Contents

1 Introduction 1
1.1 Data Mining 4
1.1.1 Association Rule Mining (ARM) 4
1.1.2 Incremental Mining 5
1.1.3 Distributed Data Mining 6
1.1.4 Sequential Mining 6
1.1.5 Clustering 6
1.1.6 Classification 8
1.1.7 Characterization 8
1.1.8 Discrimination 9
1.1.9 Deviation Mining 9
1.1.10 Evolution Mining 9
1.1.11 Prediction 10
1.1.12 Web Mining 10
1.1.13 Text Mining 11
1.1.14 Data Warehouses 11
1.2 Soft Computing 13
1.2.1 Importance of Soft Computing 13
1.2.2 Genetic Algorithms 13
1.2.3 Neural Networks 14
1.2.4 Support Vector Machines 14
1.2.5 Fuzzy Logic 15
1.2.6 Rough Sets 16
1.3 Data Mining Applications 16
References 17

2 Self Adaptive Genetic Algorithms 19
2.1 Introduction 19
2.2 Related Work 20
2.3 Overview 22
2.4 Algorithm 23

2.4.1 Problem Definition . . . 23
2.4.2 Pseudocode . . . 23
2.5 Mathematical Analysis . . . 25
2.5.1 Convergence Analysis . . . 30
2.6 Experiments . . . 32
2.7 Performance Analysis . . . 40
2.8 A Heuristic Template Based Adaptive Genetic Algorithms . . . 42
2.8.1 Problem Definition . . . 42
2.9 Example . . . 42
2.10 Performance Analysis of HTAGA . . . 44
2.11 Summary . . . 48
References . . . 49

3 Characteristic Amplification Based Genetic Algorithms . . . 51
3.1 Introduction . . . 51
3.2 Formalizations . . . 52
3.3 Design Issues . . . 54
3.4 Algorithm . . . 55
3.5 Results and Performance Analysis . . . 58
3.6 Summary . . . 61
References . . . 61

4 Dynamic Association Rule Mining Using Genetic Algorithms . . . 63
4.1 Introduction . . . 63
4.1.1 Inter Transaction Association Rule Mining . . . 64
4.1.2 Genetic Algorithms . . . 65
4.2 Related Work . . . 66
4.3 Algorithms . . . 67
4.4 Example . . . 69
4.5 Performance Analysis . . . 74
4.5.1 Experiments on Real Data . . . 78
4.6 Summary . . . 79
References . . . 79

5 Evolutionary Approach for XML Data Mining . . . 81
5.1 Semantic Search over XML Corpus . . . 82
5.2 The Existing Problem . . . 83
5.2.1 Motivation . . . 84
5.3 XML Data Model and Query Semantics . . . 85
5.4 Genetic Learning of Tags . . . 86
5.5 Search Algorithm . . . 89
5.5.1 Identification Scheme . . . 89

5.5.2 Relationship Strength . . . 90
5.5.3 Semantic Interconnection . . . 91
5.6 Performance Studies . . . 93
5.7 Selective Dissemination of XML Documents . . . 99
5.8 Genetic Learning of User Interests . . . 101
5.9 User Model Construction . . . 102
5.9.1 SVM for User Model Construction . . . 103
5.10 Selective Dissemination . . . 103
5.11 Performance Analysis . . . 105
5.12 Categorization Using SVMs . . . 108
5.12.1 XML Topic Categorization . . . 108
5.12.2 Feature Set Construction . . . 109
5.13 SVM for Topic Categorization . . . 111
5.14 Experimental Studies . . . 113
5.15 Summary . . . 116
References . . . 117

6 Soft Computing Based CBIR System . . . 119
6.1 Introduction . . . 119
6.2 Related Work . . . 120
6.3 Model . . . 121
6.3.1 Pre-processing . . . 122
6.3.2 Feature Extraction . . . 122
6.3.3 Feature Clustering . . . 126
6.3.4 Classification . . . 126
6.4 The STIRF System . . . 128
6.5 Performance Analysis . . . 129
6.6 Summary . . . 136
References . . . 136

7 Fuzzy Based Neuro - Genetic Algorithm for Stock Market Prediction . . . 139
7.1 Introduction . . . 139
7.2 Related Work . . . 140
7.3 Model . . . 141
7.4 Algorithm . . . 146
7.4.1 Algorithm FEASOM . . . 146
7.4.2 Modified Kohonen Algorithm . . . 146
7.4.3 The Genetic Algorithm . . . 148
7.4.4 Fuzzy Inference System . . . 149
7.4.5 Backpropagation Algorithm . . . 149
7.4.6 Complexity . . . 149
7.5 Example . . . 150
7.6 Implementation . . . 152
7.7 Performance Analysis . . . 154

7.8 Summary . . . 165
References . . . 165

8 Data Mining Based Query Processing Using Rough Sets and GAs . . . 167
8.1 Introduction . . . 167
8.2 Problem Definition . . . 169
8.3 Architecture . . . 170
8.3.1 Rough Sets . . . 171
8.3.2 Information Streaks . . . 174
8.4 Modeling of Continuous-Type Data . . . 175
8.5 Genetic Algorithms and Query Languages . . . 180
8.5.1 Associations . . . 181
8.5.2 Concept Hierarchies . . . 182
8.5.3 Dealing with Rapidly Changing Data . . . 185
8.6 Experimental Results . . . 186
8.7 Adaptive Data Mining Using Hybrid Model of Rough Sets and Two-Phase GAs . . . 189
8.8 Mathematical Model of Attributes (MMA) . . . 190
8.9 Two Phase Genetic Algorithms . . . 191
8.10 Summary . . . 194
References . . . 194

9 Hashing the Web for Better Reorganization . . . 197
9.1 Introduction . . . 197
9.1.1 Frequent Items and Association Rules . . . 198
9.2 Related Work . . . 200
9.3 Web Usage Mining and Web Reorganization Model . . . 200
9.4 Problem Definition . . . 202
9.5 Algorithms . . . 202
9.5.1 Classification of Pages . . . 206
9.6 Pre-processing . . . 206
9.7 Example . . . 208
9.8 Performance Analysis . . . 210
9.9 Summary . . . 214
References . . . 214

10 Algorithms for Web Personalization . . . 217
10.1 Introduction . . . 217
10.2 Overview . . . 219
10.3 Data Structures . . . 219
10.4 Algorithm . . . 221
10.5 Performance Analysis . . . 223
10.6 Summary . . . 229
References . . . 229

11 Classifying Clustered Webpages for Effective Personalization ... 231
11.1 Introduction ... 231
11.2 Related Work ... 232
11.3 Proposed System ... 233
11.4 Example ... 237
11.5 Algorithm II: Naïve Bayesian Probabilistic Model ... 239
11.6 Performance Analysis ... 241
11.7 Summary ... 246
References ... 247

12 Mining Top - k Ranked Webpages Using SA and GA ... 249
12.1 Introduction ... 249
12.2 Algorithm *TkRSAGA* ... 252
12.3 Performance Analysis ... 253
12.4 Summary ... 258
References ... 258

13 A Semantic Approach for Mining Biological Databases ... 259
13.1 Introduction ... 259
13.2 Understanding the Nature of Biological Data ... 260
13.3 Related Work ... 262
13.4 Problem Definition ... 263
13.5 Identifying Indexing Technique ... 263
13.6 LSI Model ... 265
13.7 Search Optimization Using GAs ... 266
13.8 Proposed Algorithm ... 267
13.9 Performance Analysis ... 268
13.10 Summary ... 277
References ... 277

14 Probabilistic Approach for DNA Compression ... 279
14.1 Introduction ... 279
14.2 Probability Model ... 281
14.3 Algorithm ... 284
14.4 Optimization of P' ... 285
14.5 An Example ... 286
14.6 Performance Analysis ... 287
14.7 Summary ... 288
References ... 288

15 Non-repetitive DNA Compression Using Memoization ... 291
15.1 Introduction ... 291
15.2 Related Work ... 293
15.3 Algorithm ... 294
15.4 Experimental Results ... 298
15.5 Summary ... 300
References ... 300

16 Exploring Structurally Similar Protein Sequence Motifs ... 303
16.1 Introduction ... 303
16.2 Related Work ... 305
16.3 Motifs in Protein Sequences ... 305
16.4 Algorithm ... 307
16.5 Experimental Setup ... 308
16.6 Experimental Results ... 310
16.7 Summary ... 317
References ... 317

17 Matching Techniques in Genomic Sequences for Motif Searching ... 319
17.1 Overview ... 319
17.2 Related Work ... 320
17.3 Introduction ... 321
17.4 Alternative Storage and Retrieval Technique ... 323
17.5 Experimental Setup and Results ... 327
17.6 Summary ... 329
References ... 330

18 Merge Based Genetic Algorithm for Motif Discovery ... 331
18.1 Introduction ... 331
18.2 Related Work ... 334
18.3 Algorithm ... 334
18.4 Experimental Setup ... 337
18.5 Performance Analysis ... 339
18.6 Summary ... 340
References ... 340