Contents

1 Introduction 1
1.1 Rule Induction Algorithms 2
1.2 Evolutionary Computation 4
1.2.1 Genetic Programming 4
1.3 The Motivation for Automating the Design of Classification Algorithms 7
1.3.1 The Problem of the Selective Superiority of Classification Algorithms 7
1.3.2 Human Biases in Manually Designed Algorithms 10
1.3.3 A New Level of Automation in Data Mining 11
1.4 Overview of the Proposed Genetic Programming System 12
References 15

2 Data Mining 17
2.1 Introduction 17
2.2 The Classification Task of Data Mining 18
2.2.1 On Predictive Accuracy 19
2.2.2 On Overfitting and Underfitting 22
2.2.3 On the Comprehensibility of Discovered Knowledge 23
2.3 Decision Tree Induction 25
2.4 Rule Induction via the Sequential Covering Approach 27
2.4.1 Representation of the Candidate Rules 30
2.4.2 Search Mechanism 32
2.4.3 Rule Evaluation 34
2.4.4 Rule Pruning Methods 37
2.5 Meta-learning 39
2.5.1 Meta-learning for Classification Algorithm Selection 39
2.5.2 Stacked Generalization: Meta-learning via a Combination of Base Learners' Predictions 42
2.6 Summary 42
References 43

3 Evolutionary Algorithms 47
3.1 Introduction 47
3.2 An Overview of Evolutionary Algorithms 48
3.2.1 Individual Representation 48
3.2.2 Fitness Function 49
3.2.3 Individual Selection 50
3.2.4 Genetic Operators 50
3.3 Multiobjective Optimization 52
3.3.1 The Pareto Optimality Concept 53
3.3.2 Lexicographic Multiobjective Optimization 54
3.4 Genetic Programming Versus Genetic Algorithms: A Critical Perspective 55
3.5 Genetic Programming 59
3.5.1 Terminal and Function Sets and the Closure Property 62
3.5.2 Fitness Function: An Example Involving Regression 64
3.5.3 Selection and Genetic Operators 65
3.5.4 Approaches for Satisfying the Closure Property 68
3.5.5 Bloat 68
3.6 Grammar-Based Genetic Programming 70
3.6.1 Grammars 72
3.6.2 GGP with Solution-Encoding Individual 74
3.6.3 GGP with Production-Rule-Sequence-Encoding Individual . 77
3.7 Summary 80
References 80

4 Genetic Programming for Classification and Algorithm Design 85
4.1 Introduction 85
4.2 Classification Models Versus Classification Algorithms 86
4.3 Genetic Programming for Evolving Classification Models 88
4.3.1 Evolving Classification Functions or Classification Rules ... 89
4.3.2 Evolving Decision Trees 91
4.4 Genetic Programming for Evolving Components of Rule Induction Algorithms 92
4.5 Genetic Programming for Evolving Classification Systems 95
4.6 Evolving the Design of Optimization Algorithms 97
4.6.1 Optimization Versus Classification 97
4.6.2 On Meta-heuristics and Hyper-heuristics 100
4.6.3 Evolving the Core Heuristic of Optimization Algorithms ... 101
4.6.4 Evolving an Evolutionary Algorithm for Optimization 104
4.7 Summary 105
References 106

5 Automating the Design of Rule Induction Algorithms 109
5.1 Introduction 109
5.2 The Grammar: Specifying the Building Blocks of Rule Induction Algorithms 111
5.2.1 The New Rule Induction Algorithmic Components in the Grammar 116
5.3 Individual Representation 117
5.4 Population Initialization 118
5.5 Individual Evaluation 121
5.5.1 From a Derivation Tree to Java Code 124
5.5.2 Single-Objective Fitness 126
5.5.3 Multiobjective Fitness 129
5.6 Crossover and Mutation Operations 131
5.7 Summary 133
References 133

6 Computational Results on the Automatic Design of Full Rule Induction Algorithms 137
6.1 Introduction 137
6.2 Evolving Rule Induction Algorithms Robust Across Different Application Domains 138
6.2.1 Investigating the GGP System's Sensitivity to Parameters .. 139
6.2.2 Comparing GGP-Designed Rule Induction Algorithms with Human-Designed Rule Induction Algorithms 142
6.2.3 To What Extent Are GGP-RIs Different from Manually Designed Rule Induction Algorithms? 144
6.2.4 Meta-training Set Variations 148
6.2.5 GGP System's Grammar Variations 151
6.2.6 GGP Versus Grammar-Based Hill-Climbing Search 153
6.2.7 MOGGP: A Multiobjective Version of the Proposed GGP .. 156
6.2.8 A Note on the GGP System's Execution Time 160
6.3 Evolving Rule Induction Algorithms Tailored to the Target Application Domain 161
6.3.1 Experiments with Public UCI Datasets 162
6.3.2 GGP-RIs Versus GHC-RIs 166
6.3.3 Experiments with Bioinformatics Datasets 167
6.3.4 A Note on the GGP System's Execution Time 172
6.4 Summary 173
References 174

7 Directions for Future Research on the Automatic Design of Data Mining Algorithms 177
7.1 Potential Improvements to the Current GGP System 178
7.1.1 Improving the Grammar 178
7.1.2 Modifying the GGP System's Fitness Function 179

7.2 Designing Rule Induction Algorithms Tailored to a Type of Dataset 180
7.3 Investigating Other Types of Search Methods for Automated Algorithm Design 181
7.4 Automatically Designing Other Types of Classification Algorithms . 182
7.5 Automatically Designing Other Types of Data Mining Algorithms . 183
References 184

Index 185