

Artificial Intelligence: Foundations, Theory, and Algorithms

Series Editors

Barry O'Sullivan, Cork, Ireland

Michael Wooldridge, Oxford, United Kingdom

More information about this series at <http://www.springer.com/series/13900>

Verónica Bolón-Canedo • Noelia Sánchez-Marroño
Amparo Alonso-Betanzos

Feature Selection for High-Dimensional Data

 Springer

Verónica Bolón-Canedo
Facultad de Informática
Universidad de A Coruña
A Coruña, Spain

Noelia Sánchez-Maróño
Facultad de Informática
Universidad de A Coruña
A Coruña, Spain

Amparo Alonso-Betanzos
Facultad de Informática
Universidad de A Coruña
A Coruña, Spain

ISSN 2365-3051

ISSN 2365-306X (electronic)

Artificial Intelligence: Foundations, Theory, and Algorithms

ISBN 978-3-319-21857-1

ISBN 978-3-319-21858-8 (eBook)

DOI 10.1007/978-3-319-21858-8

Library of Congress Control Number: 2015951087

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2015

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)

To our families

Foreword

The topic of variable selection in high-dimensional spaces (often with hundreds or thousands of dimensions) has attracted considerable attention in data mining research in previous years, and it is common in many real problems.

In a nutshell, feature selection is a process that chooses an optimal subset of features according to a certain criterion. The selection of the criterion must be done according to the purpose of feature selection, usually with the aim of improving the prediction accuracy of the data mining algorithm used to learn a model. Generally, the objective is to identify the features in the dataset which are important and discard others as redundant or irrelevant. The problem is especially relevant when we are managing a huge number of features and the learning algorithm loses prediction capacity using all of them. Since feature selection reduces the dimensionality of the data, the data mining algorithms can run faster and obtain better outcomes by using feature selection.

The publication of the book “Feature Selection for High-Dimensional Data” written by Verónica Bolón-Canedo, Noelia Sánchez-Marroño and Amparo Alonso-Betanzos is an important event. The book offers a coherent and comprehensive approach to feature subset selection in the scope of classification problems.

We can shortly outline the three parts found when reading the book: foundations, real application problems and challenges. First, the authors focus on the analysis and synthesis of feature selection algorithms, presenting a comprehensive review of basic concepts and experimental results of the most well known algorithms. Second, an interesting novelty and contribution of the book is how it addresses different real scenarios with high-dimensional data, showing the use of feature selection algorithms in different contexts with different requirements and information: microarray data, intrusion detection, tear film lipid layer classification and cost based features. Third, the book also delves into the scenario of big dimension, sometimes combined with massive amounts of data as big data. It pays attention to important problems under high-dimensional spaces: scalability, distributed processing and real-time processing. These are scenarios that open up new and interesting challenges for researchers.

This triple orientation makes it a different and original book, written in a very clear and comprehensive style. This book allows readers to delve into feature selection from both the theoretical and practical perspective. Furthermore, it shows the major challenges in this well established field, which in turn is also a very dynamic one because of its great importance in data preprocessing.

The book is authored by great experts in the field, who make an important contribution to the topic. It is a must read for anyone who is concerned with the design or application of data mining algorithms for getting knowledge from high-dimensional data.

Granada, Spain
March 2015

Francisco Herrera

Preface

Feature selection (FS) has been embraced as one of the high activity research areas during the last few years, due to the appearance of datasets containing hundreds of thousands of features (variables). Thus, feature selection was employed to be able to better model the underlying process of data generation, and to reduce the cost of acquiring the variables. Furthermore, from the Machine Learning algorithms viewpoint, as FS is able to reduce the dimensionality of the problem, it can be used for maintaining and, most of the time, improving the algorithms' performance, while reducing computational costs. Nowadays, the advent of Big Data has brought unprecedented challenges to machine learning researchers, who now have to deal with huge volumes of data, in terms of both instances and features, making the learning task more complex and computationally demanding. Specifically, when dealing with an extremely large number of input features, learning algorithms' performance can degenerate due to overfitting, learned models decrease their interpretability as they become more complex, and finally speed and efficiency of the algorithms decline in accordance with size.

A vast body of feature selection methods exist 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 the technical details of available algorithms. On top of this, most algorithms were developed when dataset sizes were much smaller, but, nowadays, distinct trade-offs are required for the case of small-scale and large-scale learning problems. Small-scale learning problems are subject to the usual approximation-estimation trade-off. In the case of large-scale learning problems, the trade-off is more complex because it involves not only the accuracy of the selection but also other aspects, such as stability (i.e., the sensitivity of the results to training set variations) or scalability. All these aspects are addressed in the first four chapters of the book, to give the reader not only a broad perspective of the state of the art, but also a critical review of the behavior of the methods in different situa-

tions, such as noise, correlation of the variables, redundancy, etc. In this way, the advanced reader and the researcher can have a reference against which to compare the results of new methods.

Also, the book addresses the “big” feature dimensionality factor of the data, which has received much less attention than the “Big Instance Size” factor of the data (devoted to the large number of instances). It is intended to be a comprehensive book completely devoted to analyzing the evolution of feature dimensionality in the last few decades, the state-of-the-art feature selection methods and the emerging challenges in this field. This aspect is emphasized in Chapter 4, devoted to the microarray datasets, a challenge for feature selection due to their small sample size (in the order of hundreds), but large number of variables (in the order of thousands). Chapter 5, on the other hand, explores two real applications that situate the reader in the contexts related with imbalanced datasets, the importance of the cost of selecting features, etc.

Finally, Chapter 6 is devoted to the emerging challenges of the discipline. Feature selection is often regarded as one of the most important tasks in the omnipresent actual scenario of Big Data research, since the emersion of high-dimensionality requires feature selection strategies that are capable of coping with the explosion of features. As an example, state-of-the-art feature selection methods such as minimum Redundancy Maximum Relevance (mRMR) or Support Vector Machine based on Recursive Feature Elimination (SVM-RFE) take more than a day of computational effort to deal with datasets composed of 0.5 million features. Thus, an exhaustive and updated background in the topic could be very useful for the data analytics and computational intelligence communities since it underlines the emerging trend of high-dimensionality and deliberates on how the existing methods are prepared to face the arising challenges. This book invites readers to explore the different issues associated with feature selection for high-dimensional data:

- The evolution of feature dimensionality in the last few decades.
- State-of-the-art feature selection methods.
- Adequacy of feature selection methods to solve real problems.
- Emerging challenges for feature selection.

The target audience of this book comprises anyone interested in improving their understanding of feature selection. Researchers could take advantage of this extensive review on state-of-the-art feature selection methods 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 the adventure of its production.

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. Last but not least, we are indebted and grateful to Francisco Herrera, from the University of Granada, who encouraged us to compile this book.

Preface

xi

A Coruña, Spain
March 2015

Verónica Bolón-Canedo
Noelia Sánchez-Marño
Amparo Alonso-Betanzos

Contents

1	Introduction to High-Dimensionality	1
1.1	The Need for Feature Selection	2
1.2	When Features Are Born	3
1.3	Intrinsic Characteristics of Data	4
1.3.1	Small Sample Size	4
1.3.2	Class Imbalance	5
1.3.3	Data Complexity	6
1.3.4	Dataset Shift	7
1.3.5	Noisy Data	8
1.3.6	Outliers	9
1.3.7	Feature Cost	9
1.4	A Guide for the Reader	9
	References	10
2	Foundations of Feature Selection	13
2.1	Feature Selection	14
2.1.1	Feature Relevance	14
2.1.2	Feature Redundancy	15
2.2	Feature Selection Methods	15
2.2.1	Filter Methods	17
2.2.2	Embedded Methods	24
2.2.3	Wrapper Methods	25
2.2.4	Other Approaches	26
2.3	Summary	26
	References	26
3	A Critical Review of Feature Selection Methods	29
3.1	Existing Reviews of Feature Selection Methods	30
3.2	Experimental Settings	31
3.3	Experimental Results	33
3.3.1	Dealing with Correlation and Redundancy: CorrAL	34

3.3.2	Dealing with Nonlinearity: XOR and Parity	35
3.3.3	Dealing with Noise in the Inputs: Led	35
3.3.4	Dealing with Noise in the Target: Monk3	40
3.3.5	Dealing with a Complex Dataset: Madelon	43
3.4	Case Studies	44
3.4.1	Case Study I: Different Kernels for SVM-RFE	44
3.4.2	Case Study II: mRMR vs \mathcal{M}_d	46
3.4.3	Case Study III: Subset Filters	47
3.4.4	Case Study IV: Different Levels of Noise in the Input	48
3.5	Analysis and Discussion	50
3.5.1	Analysis of Success Index	50
3.5.2	Analysis of Classification Accuracy	52
3.6	Summary	56
	References	57
4	Feature Selection in DNA Microarray Classification	61
4.1	Background: The Problem and First Attempts	63
4.2	Intrinsic Characteristics of Microarray Data	64
4.2.1	Small Sample Size	64
4.2.2	Class Imbalance	64
4.2.3	Data Complexity	65
4.2.4	Dataset Shift	65
4.2.5	Outliers	67
4.3	Algorithms for Feature Selection on Microarray Data: A Review	67
4.3.1	Filters	68
4.3.2	Wrappers	70
4.3.3	Embedded	72
4.3.4	Other Algorithms	73
4.4	A Framework for Feature Selection Evaluation in Microarray Datasets	76
4.4.1	Validation Techniques	77
4.4.2	On the Datasets Characteristics	78
4.4.3	Feature Selection Methods	79
4.4.4	Evaluation Measures	79
4.5	A Practical Evaluation: Analysis of Results	79
4.5.1	Holdout Validation Study	80
4.5.2	Cross-validation Study	83
4.6	Summary	88
	References	91
5	Application of Feature Selection to Real Problems	95
5.1	Classification in Intrusion Detection Systems	96
5.1.1	Results on the Binary Case	98
5.1.2	Results on the Multiple Class Case	101
5.2	Tear Film Lipid Layer Classification	105

- 5.2.1 Classification Accuracy 110
- 5.2.2 Robustness to Noise 110
- 5.2.3 Feature Extraction Time 111
- 5.2.4 Overall Analysis 112
- 5.2.5 The Concatenation of All Methods with CFS: A Case Study 114
- 5.3 Cost-Based Feature Selection 117
 - 5.3.1 Description of the Method 117
 - 5.3.2 Experimental Results 119
- 5.4 Summary 122
- References 123

- 6 Emerging Challenges 125**
 - 6.1 Millions of Dimensions 125
 - 6.2 Scalability 126
 - 6.3 Distributed Feature Selection 127
 - 6.4 Real-Time Processing 129
 - 6.5 Summary 130
 - References 130

- A Experimental Framework Used in This Book 133**
 - A.1 Software Tools 133
 - A.2 Datasets 133
 - A.2.1 Data Repositories 134
 - A.2.2 Synthetic Datasets 135
 - A.2.3 DNA Microarray Datasets 139
 - A.3 Validation Techniques 140
 - A.3.1 k -Fold Cross-validation 140
 - A.3.2 Leave-One-Out Cross-validation 140
 - A.3.3 Bootstrap 141
 - A.3.4 Holdout Validation 141
 - A.4 Statistical Tests 141
 - A.5 Discretization Algorithms 142
 - A.6 Classification Algorithms 143
 - A.6.1 Support Vector Machine, SVM 143
 - A.6.2 Proximal Support Vector Machine, PSVM 143
 - A.6.3 C4.5 144
 - A.6.4 Naive Bayes, NB 144
 - A.6.5 k -Nearest Neighbors, k -NN 144
 - A.6.6 One-Layer Feedforward Neural Network, One-Layer NN . . . 145
 - A.7 Evaluation Measures 145
 - A.7.1 Multiple-Criteria Decision-Making 145
 - References 146