

DECLARATION

I, Mr. Bhavani Sankar Panda, a scholar in the department of Computer Science and Engineering, Centurion University of Technology and Management, Paralakhemundi, Odisha, hereby declare that the thesis entitled “**Expert System Development for Retrieval of Missing Data through Soft Computing Techniques**” being submitted to the Faculty of Engineering & Technology, Centurion University of Technology and Management, is a record of the my own work, carried out under the supervision of Dr. Sasanko Sekhar Gantayat and Dr. Ashok Misra. I also declare that the matter embodied in this thesis has not been submitted for the award of any other degree either to this university or any other university or institute.

Bhavani Sankar

Panda

CERTIFICATE

Certified that the thesis entitled “**Expert System Development for Retrieval of Missing Data through Soft Computing Techniques**” being submitted to the Faculty of Engineering & Technology, Centurion University of Technology and Management, Paralakhemundi, by Mr. Bhavani Sankar Panda in fulfillment of the requirements for the award of the degree of “Doctor of Philosophy” in Computer Science and Engineering is a record of the scholar’s own work, carried out by him under our supervision and guidance. We certify that to the best of our knowledge, the matter embodied in this thesis has not been submitted for the award of any other degree either to this university or any other university or institute.

Dr. Sasanko Sekhar Gantayat
Dept. of Computer Science & Engg.,
Technology,
GMR Institute of Technology,
Rajam, Andhra Pradesh.

Dr. Ashok Misra
Centurian University of
Paralakhemundi, Odisha.

ACKNOWLEDGEMENTS

I am greatly indebted to express my sincere thanks and deep gratitude **The Almighty God** for blessing, protecting and guiding me throughout this period. I could never have accomplished this without the faith I have in the God.

I express my profound sense of reverence to my supervisor **Prof. Dr. Sasanko Sekhar Gantayat**, for his constant guidance, support, motivation and untiring help during the course of my Ph.D. His in-depth knowledge on a broad spectrum of Computational Intelligence topics has been extremely beneficial for me. He has given me enough freedom during my research, and he has always been nice to me. I will always remember his calm and relaxed nature. I am thankful to the Almighty for giving me a mentor like him.

I express my deepest gratitude to my co-supervisor **Prof. Dr. Ashok Misra**, Centurian University of Technology and Management, Parlakhemundi for boosting my morale throughout the course of research. He has always been caring, a source of wisdom and motivation. He is a great leader.

I would also like to thank the members of my Doctoral Committee, **Dr. M. R. Senapati** (CIT, Jatni Campus), **Dr. Anita Patra** (Dean SOET, PKD), **Dr. P.S.V Ramana Rao** (RC Member, CUTM) and **Dr. M.L. Narasimham** (Dean, Academic Research) for accepting my submission of this thesis.

I would also like to thank all the member faculties of my department for their source of help and support, not only during the thesis work, but also throughout my Ph.D. tenure. I would like to thank all those who directly or indirectly helped me throughout the project.

At last I would like to thank my wife **Meenakshi** and my son **Akash** for their love, affection, help and support without which this project has been an incomplete one.

**Bhavani Sankar
Panda**

CONTENTS

CHAPTER NO.	TITLE	PAGE
-------------	-------	------

NO.

	DECLARATION.	i
	CERTIFICATE	ii
iii	ACKNOWLEDGEMENTS	
	CONTENTS	iv
	LIST OF TABLE	x
xi	LIST OF FIGURES	
xii	LIST OF ABBRIVATIONS	
xiii	ABSTRACT	
1	CHAPTER – 1 INTRODUCTION	1
	1.1 Scope of Work and Problem Definition	2
	1.2 Objectives	2
	1.3 Contributions	2
	1.4 Organization of the Thesis	3
2	CHAPTER – 2 BACKGROUND AND LITERATURE STUDY	4
	2.1 Introduction to Missing Data	4
	2.2 Missing Data Mechanism	6
	2.2.1 Missing completely at random (MCAR)	6
	2.2.2 Missing at random (MAR)	6
	2.2.3 Missing not at random (MNAR)	6
	2.3 Problem with Incomplete Information	8
	2.4 Missing Data Techniques	8
	2.4.1 Imputation	8
	2.4.2 Mean imputation	9
	2.4.3 Imputation based on logical rules	9
	2.4.4 Matching and hot-deck imputation	9
	2.4.5 Model-based imputation	10
	2.4.6 Multiple imputations	10
	2.4.7 Maximum Likelihood	11
	2.4.8 K-means clustering	12
	2.4.9 Missing Data Imputation Based on Rough Sets	13
	2.4.10 Tolerance Relation	13

2.4.11 Similarity Relation	13
2.4.12 LERS	13
2.4.13 Predictive Value Imputation (PVI)	14
2.4.14 Distribution-Based Imputation (DBI)	14
2.4.15 Unique-Value Imputation (UVI)	14
2.4.16 Reduced-Feature Models	15
2.1.17 Other Methodologies	15
2.5 Handling Missing Values in Classification Models	15
2.6 Treatments for Missing Values at Prediction Time	16
2.7 Flexible Indiscernibility Relations for Missing Attribute Values	17
2.8 On Indiscernibility Relations for Missing Attribute Values	17
2.9 On Decomposition for Incomplete Data	18
2.10 Missing Mechanism Limitations	19
2.11 Significance of the Study	20
2.12 Conclusion	21
3 CHAPTER – 3 ROUGH SET, FUZZY SET, CBRS & SOFT SET.....	22
3.1 Introduction	22
3.2 Rough Set Concepts and Notions	23
3.3 Covering Based Rough Set	27
3.3.1 Definition and Notations	27
3.3.2 Application of Coverage based Rough Set	28
3.4 Concepts of Soft Sets	29
3.4.1 Definition and Notations	29
3.4.2 Applications of Soft Set	30
3.5 Fuzzy Sets	30
3.5.1 Definition and Notations	30
3.5.2 Fuzzy Similarity Relation	31
3.5.3 Similarity Using Fuzzy Reflexive Relation	31
3.5.4 Fuzzy Ambiguity	32
3.5.5 Generalized Definition of Rough Approximation on Fuzzy Similarity Relations	33
3.5.6 Fuzzy Partitions over Lower and Upper Approximations on X	33
3.5.7 Fuzzy Sets vs. Rough Sets	33
3.6 Rough Equality of Set	35
3.6.1 Definition	35
3.6.2 Properties	35
3.7 Rough Equivalence of Sets	38
3.7.1 Definition	38
3.7.2 Elementary Properties	38
3.8 RST Application in Data Analysis	38
3.8.1 Information Systems	39

3.9	Approximations of Classifications	40
3.10	Covering based Approximations of Classifications	41
3.10.1	Definition	41
3.10.2	Covering reducts	41
3.10.3	Indiscernibility Relation	43
3.10.4	Set Approximation	44
3.10.5	List of properties of Approximations	45
3.11	Some more Application in RST	45
3.11.1	Prediction of Business failure	45
3.11.2	Financial Investment	46
3.11.3	Bioinformatics and Medicine	47
3.11.4	Fault Diagnosis	47
3.11.5	Spatial and Meteorological Pattern	48
3.11.6	Music and Acoustics	48
3.11.7	Feature Selection	49
3.12	Conclusion	50
4	CHAPTER – 4 ROUGH SET APPROACH TO RULE INDUCTION FROM MISSING DATA ...	51
4.1	Introduction	51
4.2	Database Evaluation	51
4.3	Sequential Methods	52
4.3.1	Deleting Cases with Missing Attribute Values	52
4.3.2	The Most Common Value of an Attribute	52
4.3.3	The Most Common Value Of an Attribute Restricted To A Concept	53
4.3.4	Assigning All Possible Attribute Values to A Missing Attribute Value	53
4.3.5	Assigning all Possible Attribute Values Restricted To a Concept	54
4.3.6	Replacing Missing Attribute Values by the Attribute Mean	54
4.4	Parallel Methods	55
4.4.1	Blocks of Attribute-Value Pairs	55
4.4.2	Characteristic Sets	56
4.4.3	Lower and Upper Approximations	57
4.4.4	Rule Induction - MLEM2	57
4.5	Conclusion	58
5	CHAPTER – 5 ROUGH SET RULE BASED TECHNIQUE FOR THE RETRIEVAL OF MISSING DATA IN MALARIA DISEASES DIAGNOSIS	59
5.1	Introduction	59
5.2	Literature Review	59
5.3	Rule Induction	60
5.4	Rough set approach when values were lost	60
5.4.1	Rough set approach when values were interpreted as a “do not care condition”	60

5.4.2	RS approach when some values were lost and some interpreted as a “do not care condition”	60
5.4.3	Definition and Notations	61
5.5	Malaria Data Set with Missing Attribute	61
5.6	Proposed System	63
5.7	Cut Set for Overlapping data	63
5.8	Rough Set Rule Based Technique	64
5.9	Results and Discussion	65
5.10	Conclusion	66
6	CHAPTER – 6 APPLICATION ON ROUGH SET, COVERING BASED ROUGH SET AND SOFT SET IN MISSING INFORMATION SYSTEM	67
6.1	Introduction	67
6.2	Literature Review	68
6.3	Missing Attribute values	68
6.4	Results and Discussion	69
6.4.1	Example1	69
6.4.2	Example2	70
6.4.3	Example3	70
6.5	Conclusion	71
7	CHAPTER - 7 METHODS AND COMPARISONS OF HANDING MISSING DATA IN STUDENT DATA ANALYSIS USING ROUGH SET AND SOFT SET	72
7.1	Introduction	72
7.2	Background and Literature	72
7.3	Approaches to Handling Missing Data	73
7.3.1	Rough Set Analysis	74
7.3.2	Soft Set Analysis	76
7.3.3	Complete Delete Case (CDC) Analysis	77
7.3.4	Last Observation Carried Forward (LOCF)	78
7.4	Missing Data Generation	78
7.5	Simulation Results	79
7.6	Discussion and Conclusions	80
8	CHAPTER – 8 ROUGH SET APPROACH TO DEVELOPMENT OF A SOFT KNOWLEDGE- BASED EXPERT SYSTEM	82
8.1	Introduction	82
8.2	Expert Systems	82
8.3	Knowledge-Based Expert System	83
8.3.1	Development of a Soft Knowledge-Based Expert System	83
8.3.2	Components of a Soft Knowledge-based Expert System	83
8.4	Rough Set Approach to Design Soft K-B Expert System	85

8.4.1	Uncertainty, Inconsistency, Imprecision and Missing Information	85
8.4.2	Modeling	85
8.5	Conclusion	85
9	CHAPTER – 9 CONCLUSION	87
9.1	Summary	87
9.2	Future Research.....	87
	REFERENCES	88
	DERIVED PUBLICATION.....	99
	CURRICULUM VITAE	
	100	

LIST OF TABLES

2.1	Data set with missing attribute values	5
3.1	Basic Classes of Rough Sets	26
3.2	Information System before Iraq war	39
3.3	Information System after Iraq war	39
3.4	Information System 1.....	42
3.5	Information System 2	43

4.1 missing attribute values	51
4.2 Deleting Cases with missing attribute values	52
4.3 Most common values	52
4.4 Most common values	53
4.5 All possible values	53
4.6 All possible values	54
4.7 Replacing Missing attribute by Mean	54
4.8 Replacing Missing attribute by Mean	55
4.9 Missing Attributes	55
4.10 do-not care Attributes	56
5.2. Malaria Data set with missing attributes	62
5.3. After applying cut set	64
5.4. Imputation Result for Missing data with Actual Data	65
5.6. The Missing Data Filled with Observed Data	66
6.1. Data set with missing attributes	69
6.2. The Missing Data Filled with Observed Data	70
7.1. Student information system with missing values	74
7.2: Student information system with RST analysis	76
7.3. Analysis using Soft Set	77
7.4: Student information system with CDC analysis	77
7.5: Student information system with LOCF analysis	78
7.6: Comparison of different Methods	79

LIST OF FIGURES

3.1 Rough Set Approach	24
3.2 Approximations	44
5.2. Flow chart of the Rule Based System	63
7.1. Comparison of different Methods	80
7.2. Original Data Set	80
7.3. LOCF Analysis	80
7.4. RST Analysis	80
7.5. Soft Set Analysis	80
8.1. Basic concept of an Expert System	82
8.2. Development of a Soft K-B Expert system	84

ABBREVIATIONS

RF Set	Rough Fuzzy Set
MCAR	Missing completely at random
MAR	Missing at random
MNAR	Missing not at random
ML	Maximum Likelihood
MI	Multiple imputations
LEERS	Learning from Examples based on Rough Sets
PVI	Predictive Value Imputation
DBI	Distribution-Based Imputation
UVI	Unique-Value Imputation
LEM2	Learning from Examples Module, version 2
MLEM2	Modified Learning from Examples Module, version 2
RST	Rough Set Theory
CBRS	Covering Based Rough Set
SS	Soft Sets
KB	Knowledge Base
AI	Artificial Intelligence

ABSTRACT

Missing data is a familiar and unavoidable problem in large datasets and is widely discussed in the field of data mining and statistics. The impact of missing data on quantitative research can be serious, leading to biased estimates of parameters, loss of information, decreased statistical power, increased standard errors, and weakened generalizability of findings. All researchers have faced the problem of missing quantitative data at some point in their work. Research informants may refuse or forget to answer a survey question, files are lost, or data are not recorded properly. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

Many techniques for handling missing data have been proposed in this literature. Most of these techniques are overly complex. This thesis explores an imputation technique based on rough set computations. In this thesis, characteristic relations are introduced to describe incompletely specified decision tables. It is shown that the basic rough set idea of lower and upper approximations for incompletely specified decision tables may be defined in a variety of different ways. Empirical results obtained using real data are given and they provide a valuable and promising insight to the problem of missing data. Missing data were predicted with an accuracy of up to 99%.

The data base of a Knowledge-Based System (KBS) contains inaccuracies and uncertainties which are inherent in the description of the rules given by the expert. They are due to the difficulty of representing the facts involved in the antecedents and the consequents of the inference rules, which are expressed in most cases by ambiguous characterizations (such as a color or an age described as “young”), or by imprecise data, (for instance, a length described as “approximately equal to 15 meters”). Uncertainties appear, particularly when the expert is not certain of the validity of the rule in any cases. Another problem stems from the

utilization of these rules when the observed facts are not identical with the condition expressed in their premises, but are not too different from them. In a process using classical logic, the rule would not work in this case, but the fact is sometimes so close to the characterization indicated in the premise of a rule that it seems interesting to obtain a deduced fact, even if we must restrict its validity through a coefficient of uncertainty.