



Cairo University

A NOVEL ALGORITHM FOR FUZZY-GENETIC DISTRIBUTED DATA MINING

By

Hassan Ahmed Hassan M. Abounaser

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in
Computer Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2017

A NOVEL ALGORITHM FOR FUZZY-GENETIC DISTRIBUTED DATA MINING

By

Hassan Ahmed Hassan M. Abounaser

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in
Computer Engineering

Under the Supervision of

Prof. Dr. Ihab El-Sayed Talkhan

Prof. Dr. Ahmed Fahmy Amin

Professor
Head of Computer Engineering Department
Faculty of Engineering
Cairo University

Professor,
Computer Engineering Department
College of Engineering & Technology
Arab Academy for Science, Technology &
Maritime Transport (AASTMT), Cairo

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2017

A NOVEL ALGORITHM FOR FUZZY-GENETIC DISTRIBUTED DATA MINING

By

Hassan Ahmed Hassan M. Abounaser

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in
Computer Engineering

Approved by the
Examining Committee

Prof. Dr. Ihab El-Sayed Talkhan , Thesis Main Advisor

Prof. Dr. Ahmed Fahmy, Member
Professor, College of Engineering & Technology, Arab
Academy for Science, Technology & Maritime Transport
(AASTMT), Cairo

Prof. Dr. Mohamed Zaki Abd El-Magid , Examiner
Professor, Faculty of Engineering, Al Azhar University

Prof. Dr. Amr Anwar Badr, External Examiner
Professor, Faculty of Computers and Information, Cairo
University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2017

Engineer: Hassan Ahmed Hassan M. Abounaser
Date of Birth: 23 / 7 / 1979
Nationality: Palestinian
E-mail: hassan.abounasser@aast.edu
Phone: +201006620848
Address: 3078b, Zahraa- Nasr City, Cairo, Egypt
Registration Date: 1 / 10 / 2010
Awarding Date: 2 / 2 / 2017
Degree: Doctor of Philosophy
Department: Computer Engineering



Supervisors: Prof. Dr. Ihab El-Sayed Talkhan
Prof. Dr. Ahmed Fahmy Amin
Professor, College of Engineering & Technology- Arab Academy
for Science, Technology & Maritime Transport (AASTMT), Cairo

Examiners: Prof. Dr. Amr Anwar Badr (External Examiner)
Professor, Faculty of Computers and Information, Cairo University
Prof. Dr. Mohamed Zaki Abd El-Magid (Examiner)
Professor, Faculty of Engineering, Al Azhar University
Prof. Dr. Ihab El-Sayed Talkhan (Thesis Main Advisor)
Prof. Dr. Ahmed Fahmy Amin (Member)

Title of Thesis: A Novel Algorithm For Fuzzy-Genetic Distributed Data Mining

Key Words:- Fuzzy Classification; Rule-base; Fuzzy Logic System (FLS);
Genetic Algorithm (GA); Distributed Data Mining (DDM)

Summary:

A novel framework for a Parallel Fuzzy-Genetic Algorithm (PFGA) has been developed for classification and prediction over decentralized data sources as a main contribution to the scientific community. The model parameters are evolved using two nested genetic algorithms (GAs). The outer GA evolves the fuzzy sets whereas the inner GA evolves the fuzzy rules. During optimization, best rules are only distributed and exchanged among agents to construct the overall optimized model. Several experiments have been conducted and show that the developed model has good accuracy and more efficient in performance and comprehensibility of linguistic rules compared to some models implemented in KEEL software tool.

Acknowledgements

I would like gratefully to acknowledge all the following people who for various reasons were involved in contributing to this work, and for the help and time, they have given me over the work of this thesis.

First, I would like to thank my supervising committee, Prof. Ihab Talkhan and Prof. Ahmed Fahmy. This work would never have been successfully completed without their help, guidance and continuous support.

Second, special thanks go to my family for their love, patience, care and support during the period I spent working on this thesis. In particular, I would like to thank my parents for their endless encouragement, understanding and support.

Third, thanks also go to the staff of the Computer Engineering Department in AASTMT for the enlightening discussions and observations they made. In particular, I would like to thank Dr. Sherif Fadel for his valuable comments, suggestions and advice.

A final word of thanks is owed to my best friends, particularly, Dr. Mohamed Almoghalis, for his valuable support and advice.

Sincerely,

Hassan Ahmed Hassan M. Abounaser

Dedication

This thesis work is dedicated to my dear parents, who have always loved me unconditionally and whose good examples have taught me to work hard for the things that I aspire to achieve. This work is also dedicated to my lovely brothers and sisters who have been a constant source of support and encouragement during the challenges of life. I am truly thankful for having you in my life.

Table of Contents

ACKNOWLEDGMENTS	I
DEDICATION	II
TABLE OF CONTENTS	III
LIST OF TABLES	V
LIST OF FIGURES	VI
LIST OF ABBREVIATIONS	XI
ABSTRACT	XIII
CHAPTER 1: INTRODUCTION	1
1.1. OVERVIEW	1
1.2. MOTIVATION	1
1.3. THE CHALLENGES IN DATA MINING	4
1.4. THESIS OBJECTIVES	6
1.5. THESIS ORGANIZATION	6
CHAPTER 2: LITERATURE REVIEW	8
2.1. OVERVIEW	8
2.2. EVOLUTION OF DATA MINING STRATEGIES	8
2.2.1. CENTRALIZED APPROACH	8
2.2.2. PARALLEL APPROACH	10
2.2.3. DISTRIBUTED APPROACH	11
2.3. RELATED WORK	13
CHAPTER 3: FUZZY LOGIC SYSTEMS	16
3.1. OVERVIEW	16
3.2. FUZZY SET THEORY	16
3.3. APPLICATIONS OF FUZZY LOGIC SYSTEMS	17
3.4. COMPONENTS OF A FUZZY LOGIC SYSTEM	18
3.4.1. FUZZIFIER	18
3.4.2. KNOWLEDGE BASE	18
3.4.3. FUZZY INFERENCE ENGINE	19
3.4.4. DEFUZZIFIER	19
3.5. FUZZY KNOWLEDGE BASE REPRESENTATION METHODOLOGY	19
3.5.1. CONDITIONAL-SENTENCES REPRESENTATION METHOD	19
3.5.2. FAM MATRIX REPRESENTATION METHOD	20

3.6. PARALLELISM OF A FUZZY LOGIC SYSTEM.....	21
3.7. MAMDANI FUZZY LOGIC SYSTEM.....	22
3.8. DESIGN APPROACHES FOR A FUZZY LOGIC SYSTEM.....	22
CHAPTER 4: OPTIMIZATION USING EVOLUTIONARY ALGORITHMS	24
4.1. OVERVIEW.....	24
4.2. GENETIC ALGORITHMS THEORY.....	24
4.2.1. APPLICATIONS OF GENETIC ALGORITHMS.....	25
4.2.2. CANONICAL GENETIC ALGORITHM.....	25
CHAPTER 5: PROPOSED SYSTEMS.....	28
5.1. OVERVIEW.....	28
5.2. FUZZY LOGICCLASSIFIER AND PREDICTOR.....	28
5.2.1. MEMBERSHIP FUNCTIONS DESIGN.....	29
5.2.2. RULE-BASE REPRESENTATION METHODOLOGY.....	29
5.2.3. INFERENCE ENGINE AND DEFUZZIFICATION TECHNIQUE UTILIZED	30
5.3. PROPOSED FUZZY-GENETIC SYSTEM.....	31
5.3.1 EVOLVED FUZZY LOGIC CLASSIFIER AND PREDICTOR.....	31
5.3.2 FUZZY-GENETIC ALGORITHM AGENT.....	31
5.3.3 PARALLEL FUZZY-GENETIC FRAMEWORK STRUCTURE.....	37
CHAPTER 6: RESULTS AND DISCUSSION.....	40
6.1. OVERVIEW.....	40
6.2. RESULTS.....	40
6.3. DISCUSSION.....	41
6.4. CASE STUDY.....	61
CHAPTER 7: CONCLUSIONS.....	67
7.1. CONCLUSION AND RECOMMENDATIONS.....	67
7.2. FUTURE WORK.....	67
REFERENCES.....	69
APPENDIX A: INFORMATION REPRESENTATION AND PROCESSING	
APPROACHES.....	77

List of Tables

Table 2.1 :	List of top 10 Algorithms in DM research community.....	9
Table 2.2 :	Taxonomy of DDM Algorithms in DM research community.....	12
Table 6.1 :	List of datasets employed in experiments	40
Table 6.2 :	Results of PFGA framework versus FH-GBML and GFS-RB-MF algorithms when $N_1=N_2=20$, $N_g=500$, and $N_a=5$	57
Table 6.3 :	Results of PFGA framework versus FH-GBML and GFS-RB-MF algorithms when $N_1=N_2=20$, $N_g=500$, and $N_a=10$	58
Table 6.4 :	Results of PFGA framework versus FH-GBML and GFS-RB-MF algorithms when $N_1=N_2=40$, $N_g=1000$, and $N_a=5$	59
Table 6.5 :	Results of PFGA framework versus FH-GBML and GFS-RB-MF algorithms when $N_1=N_2=40$, $N_g=1000$, and $N_a=10$	60
Table 6.6:	Assumed integer keys for fuzzy sets as a sample case study	63
Table A.1:	Comparison between different Information Representation Approaches.....	78

List of Figures

Figure 1.1:	The brief structure of a Data Mining System as a centralized data model in distributed environment from classical technique's perspective	4
Figure 2.1:	The brief structure of a Data Mining System in centralized approach	8
Figure 2.2:	The brief structure of a Data Mining System in Distributed approach	11
Figure 3.1:	The structure of a Fuzzy Logic System (FLS) and its components interconnections	18
Figure 3.2:	A two-dimensional FAM matrix structure.....	20
Figure 3.3:	Example of membership functions for fuzzy sets (a) Trapezoid, (b) Triangular, (c) Logistic, and (d) Bell Shape.....	21
Figure 4.1:	A flow chart illustrates the Canonical GA.....	26
Figure 5.1:	The brief structure of a FGA agent constructing its local model from the dataset.....	31
Figure 5.2:	The structure of nested GAs that evolves local model parameters of FGA agent.....	32
Figure 5.3:	Example of designing inner GA chromosome encoding rule-base of 3 fuzzy rules.....	33
Figure 5.4:	Example of single-point crossover operation in inner GA where crossover points can be different positions	34
Figure 5.5:	Example of designing a structure encoding 3 fuzzy sets defined by triangular membership functions utilized for continuous input attribute.....	35
Figure 5.6:	The general structure of outer GA chromosome encoding fuzzy sets utilized in all dataset attributes along with its class attribute y.....	35
Figure 5.7:	Single-point crossover operation in outer GA where crossover points must have identical position.....	36

Figure 5.8:	The detailed structure of a FGA agent constructing its local model from the dataset.....	36
Figure 5.9:	The structure of PFGA that accepts a datasets distributed over distributed and decentralized data sources and construct the final model from the cooperative local models of FGA agents.....	37
Figure 5.10:	PFGA pseudo-code.....	38
Figure 5.11:	A flow chart illustrates the Parallel Fuzzy-Genetic Algorithm (PFGA) for classification and prediction in distributed environment.	39
Figure 6.1:	Example of best fitness data for $N_1=N_2=20$ using “banana” dataset for $N_g=500$ and $N_a=5$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	42
Figure 6.2:	Example of best fitness data for $N_1=N_2=20$ using “banana” dataset for $N_g=500$ and $N_a=10$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	42
Figure 6.3:	Example of best fitness data for $N_1=N_2=40$ using “banana” dataset for $N_g=1000$ and $N_a=5$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	44
Figure 6.4:	Example of best fitness data for $N_1=N_2=40$ using “banana” dataset for $N_g=1000$ and $N_a=10$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	44
Figure 6.5:	Example of best fitness data for $N_1=N_2=20$ using “haberman” dataset for $N_g=500$ and $N_a=5$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	45
Figure 6.6:	Example of best fitness data for $N_1=N_2=20$ using “haberman” dataset for $N_g=500$ and $N_a=10$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	45

Figure 6.7:	Example of best fitness data for $N_1=N_2=40$ using “haberman” dataset for $N_g=1000$ and $N_a=5$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	46
Figure 6.8:	Example of best fitness data for $N_1=N_2=40$ using “haberman” dataset for $N_g=1000$ and $N_a=10$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	46
Figure 6.9:	Example of best fitness data for $N_1=N_2=20$ using “saheart” dataset for $N_g=500$ and $N_a=5$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm	48
Figure 6.10:	Example of best fitness data for $N_1=N_2=20$ using “saheart” dataset for $N_g=500$ and $N_a=5$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	48
Figure 6.11:	Example of best fitness data for $N_1=N_2=20$ using “saheart” dataset for $N_g=500$ and $N_a=10$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm	49
Figure 6.12:	Example of best fitness data for $N_1=N_2=20$ using “saheart” dataset for $N_g=500$ and $N_a=10$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	49
Figure 6.13:	Example of best fitness data for $N_1=N_2=40$ using “saheart” dataset for $N_g=1000$ and $N_a=5$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm	50
Figure 6.14:	Example of best fitness data for $N_1=N_2=40$ using “saheart” dataset for $N_g=1000$ and $N_a=5$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	50
Figure 6.15:	Example of best fitness data for $N_1=N_2=40$ using “saheart” dataset for $N_g=1000$ and $N_a=10$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm	51

Figure 6.16:	Example of best fitness data for $N_1=N_2=40$ using “saheart” dataset for $N_g=1000$ and $N_a=10$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	51
Figure 6.17:	Example of best fitness data for $N_1=N_2=20$ using “car” dataset for $N_g=500$ and $N_a=5$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	52
Figure 6.18:	Example of best fitness data for $N_1=N_2=20$ using “car” dataset for $N_g=500$ and $N_a=10$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	52
Figure 6.19:	Example of best fitness data for $N_1=N_2=40$ using “car” dataset for $N_g=1000$ and $N_a=5$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	54
Figure 6.20:	Example of best fitness data for $N_1=N_2=40$ using “car” dataset for $N_g=1000$ and $N_a=10$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus FH-GBML algorithm	54
Figure 6.21:	Example of best fitness data for $N_1=N_2=20$ using “plastic” dataset for $N_g=500$ and $N_a=5$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus GFS-RB-MF algorithm	55
Figure 6.22:	Example of best fitness data for $N_1=N_2=20$ using “plastic” dataset for $N_g=500$ and $N_a=10$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus GFS-RB-MF algorithm	55
Figure 6.23:	Example of best fitness data for $N_1=N_2=40$ using “plastic” dataset for $N_g=1000$ and $N_a=5$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus GFS-RB-MF algorithm	56
Figure 6.24:	Example of best fitness data for $N_1=N_2=40$ using “plastic” dataset for $N_g=1000$ and $N_a=10$ showing the relative effect of different best fuzzy rules exchange policies on the developed algorithm versus GFS-RB-MF algorithm	56

Figure 6.25:	10 selected tuples from “plastic” real-world dataset showing the minimum (min) and maximum (max) for each attribute range as a sample case study	61
Figure 6.26:	Fuzzy sets for each dataset attribute as a sample case study	62
Figure 6.27:	Outer GA population composed of 2 chromosomes showing their structures as a sample case study	62
Figure 6.28:	Outer GA population composed of 2 chromosomes showing their assumed initial random values as a sample case study	63
Figure 6.29:	Inner GA population composed of 2 chromosomes showing their assumed initial random values as a sample case study	64
Figure 6.30:	The structure of PFGA that accepts a “plastic” real-world distributed datasets as a sample case study	66

List of Abbreviations

ACO	: Ant Colony Optimization
AdaBoost	: Adaptive Boosting
AI	: Artificial Intelligence
AIS	: Artificial Immune System
ANN	: Artificial Neural Network
ARM	: Association Rules Mining
BOAT	: Bootstrapped Optimistic Algorithm for Tree construction
CART	: Classification And Regression Trees
CBR	: Case-Based Reasoning
CI	: Computational Intelligence
CoA	: Center of Area
CoG	: Center of Gravity
CUDA	: Compute Unified Device Architecture
DDM	: Distributed Data Mining
DHT	: Distributed Hash Table
DM	: Data Mining
EC	: Evolutionary Computing
EM	: Expectation-Maximization
FAM	: Fuzzy Associative Memory
FGA	: Fuzzy-Genetic Algorithm
FH-GBML	: Fuzzy Hybrid Genetic-Based Machine Learning
FLS	: Fuzzy Logic System
FS	: Fuzzy Set
GA	: Genetic Algorithm
GFS-RB-MF	: Genetic-base Fuzzy Rule Base Construction and Membership Function tuning
GPU	: Graphics Processing Unit
KB	: Knowledge Base
KDD	: Knowledge Discovery from Data
KEEL	: Knowledge Extraction Based on Evolutionary Learning