Discovery of Data Models using Genetic Programming

W.J.L Nuwan Wijayaweera

109161N

Faculty of Information Technology

University of Moratuwa

October 2012

Discovery of Data Models using Genetic Programming

W.J.L Nuwan Wijayaweera

109161N

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of the Degree of MSc in Artificial Intelligence

October 2012

Declaration

I declare that this dissertation does not incorporate, without acknowledgment, any material previously submitted for a Degree or a Diploma in any University and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, to be made available for photocopying and for interlibrary loans, and for the title and summary to be made available to outside organization.

W.J.L.N Wijayaweera

Name of Student Signature of Student

Date:

Supervised by

Prof. Asoka S. Karunananda

Name of Supervisor(s) Signature of Supervisor(s)

Date:

Dedication

To My Grandmother

Acknowledgements

First and foremost I would like to express my sincere gratitude to my supervisor Prof. Asoka S. Karunananda for the continuous guidance and support. I would have been lost without his guidance, encouragement, sound advice, and good ideas. I am grateful to all of those who supported me in any respect during this project.

Lastly, I would like to thank my grandmother, my family and all my friends and loved ones for all their love and encouragement.

W.J.L.N Wijayaweera

October 2012

Abstract

The field of Genetic Programming in Artificial Intelligence strives to get a computer to solve a problem without explicitly coding a solution by a programmer. Genetic Programming is a relatively new technology, which comes under automatic programming. After the initial work by John R. Koza in genetic programming, many researches have been done to discover data models in various datasets. These works have been rather domain specific and little attention have been given to develop generic framework for modeling and experimenting with genetic programming solutions for real world problems.

A project has been launched to develop a visual environment to design and experiment with genetic programming solutions for real world problems. It is named as GPVLab to mean its ability to facilitate the discovery of data models for real world problems through a wide range of experiments. GPVLab takes any numerical dataset as the input. Also the user should select the reference column before the main process. If user has not specified a column, the system will automatically take the last column as the reference column. This system has two possible ways to feed datasets into the system. There is an inbuilt facility to manually enter all the data. Furthermore this system facilitates data loading from comma separated value (*.csv) files. Output of this system is an evaluable expression. Users can initiate the main process by feeding data and selecting the reference column. Then the system runs the genetic programming process by generating populations of expressions and evaluating them to find their fitness. Finally the system determines the best fit model for the dataset. If the system has found a perfect solution or the maximum number of generations has exceeded, then the system stops the process and output the best so far model. This system outputs an evaluable expression as the data model in reverse polish notation (RPN). If the resultant expression only contains basic functions, GPVLab automatically converts the resultant evaluable expression in RPN into more human readable infix notation. System is developed for any person who needs to discover a model out of a collected numerical dataset. Advanced users with the knowledge of Genetic Algorithms or Genetic Programming can use advanced settings for better results. Nevertheless the default settings will work for most of the problems. Knowledge about Genetic Programming is not a necessity.

The system has been developed using C# language with .NET framework 4.0. GPVLab also extends the AForge.NET framework to accommodate data with arbitrary number of attributes and to remove noise in data. The system has an option to use two function sets for discovering process, namely basic and extended. Basic function set contains operators such as addition, substations, multiplication and division. Extended faction set has the ability to generate models consisting of square root (sqrt), sine (sin), cosine (cos), logarithms (ln) and exponential (exp) in addition to the basic operators. Upon completion of the discovering process, the system immediately allows the user to evaluate the model by providing required parameters. The system has the facility to save resultant models and access and evaluate them via library as required. Model library is developed using SQL compact Edition database, which does not require SQL Server instance to operate. Hence, this software is highly portable and can be installed or run in any computer with .NET Framework 4.0 installed.

GPVLab has been compared with WEKA as the main evaluation. A real world noisy dataset with eight columns has been used as the main input dataset. This main experiment has proved that the error rate of the solution generated by WEKA falls between -93.74% and 52.00% but the error rate of the solution generated by GPVLab falls between -24.15% and 24.51%. Further GPVLab has successfully discovered data models in simple datasets including square root of a number, addition of three numbers and a dataset with ten columns which has a known data model. All these solutions were achieved in less than 150 generations. The experiment of finding the square root function has been done using the extended function set and it directly provided the answer using ‘sqrt’ function within the first generation. Experiments performed by tweaking advanced settings showed that all the required facilities are there in GPVLab to experiment with genetic programming problems. Furthermore the results obtained through users with no knowledge about genetic algorithms or genetic programming, proved that this can be a really good tool for the researches in non technical fields as well. GPVLab has achieved all the objectives of this project. According to the main evaluation it is evident that GPVLab can generate solutions which provide better results in 56% of the time. It is concluded that GPVLab can be used to model genetic programming application very conveniently.

Contents

List of Figures

Page

List of Tables

Page