

**FUZZY ROUGH SET ASSISTED TECHNIQUES FOR  
DATA REDUCTION TO ENHANCE PREDICTION  
PERFORMANCES**



*The thesis submitted in partial fulfilment*

*for the Award of Degree*

*DOCTOR OF PHILOSOPHY*

*by*

PANKHURI JAIN

DEPARTMENT OF MATHEMATICAL SCIENCES

INDIAN INSTITUTE OF TECHNOLOGY

(BANARAS HINDU UNIVERSITY)

VARANASI -221005

**Roll No: 18121023**

**June 2022**

## CERTIFICATE

It is certified that the work contained in this thesis entitled " Fuzzy rough set assisted techniques for data reduction to enhance prediction performances " by "Pankhuri Jain" has been carried out under my supervision and that it has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements of Comprehensive, Candidacy and SOTA.

T. Som.  
30.6.2022

Prof. Tanmoy Som

Supervisor

Professor

Head of the Department

Department of Mathematical Sciences

Indian Institute of Technology

(Banaras Hindu University)

Varanasi-221005

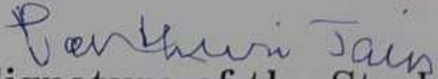
## DECLARATION BY THE CANDIDATE

---

I, **Pankhuri Jain** certify that the work embodied in this thesis is my own bonafide work and carried out by me under the supervision of Prof. Tanmoy Som from July-2018 to June-2022, at the Department of Mathematical Sciences, Indian Institute of Technology (BHU) Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully lifted up any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports dissertations, theses, etc., or available at websites and included them in this thesis and cited as my own work.

Date : 30/06/2022

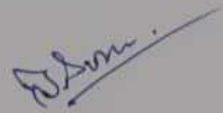
Place : Varanasi

  
Signature of the Student

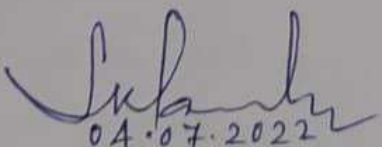
(PANKHURI JAIN)

## CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the student is correct to the best of my/our knowledge.

  
Signature of Supervisor

(Prof. Tanmoy Som)

  
04.07.2022  
Signature of Head of Department

## COPYRIGHT TRANSFER CERTIFICATE

---

Title of the Thesis: FUZZY ROUGH SET ASSISTED TECHNIQUES FOR DATA REDUCTION TO ENHANCE PREDICTION PERFORMANCES

Name of the Student: PANKHURI JAIN

### Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University) Varanasi, all rights under copyright that may exist in and for the above thesis submitted for the award of the DOCTOR OF PHILOSOPHY.

Date : 30/06/2022

Place : Varanasi

*Pankhuri Jain*  
(PANKHURI JAIN)

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

## ACKNOWLEDGEMENTS

---

I express my deep sense of gratitude to my external RPEC member **Prof. R. Srivastava**, **Prof. S. Mukhopadhyay**, Department of Mathematical Sciences, Indian Institute of Technology (BHU) and **Dr. T. Datta**, Department of Computer Engineering, Indian Institute of Technology (BHU) for their valuable suggestions, guidance, appreciation and encouragement.

I owe a lot to all my teachers who taught me Computers.

*Pankhuri Jain*  
(Pankhuri Jain)

# Contents

Certificate	ii
Declaration by the Candidate	iii
Copyright Transfer Certificate	iv
Acknowledgements	v
Contents	vi
List of Figures	xi
List of Tables	xiii
Preface	xvii
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Feature Selection . . . . .	3
1.2.1 Benefits and Application of Feature Selection . . . . .	8
1.3 Thesis Structure . . . . .	9
<b>2 Background</b>	<b>13</b>
2.1 Rough Set Theory . . . . .	13
2.2 Fuzzy Set Theory . . . . .	15
2.3 Fuzzy Rough Set Theory . . . . .	16
2.4 Fuzzy Rough Set based Feature Selection . . . . .	17
2.4.1 Related Work . . . . .	19
2.4.2 Limitation of Fuzzy Set . . . . .	20

2.5	Intuitionistic Fuzzy Set . . . . .	20
2.5.1	Intuitionistic Fuzzy Rough Set Theory . . . . .	21
2.5.2	Related Work . . . . .	22
2.6	Summary . . . . .	23
<b>3</b>	<b>Feature Selection Models and its Application</b>	<b>25</b>
3.1	Divergence based Intuitionistic Fuzzy Rough Set Model . . . . .	26
3.1.1	Feature selection using IF rough set model based on divergence measure . . . . .	35
3.1.2	Experimentation . . . . .	38
3.1.3	Application to tuberculosis treatment . . . . .	39
3.1.3.1	Result . . . . .	42
3.2	k-mean based Intuitionistic Fuzzy Rough Set Model . . . . .	49
3.2.1	Feature Selection based on Intuitionistic Fuzzy Rough Set Model based on k-means . . . . .	50
3.2.2	Application for Enhancing Prediction of Aptamer-protein Interacting Pairs . . . . .	51
3.2.2.1	Result . . . . .	52
3.3	Summary . . . . .	55
<b>4</b>	<b>A Fitting Model for Feature Selection</b>	<b>57</b>
4.1	Fitting Model based on Intuitionistic Fuzzy Rough Set . . . . .	58
4.2	Feature Selection based on Fitting Model . . . . .	61
4.3	Experimentation . . . . .	64
4.4	Summary . . . . .	68
<b>5</b>	<b>Feature Selection Model for Incomplete Data</b>	<b>69</b>
5.1	Fuzzy Rough Model for Feature Selection and Missing Value Imputation	70
5.1.1	Feature Grouping . . . . .	71
5.1.2	Missing Value Imputation . . . . .	71
5.1.3	Search Heuristic for Finding Reduct . . . . .	75
5.2	Experimentation . . . . .	80
5.2.1	Results . . . . .	81
5.2.1.1	By employing Parameter variation . . . . .	82
	By varying the level of induced missing Values . . . . .	82
	By varying the value of $k$ (number of nearest neighbour for missing value imputation) . . . . .	84
	By varying the percentage of correlated features employed for obtaining reduced dataset using proposed approach . . . . .	86
	By varying the percentage of missing entries required in an instance to ignore the instance . . . . .	88

---

5.2.1.2	Comparison with Existing Missing Value Imputation method . . . . .	89
5.2.1.3	Comparison with Other Feature Selection Algorithms . . . . .	92
5.3	Summary . . . . .	93
<b>6</b>	<b>Bireduct Model and its Application</b>	<b>95</b>
6.0.1	Bireduct formulation . . . . .	96
6.1	Intuitionistic Fuzzy Bireducts for Data Reduction . . . . .	97
6.1.1	Intuitionistic Fuzzy Feature Selection . . . . .	98
6.1.2	Intuitionistic Fuzzy Instance Selection . . . . .	99
6.1.2.1	Method I . . . . .	99
6.1.2.2	Method II . . . . .	100
6.1.3	Simultaneous Intuitionistic Fuzzy Instance and Feature Selection . . . . .	102
6.1.4	Heuristic Search Strategy for IF Bireducts . . . . .	105
6.2	Experimentation . . . . .	109
6.2.1	Results . . . . .	110
6.2.1.1	Using Parameter Variation . . . . .	111
6.2.1.2	Using Variants of Instance Selection . . . . .	116
6.2.1.3	Comparisons with Other FS Algorithms . . . . .	118
6.2.1.4	Comparison with Instance Selection and Feature Selection + Instance Selection Approaches . . . . .	121
6.2.1.5	Comparison with existing Bireduct approach . . . . .	124
6.3	Application to Cancer Treatment . . . . .	125
6.3.0.1	Results . . . . .	129
6.3.0.2	Comparison with Unreduced Dataset . . . . .	131
6.3.0.3	Comparison with Existing Approaches . . . . .	131
6.4	Summary . . . . .	132
<b>7</b>	<b>Unsupervised Feature Selection</b>	<b>135</b>
7.1	Feature Selection based on Fuzzy Rough Set in Unsupervised Domain . . . . .	137
7.1.1	Feature subset quality evaluation . . . . .	138
7.1.1.1	Dependency measure . . . . .	138
7.1.2	Search strategy . . . . .	141
7.1.2.1	Reproduction 1 . . . . .	141
7.1.2.2	Reproduction 2 . . . . .	142
7.2	Experimentation . . . . .	144
7.2.1	Results . . . . .	145
7.2.1.1	Using variants of proposed approach . . . . .	146
7.2.1.2	Comparison with state of art dependency based approach . . . . .	147
7.2.1.3	Comparison with state of art non dependency based approach . . . . .	150



---

7.2.1.4	Comparison with supervised approach . . . . .	152
7.3	Summary . . . . .	153
<b>8</b>	<b>Conclusion</b>	<b>159</b>
8.1	Feature Selection . . . . .	160
8.2	Future Work . . . . .	161
<b>A</b>	<b>Data Validation Techniques</b>	<b>163</b>
<b>B</b>	<b>Performance Evaluation Metrics</b>	<b>165</b>
<b>C</b>	<b>Statistical Testing</b>	<b>167</b>
	<b>Bibliography</b>	<b>169</b>

# List of Figures

1.1	The knowledge discovery process . . . . .	2
1.2	Feature Selection . . . . .	4
1.3	Filter approach to Feature Selection . . . . .	6
1.4	Wrapper approach to Feature Selection . . . . .	6
1.5	Various applications of Feature Selection . . . . .	10
3.1	Flowchart of proposed methodology for prediction of anti tubercular peptides . . . . .	42
3.2	AUC of eight machine learning algorithms for the dataset . . . . .	47
3.3	AUC of eight machine learning algorithms for the reduced dataset . . . . .	47
3.4	AUC of seven machine learning algorithms for the reduced dataset . . . . .	54
4.1	Variation of classification accuracy and reduct size with epsilon by proposed method . . . . .	67
5.1	The flowchart of proposed model . . . . .	80
5.2	Graphical visualization showing comparison of the classification accuracy on varying percentage of missing values . . . . .	84
5.3	Graphical visualization showing comparison of the classification accuracy by varying the value of $k$ (number of nearest neighbour for missing value imputation) . . . . .	86
5.4	Graphical visualization showing comparison of the classification accuracy by varying percentage of correlated features employed for obtaining reduced dataset using proposed approach . . . . .	86
5.5	Graphical visualization showing comparison of the classification accuracy by varying percentage of missing entries required in an instance to ignore the instance . . . . .	89
5.6	Graphical visualization showing comparison of the classification accuracy with other missing value imputation algorithms . . . . .	91
5.7	Graphical visualization showing comparison of the classification accuracy with other feature selection algorithms . . . . .	91
6.1	The flowchart of IFBRPSO-1 . . . . .	100
6.2	The flowchart of IFBRPSO-2 . . . . .	102

---

6.3	The flowchart of entire methodology to calculate $\epsilon$ -bireduct i.e for obtaining a reduced representation of the dataset . . . . .	109
6.4	Graphical visualization showing the variation of classification accuracy with dataset coverage ( $\epsilon$ ) and noise parameter $k$ . . . . .	115
6.5	Graphical visualization showing the classification accuracy with variants of IS . . . . .	116
6.6	Graphical visualization showing comparison of the classification accuracy with state of the art feature selection approach . . . . .	119
6.7	Graphical visualization showing comparison of the classification accuracy with other IS-FS combination . . . . .	123
6.8	Graphical visualization showing comparison of the classification accuracy with Bireduct approach . . . . .	125
6.9	The flowchart of proposed cancer treatment model . . . . .	130
6.10	Graphical visualization showing comparison of the classification accuracy with Bireduct approach . . . . .	132
7.1	The flowchart of entire methodology for obtaining a reduced representation of the dataset . . . . .	144

# List of Tables

2.1	Example Dataset . . . . .	19
3.1	Example dataset . . . . .	37
3.2	Divergence matrix obtained from example dataset for attribute $a_1$ . . . . .	37
3.3	Lower approximation obtained from example dataset . . . . .	38
3.4	Benchmark datasets characteristics and reduct size . . . . .	39
3.5	Comparison of classification accuracies for original datasets and reduced datasets by proposed model, Tan et. al. and Neumann et. al. approach using 10-fold cross validation . . . . .	39
3.6	Dataset characteristics and reduct size . . . . .	44
3.7	Performance evaluation parameters of learning algorithms with raw primary dataset using percentage split of 80:20 . . . . .	44
3.8	Performance evaluation parameters of learning algorithms with raw secondary dataset using percentage split of 80:20 . . . . .	44
3.9	Performance evaluation parameters of learning algorithms with reduced primary dataset using percentage split of 80:20 . . . . .	45
3.10	Performance evaluation parameters of learning algorithms with reduced secondary dataset using percentage split of 80:20 . . . . .	45
3.11	Performance evaluation parameters of learning algorithms with reduced primary dataset using 10-fold cross validation . . . . .	45
3.12	Performance evaluation parameters of learning algorithms with reduced secondary dataset using 10-fold cross validation . . . . .	46
3.13	Comparison of performance evaluation parameters of learning algorithms with reduced training set of primary dataset with percentage split of 80:20 using feature selection approaches of Tan et. al. and Neumann et. al. . . . .	46
3.14	Comparison of performance evaluation parameters of learning algorithms with reduced training set of secondary dataset with percentage split of 80:20 using feature selection approaches of Tan et. al. and Neumann et. al. . . . .	46
3.15	Comparison of the performance evaluation metrics of the current work with the previous method . . . . .	47
3.16	Performance evaluation parameters of learning algorithms with reduced Yi et. al. dataset (ACP_740) using percentage split of 80:20 . . . . .	48

3.17	Performance evaluation parameters of learning algorithms with reduced Thakur et. al. dataset (Antiviral) using percentage split of 80:20 . . . . .	48
3.18	Performance evaluation parameters of learning algorithms with reduced Manavalan et. al. dataset (Anti-hypertensive) using percentage split of 80:20 . . . . .	48
3.19	Performance evaluation parameters of learning algorithms with reduced Charoenkwan et. al. dataset (Bitter) using percentage split of 80:20 . . . . .	49
3.20	Comparison of the performance evaluation metrics of the current work with the previous method . . . . .	49
3.21	Dataset characteristics and reduct size . . . . .	53
3.22	Comparison of performance evaluation metrics for reduced training datasets by proposed model . . . . .	53
3.23	Comparison of performance evaluation metrics for reduced testing datasets by proposed model . . . . .	53
3.24	Comparison of average classification accuracies along with standard deviation for reduced training and testing datasets using proposed approach . . . . .	54
3.25	Comparison of the best values of the performance evaluation metrics of the current work with the values of previous existing methods on training dataset . . . . .	54
3.26	Comparison of the best values of the performance evaluation metrics of the current work with the values of previous existing methods on testing dataset . . . . .	54
4.1	Example dataset . . . . .	62
4.2	Similarity relation obtained from example dataset for attribute $a_1$ . . . . .	62
4.3	Granularity $[x]_{a_1}^c$ obtained from example dataset for attribute $a_1$ . . . . .	62
4.4	Intuitionistic fuzzy decision matrix . . . . .	63
4.5	Lower approximation obtained from example dataset for attribute $a_1$ . . . . .	63
4.6	Dataset characteristics and reduct size . . . . .	66
4.7	Comparison of classification accuracies for original datasets and reduced datasets by proposed model, and FMFRFS using full training . . . . .	66
4.8	Comparison of classification accuracies for original datasets and reduced datasets by proposed model, and FMFRFS using 10 fold cross validation . . . . .	66
5.1	Toy Example . . . . .	75
5.2	Benchmark datasets . . . . .	81
5.3	Benchmark datasets containing missing values . . . . .	82
5.4	Number of features selected by varying percentage of missing values . . . . .	83

5.5	Comparison of proposed approach on varying percentage of missing values . . . . .	84
5.6	Number of features selected by varying the value of $k$ (number of nearest neighbour for missing value imputation) . . . . .	85
5.7	Comparison of proposed approach on varying the value of $k$ (number of nearest neighbour for missing value imputation) . . . . .	85
5.8	Number of features selected by varying the percentage of correlated features employed for obtaining reduced dataset using proposed approach . . . . .	87
5.9	Comparison of proposed approach on varying the percentage of correlated features employed for obtaining reduced dataset using proposed approach . . . . .	87
5.10	Number of features selected by varying the percentage of missing entries required in an instance to ignore the instance . . . . .	88
5.11	Comparison of proposed approach on varying the percentage of missing entries required in an instance to ignore the instance . . . . .	89
5.12	Comparison with other missing value imputation algorithms . . . . .	90
5.13	Number of features selected on comparison with other state of art feature selection algorithms . . . . .	91
5.14	Comparison of classification accuracies with other state of art feature selection algorithms . . . . .	92
6.1	Example Dataset . . . . .	103
6.2	Lower Approximation of feature $a_1$ . . . . .	104
6.3	Benchmark Datasets . . . . .	110
6.4	IFBR results for various parameter combination . . . . .	112
6.5	IFBR classification accuracy for 90% coverage ( $\epsilon = 0.1$ ) . . . . .	113
6.6	IFBR classification accuracy for 80% coverage ( $\epsilon = 0.2$ ) . . . . .	114
6.7	IFBR classification accuracy for 70% coverage ( $\epsilon = 0.3$ ) . . . . .	115
6.8	IFBR results for variants of IS . . . . .	117
6.9	IFBR classification accuracy employing variants of IS . . . . .	117
6.10	IFBR overall reduction rate employing variants of IS . . . . .	118
6.11	Comparison with other state of art feature selection algorithms . . . . .	119
6.12	Classification accuracy comparison with other state of art feature selection algorithms . . . . .	120
6.13	Overall reduction rate comparison with other state of art feature selection algorithms . . . . .	121
6.14	Comparison with Instance Selection-Feature Selection combination . . . . .	122
6.15	Classification accuracy comparison with Instance Selection-Feature Selection combination . . . . .	123
6.16	Overall reduction rate comparison with Instance Selection-Feature Selection combination . . . . .	124
6.17	Comparison with Bireduct approach . . . . .	125

6.18	Classification accuracy comparison with Bireduct approach . . . . .	126
6.19	Overall reduction rate comparison with Bireduct approach . . . . .	127
6.20	IFBR results for various $\epsilon$ values on Anti-angiogenic dataset . . . . .	130
6.21	Performance evaluation metrics values on Anti-angiogenic dataset with IFBR . . . . .	130
6.22	Comparison of IFBR on Anti-angiogenic dataset with unreduced dataset	131
6.23	Comparison of IFBR on Anti-angiogenic dataset with HSBR, AntAn- gioCOOL [169], AntiAngioPred [123] and TargetAntiAngio [87] . . . . .	132
7.1	Example dataset . . . . .	140
7.2	Fuzzy similarity values for selected features . . . . .	140
7.3	Fuzzy similarity values for non-selected features . . . . .	140
7.4	Benchmark dataset . . . . .	146
7.5	Number of features selected by employing variants of proposed approach	147
7.6	Classification accuracy by employing variants of proposed approach .	148
7.7	Classification error results for variants of proposed approach . . . . .	149
7.8	Number of features selected on comparison with other state of art feature selection algorithm . . . . .	150
7.9	Classification error comparison with other state of art feature selec- tion algorithm . . . . .	151
7.10	Number of features selected on comparison with other state of art supervised feature selection algorithms . . . . .	153
7.11	Classification error comparison with other state of art supervised fea- ture selection algorithms . . . . .	154
7.12	Classification accuracy comparison with other state of art feature se- lection algorithm . . . . .	155
7.13	Number of features selected on comparison with other state of art non dependency based feature selection algorithms . . . . .	156
7.14	Classification accuracy comparison with other state of art non depen- dency based feature selection algorithms . . . . .	156
7.15	Classification error comparison with other state of art non dependency based feature selection algorithms . . . . .	157
7.16	Classification accuracy comparison with other state of art supervised feature selection algorithms . . . . .	158

## PREFACE

---

Due to advancement in modern technologies, various sources like network of sensors, interconnected devices, etc generate millions of data every day. This has lead to circumstances where proportion of data to the number of tools to access the same is large. Such ever expansive data is rich both in dimension and size (number of instances). Also, noise, human error in measurement, lack of proper communication, etc further lead to presence of irrelevant and redundant features, missing values in the dataset gathered. Hence, it is necessary to preprocess the datasets before applying any classification algorithm. Feature selection is a preprocessing step to remove irrelevant and/or redundant features and offers more concise and explicit descriptions of data. Feature selection has got wide applications in data mining, signal processing, bioinformatics, machine learning, etc. While instance selection removes conflicting or spurious data sample arising in the datasets. However, feature selection (FS) or instance selection (IS) alone cannot handle the ever increasing size and dimensionality of dataset. Both the aspects of data reduction must be taken into consideration for enhancing classification accuracy along with handling missing values and noise.

Rough set theory has been effectively employed as a tool for FS to solve many real-life problems without any additional parameter. However, one of the main limitations of rough set theory is discretization of data, which might lead to information loss. Fuzzy and intuitionistic fuzzy rough set comes in handy as a tool to overcome the limitations of rough set theory. Further, these tools precisely handle the vagueness and uncertainty arising in the data.

This thesis dive into the details of applying data pre-processing techniques like missing value imputation, noise removal, feature selection, data reduction or bireduct generation using fuzzy rough and intuitionistic fuzzy rough set assisted techniques.



Mathematical formulation of each concept along with underlying model construction is introduced herein. Hence, an exhaustive study is conducted covering areas such as data reduction, missing value imputation, noise removal, etc both in supervised and unsupervised domain.

*DEDICATION*  
*To*  
*My Beloved Grandfather*

