

Intelligent Systems Reference Library

Volume 147

Series editors

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

Lakhmi C. Jain, University of Canberra, Canberra, Australia;
Bournemouth University, UK;
KES International, UK
e-mail: jainlc2002@yahoo.co.uk; jainlakhmi@gmail.com
URL: <http://www.kesinternational.org/organisation.php>

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

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

© Springer International Publishing AG, part of Springer Nature 2018

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. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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
2018

Prof. Gavin Brown

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

- 1 Basic Concepts** 1
 - 1.1 What Is a Dataset, Feature and Class? 1
 - 1.2 Classification Error/Accuracy 4
 - 1.3 Training and Testing 5
 - 1.4 Comparison of Models: Statistical Tests 7
 - 1.4.1 Two Models and a Single Dataset 8
 - 1.4.2 Two Models and Multiple Dataset 9
 - 1.4.3 Multiple Models and Multiple Dataset 9
 - 1.5 Data Repositories 9
 - 1.6 Summary 10
 - References 11
- 2 Feature Selection** 13
 - 2.1 Foundations of Feature Selection 14
 - 2.2 State-of-the-Art Feature Selection Methods 16
 - 2.2.1 Filter Methods 16
 - 2.2.2 Embedded Methods 18
 - 2.2.3 Wrapper Methods 18
 - 2.3 Which Is the Best Feature Selection Method? 18
 - 2.3.1 Datasets 19
 - 2.3.2 Experimental Study 20
 - 2.4 On the Scalability of Feature Selection Methods 31
 - 2.4.1 Experimental Study 31
 - 2.5 Summary 35
 - References 35

3	Foundations of Ensemble Learning	39
3.1	The Rationale of the Approach	39
3.2	Most Popular Methods	41
3.2.1	Boosting	44
3.2.2	Bagging	46
3.3	Summary	49
	References	49
4	Ensembles for Feature Selection	53
4.1	Introduction	54
4.2	Homogeneous Ensembles for Feature Selection	56
4.2.1	A Use Case: Homogeneous Ensembles for Feature Selection Using Ranker Methods	56
4.3	Heterogeneous Ensembles for Feature Selection	73
4.3.1	A Use Case: Heterogeneous Ensemble for Feature Selection Using Ranker Methods	74
4.4	A Comparison on the Result of Both Use Cases: Homogeneous Versus Heterogeneous Ensemble for Feature Selection Using Ranker Methods	77
4.5	Summary	79
	References	80
5	Combination of Outputs	83
5.1	Combination of Label Predictions	83
5.1.1	Majority Vote	84
5.1.2	Decision Rules	85
5.2	Combination of Subsets of Features	86
5.2.1	Intersection and Union	87
5.2.2	Using Classification Accuracy	87
5.2.3	Using Complexity Measures	89
5.3	Combination of Rankings of Features	90
5.3.1	Simple Operations Between Ranks	93
5.3.2	Stuart Aggregation Method	94
5.3.3	Robust Rank Aggregation	94
5.3.4	SVM-Rank	95
5.4	Summary	95
	References	96
6	Evaluation of Ensembles for Feature Selection	97
6.1	Introduction	97
6.2	Diversity	99
6.3	Stability	103
6.3.1	Stability of Subsets of Features	103
6.3.2	Stability of Rankings of Features	104

6.4	Performance of Ensembles	106
6.4.1	Are the Selected Features the Relevant Ones?	106
6.4.2	The Ultimate Evaluation: Classification Performance	108
6.5	Summary	110
	References	111
7	Other Ensemble Approaches	115
7.1	Introduction	116
7.2	Ensembles for Classification	118
7.2.1	One-Class Classification	119
7.2.2	Imbalanced Data	120
7.2.3	Data Streaming	121
7.2.4	Missing Data	123
7.3	Ensembles for Quantification	125
7.4	Ensembles for Clustering	126
7.4.1	Types of Clustering Ensembles	130
7.5	Ensembles for Other Preprocessing Steps: Discretization	132
7.6	Summary	133
	References	133
8	Applications of Ensembles Versus Traditional Approaches: Experimental Results	139
8.1	The Rationale of the Approach	139
8.2	The Process of Selecting the Methods for the Ensemble	141
8.3	Two Filter Ensemble Approaches	142
8.3.1	Ensemble 1	142
8.3.2	Ensemble 2	143
8.4	Experimental Setup	145
8.5	Experimental Results	146
8.5.1	Results on Synthetic Data	147
8.5.2	Results on Classical Datasets	147
8.5.3	Results on Microarray Data	151
8.5.4	The Imbalance Problem	154
8.6	Summary	155
	References	156
9	Software Tools	157
9.1	Popular Software Tools	157
9.1.1	Matlab	157
9.1.2	Weka	158
9.1.3	R	160
9.1.4	KEEL	161
9.1.5	RapidMiner	162

9.1.6	Scikit-Learn	163
9.1.7	Parallel Learning	164
9.2	Code Examples	166
9.2.1	Example: Building an Ensemble of Trees	166
9.2.2	Example: Adding Feature Selection to Our Ensemble of Trees	167
9.2.3	Example: Exploring Different Ensemble Sizes for Our Ensemble	168
	References	171
10	Emerging Challenges	173
10.1	Introduction	174
10.2	Recent Contributions in Feature Selection	175
10.2.1	Applications	176
10.3	The Future: Challenges Ahead for Feature Selection	182
10.3.1	Millions of Dimensions	183
10.3.2	Scalability	184
10.3.3	Distributed Feature Selection	186
10.3.4	Real-Time Processing	189
10.3.5	Feature Cost	191
10.3.6	Missing Data	194
10.3.7	Visualization and Interpretability	195
10.4	Summary	196
	References	197