SOFT COMPUTING FOR MICROARRAY DATA ANALYSIS

P. Ganesh Kumar
(*Assistant Professor*)
Anna University,
Coimbatore, Tamil Nadu, INDIA.

T. Aruldoss Albert Victoire
(*Associate Professor*)
Anna University,
Coimbatore,
Tamil Nadu, INDIA.

D. Devaraj
(*HOD & Senior Professor*)
Kalasalingam University,
Krishnankoil,
Tamil Nadu, INDIA.

SOFT COMPUTING FOR MICROARRAY DATA ANALYSIS

Copyright © : P. Ganesh Kumar
Publishing Right Ⓟ : VSRD Academic Publishing
A Division of Visual Soft (India) Pvt. Ltd.

ISBN-13: 978-81-931580-9-8
FIRST EDITION, AUGUST 2016, INDIA

Typeset, Printed & Published by:
VSRD Academic Publishing (*A Division of Visual Soft (India) Pvt. Ltd.*)

Disclaimer: The author(s) are solely responsible for the contents of the papers compiled in this book. The publishers or its staff do not take any responsibility for the same in any manner. Errors, if any, are purely unintentional and readers are requested to communicate such errors to the Editors or Publishers to avoid discrepancies in future.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photo-copying, recording or otherwise, without the prior permission of the Publishers & Author.

Printed & Bound in India

VSRD ACADEMIC PUBLISHING
A Division of Visual Soft (India) Pvt. Ltd.

REGISTERED OFFICE
154, Tezabmill Campus, Anwarganj, KANPUR – 208 003 (UP) (IN)
Mob.: +91 99561 27040, Ph.: +91 512 6553705
Web.: www.vsrdpublishing.com, Email: vsrdpublishing@gmail.com

MARKETING OFFICE (NORTH INDIA)
Basement-2, Villa-10, Block-V, Charmwood Village, FARIDABAD–121009 (HY)(IN)
Mob.: +91 98999 36803, Ph.: +91 129 4036803
Web.: www.vsrdpublishing.com, Email: vsrdpublishing@gmail.com

MARKETING OFFICE (SOUTH INDIA)
340, FF, Adarsh Nagar, Oshiwara, Andheri(W), MUMBAI–400053 (MH)(IN)
Mob.: +91 9956127040
Web.: www.vsrdpublishing.com, Email: vsrdpublishing@gmail.com

PREFACE

The high dimensional low sample and one-sidedness nature of microarray data poses a lot of difficulties for a computer based diagnostic system. Hence careful analysis of microarray data has a significant impact on identification of diseases. This book introduces multistage mutual information (MSMI) for gene selection that computes Mutual Information (MI), initially using all the samples and based on the classification performance produced by Artificial Neural Network (ANN), MI is repeatedly calculated using only the unclassified samples until there is no improvement in the classification accuracy. A Fuzzy-Rough set based *f*-Information (FR*f*-I) approach is discussed to avoid discretization during gene selection that generates Fuzzy Equivalence Class (FEC) for every gene and then constructs a Fuzzy Equivalence Partition Matrix for estimating the Gene-Class relevance and Gene-Gene redundancy.

Even though the genes selected by the FR*f*-I method are highly informative, the classification results produced by ANN are hard to interpret as they focus only on the classification accuracy. Expert Systems based on Fuzzy Logic (FL) are highly required in medical diagnosis because of its ability to incorporate human expert knowledge in decision making. In this book, the issue of knowledge acquisition is formulated as an optimization problem and a novel Genetic Swarm Algorithm (GSA) is introduced for optimal generation of rule set and tuning of membership function. Even though the results produced by the GSA approach produces a well-informed classifier model, the knowledge produced by it is difficult

to understand. A hybrid Ant Bee Algorithm (ABA) is discussed to address this interpretability-accuracy tradeoff.

✍ *P. Ganesh Kumar*

CONTENTS

CHAPTER 1 INTRODUCTION 1

1.1 GENERAL 3

1.2 DNA MICROARRAY TECHNOLOGY 4

1.3 IMPORTANCE OF MICROARRAY DATA 6

1.4 CHALLENGES OF ANALYZING MICROARRAY DATA 8

1.5 FOCUS OF THE BOOK 9

1.6 APPROACHES FOR GENE SELECTION AND SAMPLE CLASSIFICATION 10

1.7 COMPUTATIONAL INTELLIGENT TECHNIQUES 11

1.8 METHODOLOGIES 12

1.9 ORGANIZATION OF THE BOOK 13

CHAPTER 2 ALGORITHMS FOR GENE SELECTION AND SAMPLE CLASSIFICATION 15

2.1 INTRODUCTION 17

2.2 EXISTING GENE SELECTION ALGORITHMS 18

2.2.1 STATISTICAL APPROACHES 19

2.2.2 INFORMATION THEORETIC APPROACHES 20

2.3 SUGGESTED METHODOLOGY FOR GENE SELECTION 22

2.3.1 MULTI STAGE MUTUAL INFORMATION 23

2.3.2 FUZZY-ROUGH SET BASED *F*-INFORMATION 23

2.4 EXISTING SAMPLE CLASSIFICATION ALGORITHMS 24

2.4.1 STATISTICAL APPROACHES 24

2.4.2 RULE BASED APPROACHES 24

2.5 SUGGESTED APPROACH FOR SAMPLE CLASSIFICATION 25

2.5.1 ARTIFICIAL NEURAL NETWORK 26

2.5.2 FUZZY EXPERT SYSTEM 29

2.6 SUMMARY 32

CHAPTER 3 GENE SELECTION USING MULTISTAGE MUTUAL INFORMATION 33

3.1 INTRODUCTION 35

3.2 CONCEPT OF MUTUAL INFORMATION 35

3.3 MULTI STAGE MUTUAL INFORMATION 37

3.4 ANN BASED CLASSIFICATION 40

3.5 SIMULATION RESULTS 41
3.5.1 GENE EXPRESSION DATASETS 42
3.5.2 ESTIMATION OF MI USING WHOLE SAMPLES 44
3.5.3 CLASSIFICATION OF SAMPLES USING ANN 46
3.5.4 RE-ESTIMATION OF MI USING ONLY UNCLASSIFIED SAMPLES 53
3.5.5 MONTE-CARLO CROSS VALIDATION (MCCV) 57
3.5.6 PERFORMANCE COMPARISON 60

3.6 SUMMARY 62

CHAPTER 4 FUZZY ROUGH SET APPROACH FOR *F*-INFORMATION BASED GENE SELECTION 63

4.1 INTRODUCTION 65

4.2 CONCEPT OF *F*-INFORMATION MEASURE 66

4.3 BASICS OF ROUGH SET 69

4.4 FUZZY-ROUGH SET 71

4.5 FUZZY ROUGH SET BASED *F*-INFORMATION (FR*F*-I) FOR GENE SELECTION 72
4.5.1 GENERATION OF FEPM MATRIX 73
4.5.2 COMPUTATION OF *F*-INFORMATION MEASURE 75
4.5.2.1 GENE-CLASS RELEVANCE 76
4.5.2.2 GENE-GENE REDUNDANCY 78
4.5.3 FORMATION OF CANDIDATE GENES 80

4.6 SIMULATION RESULT 81
4.6.1 COMPUTATION OF FEC AND FEPM 82
4.6.2 CALCULATION OF GENE-CLASS RELEVANCE AND GENE-GENE REDUNDANCY 84

4.6.3 CLASSIFICATION USING ANN 86
4.6.4 WILCOXON SIGNED RANK TEST 90
4.6.5 GENE ONTOLOGY (GO) ANALYSIS 94

4.7 SUMMARY 96

CHAPTER 5 FUZZY EXPERT SYSTEM BASED SAMPLE CLASSIFICATION USING GENETIC SWARM ALGORITHM 99

5.1 INTRODUCTION 101

5.2 MICROARRAY SAMPLE CLASSIFICATION USING FUZZY EXPERT SYSTEM 102

5.3 GENETIC SWARM ALGORITHM 105

5.4 GSA IMPLEMENTATION 107
5.4.1 REPRESENTATION 107
5.4.2 FITNESS FUNCTION FORMATION 110
5.4.3 OPERATIONS FOR RULE SET 111
5.4.4 OPERATIONS FOR MEMBERSHIP FUNCTION 113

5.5 SIMULATION RESULT 116
5.5.1 LEARNING ABILITY 116
5.5.2 GENERALIZATION ABILITY 123
5.5.3 SENSITIVITY – SPECIFICITY ANALYSIS 128

5.6 SUMMARY 133

CHAPTER 6 HYBRID ANT BEE ALGORITHM FOR FUZZY EXPERT SYSTEM BASED SIMPLE CLASSIFICATION 135

6.1 INTRODUCTION 137

6.2 INTERPRETABILITY – ACCURACY TRADEOFF 138

6.3 ANT BEE ALGORITHM (ABA) 140
6.3.1 MODIFIED REPRESENTATION 142
6.3.2 OPERATIONS OF ACO 143
6.3.3 OPERATIONS OF ABC 145

6.4 SIMULATION RESULT 145
6.4.1 ACCURACY 146
6.4.2 INTERPRETABILITY 149
6.4.3 ROC ANALYSIS 158
6.4.4 PERFORMANCE COMPARISON 162

6.5 SUMMARY 167

CHAPTER 7 MATLAB – A QUICK GLANCE 169

7.1 INTRODUCTION 171

7.2 MATLAB VARIABLES 171

7.3 MATRIX AND VECTOR CREATION 172

7.4 MATRIX MANIPULATION 173

7.5 SIMPLE MATLAB PROGRAMS 175

7.6 PLOTTING GRAPHS 179

7.7 MATLAB EXERCISE 187

CHAPTER 8 REFERENCES 191

REFERENCES 193