzzy systems, Evolutionary computing and the Fusion of these paradigms, Perception and Vision, Web intelligence and Multimedia.

More information about this series at http://www.springer.com/series/8578

Verónica Bolón-Canedo · Amparo Alonso-Betanzos

Recent Advances in Ensembles for Feature Selection

Springer

Verónica Bolón-Canedo
Facultad de Informática
Universidade da Coruña
A Coruña
Spain

Amparo Alonso-Betanzos
Facultad de Informática
Universidade da Coruña
A Coruña
Spain

ISSN 1868-4394 ISSN 1868-4408 (electronic)
Intelligent Systems Reference Library
ISBN 978-3-319-90079-7 ISBN 978-3-319-90080-3 (eBook)
https://doi.org/10.1007/978-3-319-90080-3

Library of Congress Control Number: 2018938365

Printed on acid-free paper

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

To our children: Iago, Alberto, Leo and Olivia

Foreword

Ensemble methods are now a cornerstone of modern Machine Learning and Data Science, the “go-to” tool that everyone uses by default, to grab that last 3–4% of predictive accuracy. Feature Selection methods too, are a critical element, throughout the data science pipeline, from exploratory data analysis to predictive model building. There can scarcely be a more generically relevant challenge, than a meaningful synthesis of the two. This is the challenge that the authors here have taken on, with gusto.

Bolón-Canedo and Alonso-Betanzos present a meticulously thorough treatment of literature to date, comparing and contrasting elements practical for applications, and interesting for theoreticians. I was surprised to find several new references I had not found myself, in several years of working on these topics.

The first half of the book presents tutorials, cross referenced to current literature and thinking—this should prove a very useful launch-pad for students wanting to get into the area. The second half presents more advanced topics and issues—from appropriate evaluation protocols (it’s really not simple, and certainly not a done deal yet), to still quite open questions (e.g. combination of ranks and the stability of algorithms), through to software tips and tools for practitioners. I particularly like Chapter 10, on emerging challenges. This sort of chapter points the way for new PhDs, providing inspiration and confidence that your research is moving in the right direction.

In summary, Bolón-Canedo and Alonso-Betanzos have authored an eloquent and authoritative treatment of this important area—something I will be recommending to my students and colleagues as essential reference material.

University of Manchester Prof. Gavin Brown
2018

Preface

Classically, machine learning methods have used a single learning model to solve a given problem. However, the technique of using multiple prediction models for solving the same problem, known as ensemble learning, has proven its effectiveness over the last few years. The idea builds on the assumption that combining the output of multiple experts is better than the output of any single expert. Classifier ensembles have flourished into a prolific discipline; in fact, there is a series of workshops on Multiple Classifier Systems (MCSs) run since 2000 by Fabio Roli and Josef Kittler.

However, ensemble learning can be also thought as means of improving other machine learning disciplines such as feature selection, which has not received yet the same amount of attention. There exists a vast body of feature selection methods in the literature, including filters based on distinct metrics (e.g. entropy, probability distributions or information theory) and embedded and wrapper methods using different induction algorithms. The proliferation of feature selection algorithms, however, has not brought about a general methodology that allows for intelligent selection from existing algorithms. In order to make a correct choice, a user not only needs to know the domain well but also is expected to understand technical details of available algorithms.

Ensemble feature selection can be a solution for the aforementioned problem since, by combining the output of several feature selectors, the performance can be usually improved and the user is released from having to choose a single method. This book aims at offering a general and comprehensive overview of ensemble learning in the field of feature selection.

Ensembles for feature selection can be classified into homogeneous (the same base feature selector) and heterogeneous (different feature selectors). Moreover, it is necessary to combine the partial outputs that can be either in the form of subsets of features or in the form of rankings of features. This book stresses the gap with standard ensemble learning and its application to feature selection, showing the particular issues that researchers have to deal with. Specifically, it reviews different techniques for combination of partial results, measures of diversity and evaluation of the ensemble performance. Finally, this book also shows examples of problems in which ensembles for feature selection have applied in a successful way and introduces the new challenges and possibilities that researchers must acknowledge, especially since the advent of Big Data.

The target audience of this book comprises anyone interested in the field of ensembles for feature selection. Researchers could take advantage of this extensive review on recent advances on the field and gather new ideas from the emerging challenges described. Practitioners in industry should find new directions and opportunities from the topics covered. Finally, we hope our readers enjoy reading this book as much as we enjoyed writing it.

We are thankful to all our collaborators, who helped with some of the research involved in this book. We would also like to acknowledge our families and friends for their invaluable support, and not only during this writing process.

A Coruña, Spain
March 2018

Verónica Bolón-Canedo
Amparo Alonso-Betanzos

Contents