Artificial Intelligence: Foundations, Theory, and Algorithms

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

Feature Selection for High-Dimensional Data

Springer

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

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

Noelia Sánchez-Maroño
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

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-Maroño 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.

A Coruña, Spain
March 2015

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

Contents