

# Pattern Mining with Evolutionary Algorithms



Sebastián Ventura • José María Luna

# Pattern Mining with Evolutionary Algorithms

 Springer

Sebastián Ventura  
Department of Computer Science  
and Numerical Analysis  
University of Cordoba  
Cordoba, Spain

José María Luna  
Department of Computer Science  
and Numerical Analysis  
University of Cordoba  
Cordoba, Spain

ISBN 978-3-319-33857-6

ISBN 978-3-319-33858-3 (eBook)

DOI 10.1007/978-3-319-33858-3

Library of Congress Control Number: 2016939025

© Springer International Publishing Switzerland 2016

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

This Springer imprint is published by Springer Nature  
The registered company is Springer International Publishing AG Switzerland

*–Success is a journey, not a destination. The  
doing is often more important than the  
outcome–*

*Arthur Ashe.*

*To our families.*



# Preface

This book is intended to provide a general and comprehensible overview of the field of pattern mining with evolutionary algorithms. To do so, the book provides formal definitions about patterns, pattern mining, type of patterns, and the usefulness of patterns in the knowledge discovery process. As it is described within the book, the discovery process suffers from both high runtime and memory requirements, especially when high-dimensional datasets are analyzed. To solve this issue, many pruning strategies have been developed. Nevertheless, with the growing interest in the storage of information, more and more datasets comprise such a dimensionality that the discovery of interesting patterns becomes a hard process. In this regard, the use of evolutionary algorithms for mining pattern enables the computation capacity to be reduced, providing sufficiently good solutions.

The book also provides a survey on evolutionary computation with particular emphasis on genetic algorithms and genetic programming. Additionally, this book carries out an analysis of the set of quality measures most widely used in the field of pattern mining with evolutionary algorithms. This book serves as a good review on the most important evolutionary algorithms for pattern mining. In this sense, it considers the analysis of different algorithms for mining different types of patterns and relationships between patterns, such as frequent patterns, infrequent patterns, patterns defined in a continuous domain, or even positive and negative patterns.

The book also introduces a completely new problem in the pattern mining field, which is known by the name of the mining of exceptional relationships between patterns. In this problem, the goal is to identify patterns where distribution is exceptionally different from the distribution in the complete set of data records. Finally, this book deals with the subgroup discovery task, a method to identify a subgroup of interesting patterns that is related to a dependent variable or target attribute. This subgroup of patterns satisfies two essential conditions: interpretability and interestingness.

Cordoba, Spain  
February 2016

Sebastián Ventura  
José Mará Luna





# Acknowledgments

We would like to thank the Springer editorial team for giving us the opportunity to publish this book, for their great support toward the preparation and completion of this work, and for their valuable editing suggestions to improve the organization and readability of the manuscript. We also want to thank our colleagues for their valuable help during the preparation of the book, whose comments were very helpful for improving its quality.

This work was supported by the Spanish Ministry of Economy and Competitiveness under the project TIN2014-55252-P and FEDER funds.



# Contents

<b>1</b>	<b>Introduction to Pattern Mining</b>	1
1.1	Definitions	1
1.2	Type of Patterns	3
1.2.1	Frequent and Infrequent Patterns	3
1.2.2	Closed and Maximal Frequent Patterns	6
1.2.3	Positive and Negative Patterns	7
1.2.4	Continuous Patterns	9
1.2.5	Colossal Patterns	10
1.2.6	Sequential Patterns	11
1.2.7	Spatio-Temporal Patterns	12
1.3	Pattern Space Pruning	13
1.4	Traditional Approaches for Pattern Mining	15
1.5	Association Rules	22
	References	24
<b>2</b>	<b>Quality Measures in Pattern Mining</b>	27
2.1	Introduction	27
2.2	Objective Interestingness Measures	28
2.2.1	Quality Properties of a Measure	30
2.2.2	Relationship Between Quality Measures	35
2.2.3	Other Quality Properties	38
2.3	Subjective Interestingness Measures	41
	References	42
<b>3</b>	<b>Introduction to Evolutionary Computation</b>	45
3.1	Introduction	45
3.2	Genetic Algorithms	48
3.2.1	Standard Procedure	48
3.2.2	Individual Representation	50
3.2.3	Genetic Operators	50
3.3	Genetic Programming	53
3.3.1	Individual Representation	53

- 3.3.2 Genetic Operators ..... 55
    - 3.3.3 Code Bloat ..... 57
  - 3.4 Other Bio-Inspired Algorithms ..... 58
  - References ..... 59
- 4 Pattern Mining with Genetic Algorithms ..... 63**
  - 4.1 Introduction ..... 63
  - 4.2 General Issues ..... 65
    - 4.2.1 Pattern Encoding ..... 66
    - 4.2.2 Genetic Operators ..... 71
    - 4.2.3 Fitness Function ..... 73
  - 4.3 Algorithmic Approaches ..... 76
  - 4.4 Successful Applications ..... 82
  - References ..... 83
- 5 Genetic Programming in Pattern Mining ..... 87**
  - 5.1 Introduction ..... 87
  - 5.2 General Issues ..... 89
    - 5.2.1 Canonical Genetic Programming ..... 89
    - 5.2.2 Syntax-Restricted Programming ..... 93
  - 5.3 Algorithmic Approaches ..... 97
    - 5.3.1 Frequent Patterns ..... 97
    - 5.3.2 Infrequent Patterns ..... 102
    - 5.3.3 Highly Optimized Continuous Patterns ..... 107
    - 5.3.4 Mining Patterns from Relational Databases ..... 110
  - 5.4 Successful Applications ..... 114
  - References ..... 116
- 6 Multiobjective Approaches in Pattern Mining ..... 119**
  - 6.1 Introduction ..... 119
  - 6.2 General Issues ..... 120
    - 6.2.1 Multiobjective Optimization ..... 121
    - 6.2.2 Quality Indicators of the Pareto Front ..... 122
    - 6.2.3 Quality Measures to Optimize in Pattern Mining ..... 125
  - 6.3 Algorithmic Approaches ..... 127
    - 6.3.1 Genetic Algorithms ..... 127
    - 6.3.2 Genetic Programming ..... 131
    - 6.3.3 Other Algorithms ..... 135
  - 6.4 Successful Applications ..... 137
  - References ..... 137
- 7 Supervised Local Pattern Mining ..... 141**
  - 7.1 Introduction ..... 141
  - 7.2 Subgroup Discovery ..... 143
    - 7.2.1 Problem Definition ..... 143
    - 7.2.2 Quality Measures ..... 144

- 7.2.3 Deterministic Algorithms ..... 146
- 7.2.4 Evolutionary Algorithms ..... 148
- 7.3 Other Supervised Local Pattern Mining Approaches ..... 157
- References ..... 159
- 8 Mining Exceptional Relationships Between Patterns ..... 163**
  - 8.1 Introduction ..... 163
  - 8.2 Mining the Exceptionableness ..... 165
    - 8.2.1 Exceptional Model Mining Problem ..... 165
    - 8.2.2 Exceptional Relationship Mining ..... 167
  - 8.3 Algorithmic Approach ..... 169
  - 8.4 Successful Applications ..... 173
  - References ..... 175
- 9 Scalability in Pattern Mining ..... 177**
  - 9.1 Introduction ..... 177
  - 9.2 Traditional Methods for Speeding Up the Mining Process ..... 179
    - 9.2.1 The Role of Evolutionary Computation in Scalability Issues ..... 179
    - 9.2.2 Parallel Algorithms ..... 181
    - 9.2.3 New Data Structures ..... 183
  - 9.3 New Trends in Pattern Mining: Scalability Issues ..... 185
  - References ..... 188