

Understanding and Using Rough Set Based Feature Selection: Concepts, Techniques and Applications

Muhammad Summair Raza ·
Usman Qamar

Understanding and Using Rough Set Based Feature Selection: Concepts, Techniques and Applications

Second Edition

 Springer

Muhammad Summair Raza
Department of Computer and Software
Engineering, College of Electrical
and Mechanical Engineering
National University of Sciences
and Technology (NUST)
Islamabad, Pakistan

Usman Qamar
Department of Computer and Software
Engineering, College of Electrical
and Mechanical Engineering
National University of Sciences
and Technology (NUST)
Islamabad, Pakistan

ISBN 978-981-32-9165-2

ISBN 978-981-32-9166-9 (eBook)

<https://doi.org/10.1007/978-981-32-9166-9>

1st edition: © The Editor(s) (if applicable) and The Author(s) 2017

2nd edition: © Springer Nature Singapore Pte Ltd. 2019

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, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd. The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

This book is dedicated to the memory of Prof. Zdzisław Pawlak (1926–2006), the father of Rough Set Theory. He was the founder of the Polish school of Artificial Intelligence and one of the pioneers in Computer Engineering and Computer Science.

Preface

Rough set theory, proposed in 1982 by Zdzislaw Pawlak, is in constant development. It is concerned with the classification and analysis of imprecise or uncertain information and knowledge. It has become a prominent tool for data analysis. This book provides a comprehensive introduction to rough set-based feature selection. It enables the reader to systematically study all topics of rough set theory (RST) including preliminaries, advance concepts, and feature selection using RST. This book is supplemented with RST-based API library that can be used to implement several RST concepts and RST-based feature selection algorithms.

The book is intended to provide an important reference material for students, researchers and developers working in the areas of feature selection, knowledge discovery and reasoning with uncertainty, especially for those who are working in RST and granular computing. The primary audience of this book is the research community using rough set theory (RST) to perform feature selection (FS) on large-scale datasets in various domains. However, any community interested in feature selection such as medical, banking, finance can also benefit from the book.

The second edition of the book now also covers dominance-based rough set approach and fuzzy rough sets. Dominance-based rough set approach (DRSA) is an extension to the conventional rough set approach which supports the preference order using dominance principle. Fuzzy rough sets are fuzzy generalization of rough sets. API library of dominance-based rough set approach is also provided with the second edition of the book.

Islamabad, Pakistan

Muhammad Summair Raza
Usman Qamar

Contents

1	Introduction to Feature Selection	1
1.1	Feature	1
1.1.1	Numerical	2
1.1.2	Categorical Attributes	3
1.2	Feature Selection	3
1.2.1	Supervised Feature Selection	4
1.2.2	Unsupervised Feature Selection	6
1.3	Feature Selection Methods	8
1.3.1	Filter Methods	8
1.3.2	Wrapper Methods	10
1.3.3	Embedded Methods	10
1.4	Objective of Feature Selection	11
1.5	Feature Selection Criteria	13
1.5.1	Information Gain	14
1.5.2	Distance	14
1.5.3	Dependency	14
1.5.4	Consistency	15
1.5.5	Classification Accuracy	15
1.6	Feature Generation Schemes	15
1.6.1	Forward Feature Generation	15
1.6.2	Backward Feature Generation	16
1.6.3	Random Feature Generation	17
1.7	Related Concepts	18
1.7.1	Search Organization	18
1.7.2	Generation of a Feature Selection Algorithm	18
1.7.3	Feature Relevance	19
1.7.4	Feature Redundancy	19

1.7.5	Applications of Feature Selection	20
1.7.6	Feature Selection: Issues	21
1.8	Summary	22
	References	23
2	Background	27
2.1	Curse of Dimensionality	27
2.2	Transformation-Based Reduction	28
2.2.1	Linear Methods	29
2.2.2	Non-linear Methods	32
2.3	Selection-Based Reduction	35
2.3.1	Feature Selection in Supervised Learning	36
2.3.2	Filter Techniques	36
2.3.3	Wrapper Techniques	39
2.3.4	Feature Selection in Unsupervised Learning	40
2.4	Correlation-Based Feature Selection	42
2.4.1	Correlation-Based Measures	43
2.4.2	Efficient Feature Selection Based on Correlation Measure (ECMBF)	45
2.5	Mutual Information-Based Feature Selection	46
2.5.1	A Mutual Information-Based Feature Selection Method (MIFS-ND)	47
2.5.2	Multi-objective Artificial Bee Colony (MOABC) Approach	48
2.6	Summary	49
	References	49
3	Rough Set Theory	53
3.1	Classical Set Theory	53
3.1.1	Sets	53
3.1.2	SubSets	54
3.1.3	Power Sets	54
3.1.4	Operators	55
3.1.5	Mathematical Symbols for Set Theory	56
3.2	Knowledge Representation and Vagueness	56
3.3	Rough Set Theory (RST)	57
3.3.1	Information Systems	58
3.3.2	Decision Systems	58
3.3.3	Indiscernibility	58
3.3.4	Approximations	59
3.3.5	Positive Region	61
3.3.6	Discernibility Matrix	62
3.3.7	Discernibility Function	63
3.3.8	Decision-Relative Discernibility Matrix	64

3.3.9	Dependency	66
3.3.10	Reducts and Core	68
3.4	Discretization Process	70
3.5	Miscellaneous Concepts	72
3.6	Applications of RST	73
3.7	Summary	76
	References	76
4	Advanced Concepts in Rough Set Theory	81
4.1	Fuzzy Set Theory	81
4.1.1	Fuzzy Set	81
4.1.2	Fuzzy Sets and Partial Truth	82
4.1.3	Membership Function	83
4.1.4	Fuzzy Operators	84
4.1.5	Fuzzy Set Representation	86
4.1.6	Fuzzy Rules	87
4.2	Fuzzy Rough Set Hybridization	89
4.2.1	Supervised Learning and Information Retrieval	89
4.2.2	Feature Selection	89
4.2.3	Rough Fuzzy Set	91
4.2.4	Fuzzy Rough Set	92
4.3	Dependency Classes	92
4.3.1	Incremental Dependency Classes (IDC)	93
4.3.2	Direct Dependency Classes (DDC)	98
4.4	Redefined Approximations	101
4.4.1	Redefined Lower Approximation	101
4.4.2	Redefined Upper Approximation	104
4.5	Summary	107
	References	107
5	Rough Set Theory Based Feature Selection Techniques	109
5.1	Quick Reduct	109
5.2	Hybrid Feature Selection Algorithm Based on Particle Swarm Optimization (PSO)	112
5.3	Genetic Algorithm	113
5.4	Incremental Feature Selection Algorithm (IFSA)	116
5.5	Feature Selection Method Using Fish Swarm Algorithm (FSA)	117
5.5.1	Representation of Position	118
5.5.2	Distance and Center of Fish	119
5.5.3	Position Update Strategies	119
5.5.4	Fitness Function	119
5.5.5	Halting Condition	119

5.6	Feature Selection Method Based on Quick Reduct and Improved Harmony Search Algorithm (RS-IHS-QR)	120
5.7	A Hybrid Feature Selection Approach Based on Heuristic and Exhaustive Algorithms Using Rough Set Theory (FSHEA)	120
5.7.1	Feature Selection Preprocessor	120
5.7.2	Using Relative Dependency Algorithm to Optimize the Selected Features	123
5.8	A Rough Set Based Feature Selection Approach Using Random Feature Vectors	127
5.9	Heuristic-Based Dependency Calculation Technique	130
5.10	Parallel Dependency Calculation Method for Feature Selection.	131
5.11	Summary	131
	References	134
6	Unsupervised Feature Selection Using RST	135
6.1	Unsupervised Quick Reduct Algorithm (USQR)	135
6.2	Unsupervised Relative Reduct Algorithm	139
6.3	Unsupervised Fuzzy-Rough Feature Selection	141
6.4	Unsupervised PSO Based Relative Reduct (US-PSO-RR)	142
6.5	Unsupervised PSO Based Quick Reduct (US-PSO-QR)	145
6.6	Summary	147
	References	147
7	Critical Analysis of Feature Selection Algorithms	149
7.1	Pros and Cons of Feature Selection Techniques	149
7.1.1	Filter Methods	149
7.1.2	Wrapper Methods	150
7.1.3	Embedded Methods	150
7.2	Comparison Framework.	151
7.2.1	Percentage Decrease in Execution Time	151
7.2.2	Memory Usage	151
7.3	Critical Analysis of Various Feature Selection Algorithms	152
7.3.1	Quick Reduct	152
7.3.2	Rough Set Based Genetic Algorithm	153
7.3.3	PSO-QR	154
7.3.4	Incremental Feature Selection Algorithm (IFSA).	155
7.3.5	AFSA	155
7.3.6	Feature Selection Using Exhaustive and Heuristic Approach	156
7.3.7	Feature Selection Using Random Feature Vectors	157
7.4	Summary	157
	References	157

8	Dominance-Based Rough Set Approach	159
8.1	Introduction	159
8.2	Dominance-Based Rough Set Approach	160
8.2.1	Decision Table	160
8.2.2	Dominance	161
8.2.3	Decision Classes and Class Unions	162
8.2.4	Lower Approximations	163
8.2.5	Upper Approximations	164
8.3	Some DRSA-Based Approaches	166
8.4	Summary	174
	References	175
9	Fuzzy Rough Sets	179
9.1	Fuzzy Rough Set Model	179
9.1.1	Fuzzy Approximations	179
9.1.2	Fuzzy Positive Region	180
9.2	Fuzzy Rough Set Based Approaches	181
9.3	Summary	187
	References	187
10	Introduction to Classical Rough Set Based APIs Library	189
10.1	A Simple Tutorial	189
10.1.1	Variable Declaration	189
10.1.2	Array Declaration	190
10.1.3	Comments	190
10.1.4	If-Else Statement	191
10.1.5	Loops	191
10.1.6	Functions	192
10.1.7	LBound and UBound Functions	192
10.2	How to Import the Source Code	192
10.3	Calculating Dependency Using Positive Region	200
10.3.1	Main Function	200
10.3.2	CalculateDRR Function	202
10.3.3	SetDClasses Method	203
10.3.4	FindIndex Function	204
10.3.5	ClrTCC Function	205
10.3.6	AlreadyExists Method	206
10.3.7	InsertObject Method	207
10.3.8	MatchCClasses Function	207
10.3.9	PosReg Function	208
10.4	Calculating Dependency Using Incremental Dependency Classes	209
10.4.1	Main Function	209
10.4.2	CalculateDID Function	209

10.4.3	Insert Method	212
10.4.4	MatchChrom Method	213
10.4.5	MatchDClass Method	213
10.5	Lower Approximation Using Conventional Method	214
10.5.1	Main Method	214
10.5.2	CalculateLAObjects Method	215
10.5.3	FindLAO Method	217
10.5.4	SetDConcept Method	218
10.6	Lower Approximation Using Redefined Preliminaries	218
10.7	Upper Approximation Using Conventional Method	221
10.8	Upper Approximation Using Redefined Preliminaries	221
10.9	Quick Reduct Algorithm	223
10.9.1	Miscellaneous Methods	225
10.9.2	Restore Method	226
10.9.3	C_R Method	226
10.10	Summary	227
11	Dominance Based Rough Set APIs Library	229
11.1	Lower Approximations	229
11.1.1	Function: Find_PL_L_t ()	230
11.1.2	Function: Get_Cl_LE_t	231
11.1.3	Function: DP_N_X	231
11.1.4	Function: Find_P_L_G_T	232
11.2	Upper Approximations	234
11.3	Summary	236

About the Authors

Muhammad Summair Raza has Ph.D. specialization in software engineering from the National University of Sciences and Technology (NUST), Pakistan. He completed his MS from International Islamic University, Pakistan, in 2009. He is also associated with the Virtual University of Pakistan as assistant professor. He has published various papers in international-level journals and conferences with a focus on rough set theory. His research interests include feature selection, rough set theory, trend analysis, software architecture, software design and non-functional requirements.

Dr. Usman Qamar is currently a tenured associate professor at National University of Sciences and Technology (NUST) having over 15 years of experience in data engineering and decision sciences both in academia and industry having spent nearly 10 years in the UK. He has a Masters in Computer Systems Design from University of Manchester Institute of Science and Technology (UMIST), UK. His M.Phil. in Computer Systems was a joint degree between UMIST and University of Manchester which focused on feature selection in big data. In 2008/09, he was awarded Ph.D. from University of Manchester, UK. His Ph.D. specialization is in Data Engineering, Knowledge Discovery and Decision Science. His Post Ph.D. work at University of Manchester, involved various research projects including hybrid mechanisms for statistical disclosure (feature selection merged with outlier analysis) for Office of National Statistics (ONS), London, UK, churn prediction for Vodafone UK and customer profile analysis for shopping with the University of Ghent, Belgium. He has also done a post graduation in Medical and Health Research, from University of Oxford, UK, where he worked on evidence-based health care, thematic qualitative data analysis and healthcare innovation and technology. He is director of Knowledge and Data Science Research Centre, a Centre of Excellence at NUST, Pakistan and principal investigator of Digital Pakistan Lab, which is part of National Centre for Big Data and Cloud Computing. He has authored over 150 peer reviewed publications which includes 2 books published by Springer & Co. He has successfully supervised 5 Ph.D. students and over 70 master students. Dr. Usman has been able to acquire nearly PKR 100 million in research grants. He has received multiple research awards, including Best Researcher of

Pakistan 2015/16 by Higher Education Commission (HEC), Pakistan as well as gold in Research and Development category by Pakistan Software Houses Association (P@SHA) ICT Awards 2013 and 2017 and Silver award in APICTA (Asia Pacific ICT Alliance Awards) 2013 in category of R&D hosted by Hong Kong. He is also recipient of the prestigious Charles Wallace Fellowship 2016/17 as well as British Council Fellowship 2018, visiting research fellow at Centre of Decision Research, University of Leeds, UK and visiting senior lecturer at Manchester Metropolitan University, UK. Finally, he has the honour of being the finalist of the British Council's Professional Achievement Award 2016/17.