

AIN SHAMS UNIVERSITY
Faculty of Computer
& Information Science
Computer Science Department



A HYBRID GENETIC ALGORITHM- DECISION TREE METHOD FOR KNOWLEDGE DISCOVERY

A Thesis

Submitted to Computer Science Department,
Faculty of Computer & Information Sciences,
Ain Shams University

In partial fulfillment of the requirements for
Master of Science Degree

By

Abeer Mahmoud Mahmoud

B.Sc. in Computer Science, 2000.
Demonstrator, Computer Science Department,
Faculty of Computer & Information Sciences,
Ain Shams University, Cairo, Egypt.

Under Supervision of

Prof. Dr. Abdel-Badeeh M. Salem

Professor of Computer Science, Computer Science
Department, Faculty of Computer & Information
Sciences, Ain shams University, Cairo, Egypt.

Dr. Khaled Ahmed Nagaty

Lecturer, Computer Science Department,
Faculty of Computer & Information Sciences,
Ain shams University, Cairo, Egypt.

April 2004

Acknowledgement

First and foremost, I am implementing my deep thanks to Allah for giving me the opportunity and the strength to accomplish this work,

I'm dedicating a special thanks to our dean Prof.Dr. Mohamed Saied Abdel Wahab for his valuable and helpful support.

I would like to express my deep appreciation to Prof. Dr. Abdel-badeeh Mohammed Salem for his full supervision of this work and to his planning, on going advises, guidance and for his comprehensive help in the interpretation of the outcome results by valuable suggestions in writing this thesis.

I am indebted to Dr. Khaled Ahmed Nagaty for his help in reviewing the work presented in this thesis.

Finally, I would like to dedicate my deep thanks to Prof. Dr. Ali Elnaiem my fiancé Mr. Asraf El Asmarwe My Family and my colleague friend, Mr. Emad Moneer those were encouraging me strongly during the execution of this work,

Publications

- ✓ Abdel-Badeeh M.Salem and Abeer M.Mahmoud, "*A Hybrid Genetic Algorithm-Decision Tree Classifier*", Proceedings of the 3rd International Conference on New Trends in Intelligent Information Processing and Web Mining, Zakopane, Poland, PP 221-232, June 2-5, 2003.
- ✓ Abdel-Badeeh M.Salem and Abeer M.Mahmoud, "*A Hybrid Genetic Algorithm- Decision Tree Classifier*", Journal of Intelligent Computing & Information Science, Cairo, Egypt, Volume 2, Number 2, PP 1-12, July 2002.
- ✓ Abdel-Badeeh M.Salem, Khaled Nagaty and Abeer M.Mahmoud, "*Genetic Algorithm Based Classifier for Breast Cancer Disease* ", Proceedings of the 9th International Conference on Soft Computing, Mendel, Brno, Czech Republic, PP 142-147, June 4-6, 2003.
- ✓ Abdel-Badeeh M.Salem and Abeer M.Mahmoud, "*Applying The Genetic Algorithms Approach for Data Mining Classification Task* ", Accepted for publication IFIP WG12.6, First IFIP Conference on Artificial Intelligence Applications and Innovations, Toulouse France, August, 22-27, 2004.

Abstract

Knowledge discovery is a multidisciplinary field. It includes database, visualization, statistics, machine learning and expert systems. Knowledge discovery process consists of six stages: data selection, cleaning, enrichment, coding, data mining and reporting. Data mining stage is the process of discovering useful patterns in large data sets. There are various mining techniques used for different purpose such as query tools, statistical techniques, online analytical processing (OLAP), case-based learning, decision trees, association rules, neural networks and genetic algorithms. Data mining is supported by hosting models or tasks such as: clustering, regression, summarization and classification models.

Classification is an important data-mining task that has a wide range of applications; one of them is medical diagnosis. The goal of classification is to build a model that is used to assign class labels to a database of testing records, where the values of the predictor attributes are known but the value of the class label is unknown. A variety of classification algorithms have been used in the literature. These algorithms can be divided into four main categories, which are decision tree based classification algorithms, neural network based classification algorithms, statistical based algorithms and Bayesian learning based algorithms.

The main objective of this study was to explore a new method integrating genetic algorithms and decision tree approaches, for data mining classification task.

Decision trees' learning is one of the most widely used and practical methods for data mining classification task. It is a method for approximating discrete-valued functions that

is robust to noisy data and capable of learning disjunctive expressions. Genetic algorithms provide an approach to learning that is based loosely on simulated evolution. The search for an appropriate hypothesis begins with a population of initial hypotheses. Members of the current population give rise to the next generation population by means of operations such as evaluation by fitness, selection, mutation and crossover.

This research demonstrates the usefulness of applying genetic algorithms approach in improving classification rates over the well known decision tree algorithm C4.5 (Quinlan, 1993). The study presents a new approach for developing two classifiers based on algorithm C4.5. The first classifier (RFC4.5) uses the RainForest framework database access method and replacing C4.5 pruning algorithm with a simple pruning algorithm. The second classifier (GARFC4.5) uses genetic algorithms approach. The two developed classifiers have been applied to large medical database for thrombosis diseases of 20MB size. The results show that RFC4.5 classifier with the simple pruning algorithm improves the classification rate from 81% to 93% over traditional C4.5. Also, adding genetic algorithms approach, GARFC4.5 classifier enhances the classification accuracy from 81% and 93% to 94% over traditional C4.5 and RFC4.5 classifiers respectively.

Moreover the study includes the application of our developed GARFC4.5 classifier on another database for breast cancer disease characterized by numerical attributes. Also a comparison have been done between our developed classifier and thirty-three classification algorithms, based on different learning methodologies, published recently by (Lim et al, 2000). The results show that GARFC4.5 classifier gives a reasonable classification rates comparing to those algorithms.

Table of Contents

Acknowledgment.	II
Publications.....	III
Abstract.....	IV
Table of Contents	VI
List of Figures.....	X
List of Tables.....	XII

CHAPTER	Page
1- Introduction.....	2
1.1 Problem Overview.....	2
1.2 Thesis Objectives.....	4
1.3 Thesis Organization.....	5
2- Knowledge Discovery and Data Mining.....	9
2.1 Introduction.....	9
2.2 Knowledge Discovery	10
2.2.1 Knowledge Discovery Process.....	11
2.2.2 Knowledge Discovery Systems.....	13
2.3 Data Mining Process	17
2.3.1 Data Mining Tasks.....	18
2.3.2 Data Mining Techniques	19
2.4 Summary.....	23
3- Classification Algorithms.....	24
3.1 Introduction.....	24
3.2 Decision Tree Based Algorithms.....	24
3.3 Neural Networks Based Algorithms.....	32
3.4 Statistical Based Algorithms.....	34
3.5 Data Mining Algorithms Selection.....	36

3.6	Summary.....	40
4-	Decision Trees Classification Algorithms...	41
4.1	Introduction	41
4.2	Decision Tree Representation.....	42
4.3	Appropriate Problems for Decision Tree Learning.....	43
4.4	ID3 Algorithm.....	45
4.4.1	Which Attribute is The Best Classifier?.....	47
4.4.2	An Illustrative Example	48
4.5	C4.5 Algorithm.....	49
4.5.1	Avoiding Overfitting the Data.....	50
4.5.2	Continuous-Valued Attributes.....	53
4.5.3	Alternative Measures for Selecting Attribute.....	54
4.5.4	Handling Training Set with Missing Attribute Values.....	56
4.5.5	C4.5 Algorithm Notations.....	57
4.6	RainForest-Fast Tree Construction Framework.....	60
4.6.1	RainForest Framework.....	60
4.6.2	RainForest & Main Memory.....	63
4.6.3	RainForest RF-Write Algorithm.....	63
4.7	Summary.....	65
5-	Evolutionary Algorithms.....	67
5.1	Introduction.....	67
5.2	Evolutionary Strategies.....	68
5.3	Evolutionary Programming.....	69
5.4	Genetic Algorithms.....	71

5.4.1	Genetic Algorithm Biological Background.....	72
5.4.2	Genetic Algorithm Motivations.....	74
5.4.3	Genetic Algorithm Requirements...	75
5.4.4	Genetic Algorithms Types	82
5.5	Summary.....	85
6-	A Hybrid Genetic Algorithm-Decision Tree Classifier.....	86
6.1	Introduction.....	86
6.2	The C4.5 Algorithm.....	87
6.3	The Proposed RFC4.5 Classifier.....	90
6.4	The Proposed GARFC4.5 Classifier.....	92
6.5	Experimental Results.....	96
6.5.1	Database Domain.....	96
6.5.2	Database Description.....	97
6.5.3	Database Preparation.....	101
6.5.4	Database Results.....	103
6.6	Conclusions.....	110
7-	Testing The Applicability of GARFC4.5 Classifier on Numerical	112
7.1	Introduction.....	112
7.2	Breast Cancer Database.....	113
7.3	Application of GARFC4.5 on Breast Cancer	116
7.4	Comparison of GARFC4.5 Classifier with Different Classification Algorithms.....	119
7.4.1	Decision Tree Based Algorithms...	119
7.4.2	Statistical Based Algorithms.....	122
7.4.3	Neural Networks Based Algorithms..	124
7.5	Conclusions.....	126

8- Summary, Conclusions and Future Work	127
8.1 Summary.....	127
8.2 Conclusions.....	129
8.3 Future Work.....	130
References.....	131
Appendix A: Object Oriented Classes of RFC4.5 Classifier	137
Appendix B: Object Oriented Classes of GARFC4.5 Classifier	138
Appendix C: Simple Manual for Using RFC4.5 and GARFC4.5 Classifiers.....	139

List of Figures

Figure 2.1	The steps of knowledge discovery process.....	12
Figure 2.2	Data mining Techniques for knowledge discovery.....	20
Figure 3.1	Different learning algorithms compared with different types of tasks.....	37
Figure 3.2	Selection of data mining algorithm	39
Figure 4.1	Decision tree for predicting a tennis game	43
Figure 4.2	Pseudo code of ID3 algorithm.....	46
Figure 4.3	Pseudo-code of the C4.5 algorithm.....	59
Figure 4.4	Tree induction schema and refinement.....	61
Figure 5.1	Problem solution using genetic algorithms	72
Figure 5.2	DNA Structure.....	73
Figure 5.3	Pseudo code for genetic algorithm.....	76
Figure 5.4	Chromosome representations.....	77
Figure 5.5	Mutation operator for a string.....	78
Figure 5.6	Crossover operator example.....	79
Figure 5.7	Crossover and mutation methods.....	80
Figure 5.8	Genetic algorithm hierarchy types.....	82
Figure 6.1	Pseudo-code of the C4.5 algorithm.....	89
Figure 6.2	Simple tree-pruning algorithm.....	91
Figure 6.3	GARFC4.5 classifier initialization operation.....	93
Figure 6.4	GARFC4.5 classifier crossover operation	93
Figure 6.5	GARFC4.5 classifier mutation operation	94
Figure 6.6	Block diagram of GARFC4.5 classifier....	94
Figure 6.7	Pseudo code for object oriented GA.....	96

Figure 6.8	RFC4.5 classifier behaviors at different population size.....	106
Figure 6.9	RFC4.5 classifier behaviors at different sample size.....	106
Figure 6.10	RFC4.5 and GARFC4.5 classification rates at population size 10.....	107
Figure 6.11	RFC4.5 and GARFC4.5 classification rates at population size 50.....	107
Figure 6.12	RFC4.5 and GARFC4.5 classification rates at population size 100.....	107
Figure 6.13	GARFC4.5 classifier behaviors at different population size.....	108
Figure 6.14	GARFC4.5 classifier behaviors at different sample size	108
Figure 6.15	GARFC4.5 classifier run time at different population size.....	109
Figure 6.16	GARFC4.5 classifiers run time at different sample size	109
Figure 7.1	RFC4.5 and GARFC4.5 classification rates at population size 10	118
Figure 7.2	RFC4.5 and GARFC4.5 classification rates at population size 20.....	118
Figure 7.3	RFC4.5 and GARFC4.5 classification rates at population size 50	118
Figure A.1	Object oriented design of Feature class....	137
Figure A.2	Object oriented design of Record class.....	137
Figure A.3	Object oriented design of RecordSet class.	138
Figure A.4	Object oriented design of Node class....	138
Figure A.5	Object-oriented design of Decision Tree class.....	138

Figure B.1	Object-oriented design of Chromosome Class.....	139
Figure B.2	Object-oriented design of Population class	139
Figure C.1	The interface for the hybrid classifiers.....	140
Figure C.2	RFC4.5 classifier functions.....	141
Figure C.3	Example of drawn tree by the classifier...	142
Figure C.4	GARFC4.5 classifiers parameter setting...	143

List of Tables

Table 4.1	Training Examples for The Target Concept Playtennis.....	50
Table 4.2	RainForest Algorithms States and Processing Behavior.....	64
Table 6.1	Schemata for first database-table (TSUM_A.CSV).....	98
Table 6.2	Schemata for second database table (TSUM_B.CSV).....	99
Table 6.3	Schemata for third database table (TSUM_C.CSV).....	100
Table 6.4	Computational Results for RFC4.5 and GARFC4.5 Classifiers.....	105
Table 6.5	Average Classification Rate of C4.5, RFC4.5 and GARFC4.5 Classifiers on Thrombosis Database.....	105
Table 7.1	Schemata for The Breast Cancer Database...	114
Table 7.2	Breast Cancer Database Records	115
Table 7.3	Computational Results for Breast Cancer Database.....	117
Table 7.4	Results of Decision Tree Based Algorithms on Breast Cancer Database	122
Table 7.5	Results of Statistical Based Algorithms on Breast Cancer Database	124
Table 7.6	Results of Neural Networks Based Algorithms on Breast Cancer Database.....	125
Table 7.7	A Comparison of Prediction Accuracy of Different Learning Approaches on Breast Cancer Database.....	125

Chapter 1

A
H
Y
B
R
I
D

G
E
N
E
T
I
C

A
L
G
O
R
I
T
H
M

D
E
C
I
S
I
O
N

T
R
E
E

M
E
T
H
O
D

F
O
R

K
N
O
W
L
E
D
G
E

D
I
S
C
O
V
E
R
Y

Introduction

Chapter 1

Introduction

1.1 Problem Overview

We are living in information age. During the last couple of decades, we exercised an on going evolving growth for our capabilities in both collecting and generating informative data. Each second, thousands of new information records are being generated. Information became an important commodity. This information needs to be summarized and synthesized in order to support decision-making. There is an urgent need to make sense of large amounts of data.

Knowledge discovery is a multidisciplinary field because it exploits several research disciplines of artificial intelligence such as machine learning, pattern recognition, expert systems, knowledge acquisition, as well as mathematical disciplines such as statistics, theory of information, uncertainty processing and others. Knowledge discovery process consists of six stages: data selection, cleaning, enrichment, coding, data mining and reporting. Data mining stage is the phase of real discovery; it deals with discovery of hidden knowledge, unexpected pattern and new rules from large databases. There are various mining techniques used for different purpose such as query tools, statistical techniques, online analytical processing (Olap), case based learning (K-nearest neighbor), decision trees , association rules, neural networks and genetic algorithm.