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. Ashraf El Ashmawe, 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

List of Figures

List of Tables

Chapter 1

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.