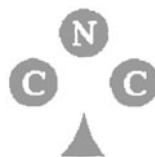


Natural Computing Series

Series Editors: G. Rozenberg (Managing)
Th. Bäck A.E. Eiben J.N. Kok H.P. Spaink
Leiden Center for Natural Computing



Advisory Board: S. Amari G. Brassard M. Conrad
K.A. De Jong C.C.A.M. Gielen T. Head L. Kari
L. Landweber T. Martinetz Z. Michalewicz M.C. Mozer
E. Oja G. Păun J. Reif H. Rubin A. Salomaa M. Schoenauer
H.-P. Schwefel D. Whitley E. Winfree J.M. Zurada

Alex A. Freitas

Data Mining and Knowledge Discovery with Evolutionary Algorithms

With 74 Figures and 10 Tables



Springer

Author

Dr. Alex A. Freitas
Computing Laboratory, University of Kent
Canterbury CT2 7NF, UK
A.A.Freitas@ukc.ac.uk

Series Editors

G. Rozenberg (Managing Editor)
Th. Bäck, A.E. Eiben, J.N. Kok, H.P. Spink

Leiden Center for Natural Computing, Leiden University
Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
rozenber@cs.leidenuniv.nl

Library of Congress Cataloging-in-Publication Data

Freitas, Alex. A., 1964–

Data mining and knowledge discovery with evolutionary algorithms/Alex A. Freitas.
p.cm. – (Natural computing series)

Includes bibliographical references and index.

ISBN 978-3-642-07763-0 ISBN 978-3-662-04923-5 (eBook)

DOI 10.1007/978-3-662-04923-5

1. Data mining. 2. Database searching. 3. Computer algorithms. I. Title. II. Series.

QA76.9.D343F722002

006.3–dc21

2002021728

ACM Computing Classification (1998): I., I.2, I.2.6

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German copyright law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

<http://www.springer.de>

© Springer-Verlag Berlin Heidelberg 2002

Originally published by Springer-Verlag Berlin Heidelberg New York in 2002

Softcover reprint of the hardcover 1st edition 2002

The use of general descriptive names, trademarks, 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.

Cover design: KünkellOpka, Heidelberg

Typesetting: Steingraeber, Heidelberg

Printed on acid-free paper SPIN: 10986554 45/3111 GF– 54321

This book is dedicated to all the people
who believe that learning is not only
one of the most necessary but also
one of the noblest human activities.

Preface

This book addresses the integration of two areas of computer science, namely data mining and evolutionary algorithms. Both these areas have become increasingly popular in the last few years, and their integration is currently an area of active research.

In essence, data mining consists of extracting valid, comprehensible, and interesting knowledge from data. Data mining is actually an interdisciplinary field, since there are many kinds of methods that can be used to extract knowledge from data. Arguably, data mining mainly uses methods from machine learning (a branch of artificial intelligence) and statistics (including statistical pattern recognition). Our discussion of data mining and evolutionary algorithms is primarily based on machine learning concepts and principles. In particular, in this book we emphasize the importance of discovering comprehensible, interesting knowledge, which the user can potentially use to make intelligent decisions.

In a nutshell, the motivation for applying evolutionary algorithms to data mining is that evolutionary algorithms are robust search methods which perform a global search in the space of candidate solutions (rules or another form of knowledge representation). In contrast, most rule induction methods perform a local, greedy search in the space of candidate rules. Intuitively, the global search of evolutionary algorithms can discover interesting rules and patterns that would be missed by the greedy search.

This book initially presents a comprehensive review of basic concepts from both data mining and evolutionary algorithms, and then it discusses significant advances in the integration of these two areas. It is a self-contained book, explaining both basic concepts and advanced topics in a clear and informal style.

My research on data mining with evolutionary algorithms has been done with the help of many people. I am grateful to my research students for their hard work. They had to carry out a time-consuming iterative process of studying a lot about both data mining and evolutionary algorithms, developing an evolutionary algorithm for data mining, performing several experiments to evaluate the algorithm, analyzing the results, refining the algorithm, etc. That was not an easy task. In particular, I would like to thank (in alphabetical order of first name) Celia C. Bojarczuk, Deborah R. Carvalho, Denise F. Tsunoda, Dieferson L.A. Araujo, Edgar Noda, Fabricio B. Voznika, Fernando E.B. Otero, Gisele L. Pappa, Marcus V. Fidelis, Monique M.S. Silva, Otavio Larsen, Rafael S. Parpinelli, Raul T. Santos, Roberto R.F. Mendes, and Wesley Romao.

I am also grateful to my colleague Dr. Heitor S. Lopes, with whom I have had many interesting discussions about evolutionary algorithms for data mining and have shared the supervision of several of the above-mentioned students.

In addition, I thank an anonymous reviewer for reviewing the manuscript of the book, my colleague Dr. L. Valeria R. Arruda for reviewing the first draft of chapter 11, and Prof. Dr. A.E. Eiben for reviewing chapter 12. I am also grateful to the team at Springer-Verlag for their assistance throughout the project.

Most of the research that led to the writing of this book has been carried out in two universities, both located in Curitiba, Brazil. More precisely, I was with CEFET-PR (Centro Federal de Educacao Tecnologica do Parana) in 1998, and since 1999 I have been with PUC-PR (Pontificia Universidade Catolica do Parana). I thank my colleagues from both institutions for creating a productive, friendly research environment.

Finally, during my research on data mining with evolutionary algorithms I have been partially financially supported by a grant from the Brazilian government's National Council of Scientific and Technological Development (CNPq), process number 300153/98-8.

Curitiba, Brazil
March 2002

Alex A. Freitas

Table of Contents

1	Introduction	1
1.1	Data Mining and Knowledge Discovery	1
1.1.1	Desirable Properties of Discovered Knowledge	2
1.2	Knowledge Representation	3
1.2.1	Prediction Rules	4
1.3	An Overview of Data Mining Paradigms	5
1.3.1	Rule Induction.....	7
1.3.2	Evolutionary Algorithms.....	9
	References	10
2	Data Mining Tasks and Concepts	13
2.1	Classification.....	13
2.1.1	The Non-determinism of the Classification Task.....	17
2.1.2	Overfitting and Underfitting	17
2.1.3	Estimating Accuracy via Training and Testing.....	19
2.2	Dependence Modeling.....	22
2.2.1	Dependence Modeling vs Association-Rule Discovery	23
2.3	The Challenge of Measuring Prediction-Rule Quality	24
2.3.1	Measuring Predictive Accuracy	24
2.3.2	Measuring Rule Comprehensibility	26
2.3.3	Measuring Rule Interestingness	27
2.4	Clustering	31
2.4.1	Hierarchical Agglomerative Clustering	33
2.4.2	The <i>K</i> -Means Algorithm	34
2.4.3	The Challenge of Measuring Clustering Quality	35
2.5	Inductive Bias.....	36
2.5.1	The Domain-Dependent Effectiveness of a Bias	37
2.5.2	The Simplicity Bias of Occam's Razor.....	38
2.5.3	The Minimum Description Length (MDL) Principle.....	39
	References	40

3	Data Mining Paradigms	45
3.1	Decision-Tree Building Algorithms	45
3.1.1	Decision-Tree Building	47
3.1.2	Decision-Tree Pruning	49
3.1.3	Pros and Cons of Decision-Tree Algorithms	50
3.1.4	The Inductive Bias of Decision-Tree Algorithms	52
3.1.5	Linear (Oblique) Decision Trees	53
3.2	Rule Induction Algorithms	54
3.2.1	The Pitfall of Attribute Interaction for Greedy Algorithms	55
3.3	Instance-Based Learning (Nearest Neighbor) Algorithms	58
	References	60
4	Data Preparation	65
4.1	Attribute Selection	65
4.1.1	The Motivation for Attribute Selection	66
4.1.2	Filter and Wrapper Approaches	66
4.1.3	Attribute Selection as a Particular Case of Attribute Weighting	71
4.2	Discretization of Continuous Attributes	71
4.2.1	An Overview of Discretization Methods	72
4.2.2	The Pros and Cons of Discretization	73
4.3	Attribute Construction	74
4.3.1	Categorizing Attribute Construction Methods	75
	References	76
5	Basic Concepts of Evolutionary Algorithms	79
5.1	An Overview of Evolutionary Algorithms (EAs)	79
5.1.1	Key Issues in the Design of EAs	81
5.2	Selection Methods	83
5.3	Genetic Algorithms (GA)	84
5.3.1	Individual (“Chromosome”) Representation	84
5.3.2	Crossover Operators	86
5.3.3	Mutation Operators	88
5.4	Genetic Programming (GP)	89
5.4.1	Function Set	90
5.4.2	Crossover Operators	92
5.4.3	Mutation Operators	95

5.4.4	The Standard GP Approach for Classification.....	99
5.4.5	Code Bloat	100
5.5	Niching.....	101
References	103
6	Genetic Algorithms for Rule Discovery	107
6.1	Individual Representation	107
6.1.1	Pittsburgh vs Michigan Approach.....	107
6.1.2	Encoding a Rule Antecedent.....	110
6.1.3	Encoding a Rule Consequent	116
6.2	Task-Specific Generalizing/Specializing Operators.....	119
6.2.1	Generalizing/Specializing Crossover	119
6.2.2	Generalizing/Specializing Mutation.....	122
6.2.3	Condition Removal/Insertion.....	123
6.2.4	Rule Insertion/Removal	124
6.2.5	Discussion.....	124
6.3	Task-Specific Population Initialization and Seeding.....	126
6.4	Task-Specific Rule-Selection Methods	127
6.5	Fitness Evaluation	129
References	134
7	Genetic Programming for Rule Discovery	139
7.1	The Problem of Closure in GP for Rule Discovery	139
7.2	Booleanizing All Terminals	140
7.2.1	Methods for Booleanizing Attributes.....	142
7.2.2	Examples of GP Systems Booleanizing Terminals.....	145
7.2.3	A Note on Semantic Constraints.....	145
7.3	Constrained-Syntax and Strongly-Typed GP	146
7.4	Grammar-Based GP for Rule Discovery	150
7.5	GP for Decision-Tree Building	153
7.5.1	Evolving Decision Trees with First-Order-Logic Conditions	154
7.5.2	Evolving Linear Decision Trees	155
7.5.3	Hybrid GP, Cellular Automaton, and Simulated Annealing	157
7.5.4	Using GP for Implementing Decision-Tree Nodes.....	157
7.6	On the Quality of Rules Discovered by GP.....	158
References	161

8	Evolutionary Algorithms for Clustering	165
8.1	Cluster Description-Based Individual Representation	165
8.1.1	Individual Representation	166
8.1.2	Genetic Operators	167
8.2	Centroid/Medoid-Based Individual Representation	169
8.2.1	Centroid-Based Individual Representation	169
8.2.2	Medoid-Based Individual Representation.....	171
8.3	Instance-Based Individual Representation	172
8.3.1	Hybrid Genetic Algorithm/K-Means.....	172
8.3.2	Graph-Based Individual Representation	173
8.4	Fitness Evaluation	174
8.4.1	Coping with Degenerate Partitions	176
8.5	EAs vs Conventional Clustering Techniques	177
	References	177
9	Evolutionary Algorithms for Data Preparation.....	179
9.1	EAs for Attribute Selection	179
9.1.1	Individual Encoding and Genetic Operators	180
9.1.2	Fitness Evaluation.....	182
9.1.3	Attribute Selection via Search for Partially Defined Instances.....	187
9.1.4	Joint Attribute Selection and Instance Selection.....	187
9.1.5	Attribute Selection for Clustering	189
9.1.6	Discussion.....	190
9.2	EAs for Attribute Weighting	191
9.2.1	Attribute Weighting	192
9.2.2	Towards Constructive Induction.....	193
9.3	Combining Attribute Selection and Attribute Weighting.....	194
9.4	GP for Attribute Construction	196
9.4.1	Constructing Combinations of Boolean Attributes	197
9.4.2	Discovering Specialized Functions	199
9.5	Combining Attribute Selection and Construction with a Hybrid GA/GP.....	200
	References	201
10	Evolutionary Algorithms for Discovering Fuzzy Rules.....	205
10.1	Basic Concepts of Fuzzy Sets	205
10.1.1	Fuzzy Sets	205
10.1.2	Operations on Fuzzy Sets.....	207

10.1.3 Membership Functions.....	209
10.2 Fuzzy Prediction Rules vs Crisp Prediction Rules	213
10.3 A Simple Taxonomy of EAs for Fuzzy-Rule Discovery.....	215
10.4 Using EAs for Generating Fuzzy Rules	215
10.4.1 Individual Encoding.....	216
10.4.2 Determining the Degree of Matching Between a Fuzzy Rule Antecedent and a Data Instance	220
10.4.3 Using Fuzzy Rules to Classify a Data Instance.....	221
10.4.4 Specifying the Shape and the Number of Membership Functions	222
10.5 Using EAs for Tuning Membership Functions.....	224
10.6 Using EAs for Both Generating Fuzzy Rules and Tuning Membership Functions.....	225
10.7 Fuzzy Fitness Evaluation	228
References	230
11 Scaling up Evolutionary Algorithms for Large Data Sets	233
11.1 Using Data Subsets in Fitness Evaluation	233
11.1.1 Random Training-Subset Selection.....	234
11.1.2 Adaptive Training-Subset Selection	235
11.2 An Overview of Parallel Processing.....	237
11.2.1 Basic Concepts of Parallel Processing	237
11.2.2 Load Balancing	239
11.2.3 Data Parallelism vs Control Parallelism.....	240
11.2.4 Speed up and Efficiency Measures	242
11.3 Parallel EAs for Data Mining.....	243
11.3.1 Exploiting Control Parallelism.....	246
11.3.2 Exploiting Data Parallelism	248
11.3.3 Exploiting Hybrid Control/Data-Parallelism	250
References	253
12 Conclusions and Research Directions.....	255
12.1 General Remarks on Data Mining with EAs	255
12.1.1 On Predictive Accuracy	255
12.1.2 On Knowledge Comprehensibility.....	256
12.1.3 On Computational Time	256
12.2 Research Directions.....	257
12.2.1 Developing EAs for Data Preparation.....	257

12.2.2 Developing Multi-objective EAs for Data Mining	258
12.2.3 Developing a GP System for Algorithm Induction.....	258
References	260
Index.....	263