GPTIPS:An Open Source Genetic Programming Toolbox For Multigene Symbolic Regression

Dominic P. Searson, David E. Leahy and Mark J. Willis

***Abstract*— In this contribution GPTIPS, a free, open source MATLAB toolbox for performing symbolic regression by genetic programming (GP) is introduced. GPTIPS is specifically designed to evolve mathematical models of predictor response data that are "multigene" in nature, i.e. linear combinations of low order non-linear transformations of the input variables. The functionality of GPTIPS is demonstrated by using it to generate an accurate, compact QSAR (quantitative structure activity relationship) model of existing toxicity data in order to predict the toxicity of chemical compounds. It is shown that the low-order "multigene" GP methods implemented by GPTIPS can provide a useful alternative, as well as a complementary approach, to currently accepted empirical modelling and data analysis techniques. GPTIPS and documentation is available for download at http://sites.google.com/site/gptips4matlab/.**

***Index Terms*— genetic programming, symbolic regression, QSAR, toxicity, *T. pyriformis*.**

I. Introduction

Genetic programming [1] is a biologically inspired machine learning method that evolves computer programs to perform a task. It does this by randomly generating a population of computer programs (represented by tree structures) and then mutating and crossing over the best performing trees to create a new population. This process is iterated until the population contains programs that (hopefully) solve the task well.

When the task is building an empirical mathematical model of data acquired from a process or system, the GP is often known as symbolic regression. Unlike traditional regression analysis (in which the user must specify the structure of the model), GP automatically evolves both the structure and the parameters of the mathematical model. Symbolic regression has had both successful academic [2] and industrial applications [3].

The purpose of this paper is to introduce a free open source MATLAB toolbox called GPTIPS [4] that was written for the specific purpose of performing symbolic regression. GPTIPS employs a unique type of symbolic regression called multigene symbolic regression [5], [6] that evolves linear combinations of non-linear transformations of the input variables. When the transformations are forced to be low order (by restricting the GP tree depth) this, in contrast to "standard" symbolic regression, allows the evolution of accurate, relatively compact mathematical models of predictor – response (input – output) data sets, even when there are a large number of input variables. Hence, the authors believe that GPTIPS provides a useful, free and complementary alternative to current data analysis techniques and has a wide domain of applicability.

This paper is structured as follows. Section II provides a brief overview of the multigene low order GP approach that GPTIPS implements. Next, in section III, some of the features of GPTIPS are described. In sections IV -VII, the capabilities of GPTIPS are demonstrated by using it to evolve an accurate, relatively compact mathematical model to predict the toxicity of chemical compounds using an data set from the literature containing over 1000 compounds along with measured toxicity values. Finally, in section VIII we provide some concluding remarks. The following material assumes a basic familiarity with GP. If this is not the case then an excellent, free to download introduction and review of the literature is provided by [7].

II. Multigene symbolic regression

Typically, symbolic regression is performed by using GP to evolve a population of trees, each of which encodes a mathematical equation that predicts a ($N \times 1$) vector of outputs $\mathbf{y}$ using a corresponding ($N \times M$) matrix of inputs $\mathbf{X}$ where N is the number of observations of the response variable and M is the number of input (predictor) variables. i.e. the ith column of $\mathbf{X}$ comprises the N input values for the ith input variable and may be designated as the input variable x_i.

In contrast, in multigene symbolic regression each symbolic model (and each member of the GP population) is a weighted linear combination of the outputs from a number of GP trees, where each tree may be considered to be a "gene". For example, the multigene model shown in Fig. 1 predicts an output variable using input variables x_1, x_2 and x_3.

This model structure contains non-linear terms (e.g. the hyperbolic tangent) but is linear in the parameters with respect to the coefficients d_0, d_1 and d_2. In practice, the user specifies the maximum number of genes $G_{\max}$ a model is allowed to have and the maximum tree depth $D_{\max}$ any gene may have and therefore can exert control over the maximum complexity of the evolved models. In particular, we have found that enforcing stringent tree depth restrictions (i.e. maximum depths of 4 or 5 nodes) often allows the evolution of relatively compact models that are linear combinations of

Manuscript received December 2, 2009.

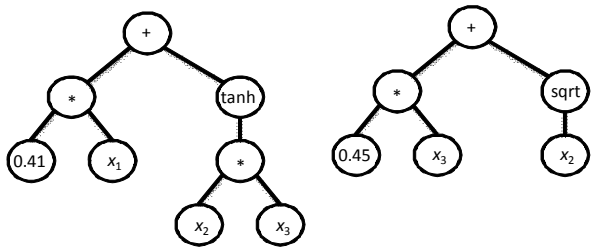
Dominic Searson is with the Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne, U.K

David Leahy is with the Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne, U.K

Mark Willis is with the School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne, U.K (e-mail: mark.willis@ncl.ac.uk).

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

low order non-linear transformations of the input variables.



$$y = d_0 + d_1(0.41x_1 + \tanh(x_2x_3) + d_2(0.45x_3 + \text{sqrt}(x_2))$$

Fig. 1. Example of a multigene symbolic model.

For each model, the linear coefficients are estimated from the training data using ordinary least squares techniques. Hence, multigene GP combines the power of classical linear regression with the ability to capture non-linear behaviour without needing to pre-specify the structure of the non-linear model. In [5] it was shown that multigene symbolic regression can be more accurate and computationally efficient than the standard GP approach for symbolic regression and [6] demonstrated that the multigene approach could be successfully embedded within a non-linear partial least squares algorithm.

In GPTIPS, the initial population is constructed by creating individuals that contain randomly generated GP trees with between 1 and $G_{\max}$ genes. During a GPTIPS run, genes are acquired and deleted using a tree crossover operator called two point high level crossover. This allows the exchange of genes between individuals and it is used in addition to the "standard" GP recombination operators. If the *i*th gene in an individual is labelled G_i then a two point high level crossover is performed as in the following example. Here, the first parent individual contains the genes (G_1 G_2 G_3) and the second contains the genes (G_4 G_5 G_6 G_7) where $G_{\max} = 5$. Two randomly selected crossover points are created for each individual. The genes enclosed by the crossover points are denoted by < … >.

(G_1 < $\boldsymbol{G_2}$ > G_3) (G_4 < $\boldsymbol{G_5}$ $\boldsymbol{G_6}$ $\boldsymbol{G_7}$ >)

The genes enclosed by the crossover points are then exchanged resulting in the two new individuals below.

(G_1 G_5 G_6 G_7 G_3) (G_4 G_2)

Two point high level crossover allows the acquisition of new genes for both individuals but also allows genes to be removed. If an exchange of genes results in an individual containing more genes than $G_{\max}$ then genes are randomly selected and deleted until the individual contains $G_{\max}$ genes.

In GPTIPS, standard GP subtree crossover is referred to as low level crossover. In this case, a gene is selected randomly from each parent individual, standard subtree crossover is performed and the resulting trees replace the parent trees in the otherwise unaltered individual in the next generation. GPTIPS also provides several methods of mutating trees.

The user can set the relative probabilities of each of these recombinative processes. These processes are grouped into categories called events. The user can then specify the probability of crossover events, direct reproduction events and mutation events. These must sum to one. The user can also specify the probabilities of event subtypes, e.g. the probability of a two point high level crossover taking place once a crossover event has been selected, or the probability of a subtree mutation once a mutation event has been selected. However, GPTIPS provides default values for each of these probabilities so the user does not need to explicitly set them.

III. GPTIPS FEATURES

GPTIPS is a predominantly command line driven open source toolbox that requires only a basic working knowledge of MATLAB. A run is configured by a simple configuration M file and there are a number of command line functions to facilitate post-run analyses of the results. Whilst not an exhaustive list, GPTIPS currently contains the following configurable GP features: tournament selection & plain lexicographic tournament selection [8], elitism, three different tree building methods (full, grow and ramped half and half) and six different mutation operators: (1) subtree mutation (2) mutation of constants using an additive Gaussian perturbation (3) substitution of a randomly selected input node with another randomly selected input node (4) set a randomly selected constant to zero (5) substitute a randomly selected constant with another randomly generated constant (6) set a randomly selected constant to one. In addition, GPTIPS can, without modification in the majority of cases, use nearly any built in MATLAB function as part of the function set for a run. The user can also write bespoke function node M files and fitness functions.

In addition, GPTIPS has a number of features that are specifically aimed at the creation, analysis and simplification of multigene symbolic regression models. These include: (1) use of a 'holdout' validation set during training to mitigate the effects of overfitting (2) graphical display of the results of symbolic regression for any multigene model in the final population (3) mathematical simplification of any model (4) conversion to LaTex format of any model (5) conversion to PNG (portable network graphics) file of the simplified equation of any model (6) conversion of any model to standalone M file for use outside GPTIPS (7) graphical display of the statistical significance of each gene in a model (8) functions to reduce the complexity of any model using "gene knockouts" to explore the trade off of model accuracy against complexity (9) graphical population browser to explore the trade off surface of complexity/accuracy (10) graphical input frequency analysis of individual models or of a user specified fraction of the population to facilitate the identification of input variables that are relevant to the output.

The Symbolic Math toolbox (a commercial toolbox available from the vendors of MATLAB) is required for the majority of the post run simplification and model conversion features and the Statistics Toolbox is required for the display of gene statistical significance. The core functionality of GPTIPS and the ability to evolve multigene models does not, however, require any specific toolboxes. In the following section, some of these features will be demonstrated using a real world modelling example.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

IV. Evolution of a Predictive Model of Aqueous Chemical Toxicity Using GPTIPS

QSAR (Quantitative Structure Activity Relationships) is a well established technique for deriving structure property relationships for chemical compounds that can be used to predict the properties of novel chemical structures. Chemical compounds can be represented by a large number of computed numerical values, called "descriptors", each of which in some way characterises the structure or behaviour of the compound. The idea of QSAR is to build empirical or semi-empirical models that relate the descriptors of a compound to some physical, chemical or biological property. A number of software packages are available to compute descriptor values for compounds with a known structure. Many of these are commercial products (e.g. DRAGON) but there are also free/open source packages (e.g. the Chemical Descriptors Library (CDL; [9]) and the Chemical Development Kit (CDK; [10]).

A QSAR modelling scenario involves a data set of known chemical compounds and a measured endpoint for each compound. The measured endpoint is the property of interest. Typical properties of interest are those related to pharmaceutical drug development. These include biological activities representing the ability of a drug candidate to perform its desired function (e.g. IC50, the concentration of a compound required to inhibit a particular biological or biochemical function by half) and the ADME properties (adsorption, distribution, metabolism and excretion) which characterise the behaviour of a of a pharmaceutical drug compound within the organism.

The prediction of chemical toxicity is another chemical property that is of vital importance in both pharmaceutical drug development and managing the environmental risk of chemical compounds. In the latter case there are legal regulatory structures (e.g. the REACH regulations in the European Union - EC 1907/2006) that specify that QSAR models should play a part in managing this risk in order to reduce the costs of experimental toxicity measurement. Hence, the development of effective QSAR modelling methods continues to present a very real and relevant challenge.

There are a number of strategies & protocols for experimentally evaluating chemical toxicity. One commonly accepted method is the measurement of the growth inhibition of ciliated protozoan *T. pyriformis* [11]. There are freely available aquatic toxicity data for more than 1000 compounds, due to the efforts of Schultz and colleagues (e.g. see [12]). The authors of [11] have used this to compile a data set of 1093 unique compounds and have developed a number of predictive QSAR models using various descriptor packages and modelling methodologies. Here, the use of GPTIPS to evolve a predictive model of chemical toxicity using this data set is demonstrated (using the descriptors from the commercial DRAGON package) and the results compared with those published in [11].

V. Data

The *T. pyriformis* toxicity values (i.e. the response **y** data) are measured as the logarithm of the 50% growth inhibition concentration $\log(IGC50^{-1})$. The data available for training QSAR models contains 644 compounds and another 449 compounds are used an external test/validation data set to verify the predictive ability of the models. For each compound 1664 DRAGON descriptor values are used as the predictor data (i.e. the input **X** data contains 1664 input variables) - compound structures, toxicity and descriptor values are available from the EU CADASTER website at http://www.cadaster.eu/node/65. To mitigate against the effects of overfitting, 128 compounds (approximately 20%) in the training data set were randomly selected for use as a holdout validation data set leaving the training data containing 516 compounds. In GPTIPS, holdout validation is performed as follows: at the end of each generation, the "best" individual (as evaluated on the training data) is then evaluated on the holdout validation set. The individual that performs best on the holdout set (over the course of the run) is stored and may be accessed after the run.

VI. GPTIPS Run Settings

A GPTIPS run with the following settings was performed: Population size = 500, Number of generations = 500, Tournament size = 12 (with lexicographic selection pressure), $D_{max} = 4$, $G_{max} = 8$, Elitism = 0.01 % of population, function node set = {plus, minus, times, tanh, sin}, terminal node set = {1664 DRAGON descriptors $x_1 - x_{1664}$, ephemeral random constants in the range [-10 10]}. The default GPTIPS multigene symbolic regression function was used in order to minimise the root mean squared prediction error on the training data.

The following (default) recombination operator event probabilities were used: Crossover events = 0.85, mutation events = 0.1, direct reproduction = 0.05. The following sub-event probabilities were used: high level crossover = 0.2, low level crossover = 0.8, subtree mutation = 0.9, replace input terminal with another random terminal = 0.05, Gaussian perturbation of randomly selected constant = 0.05 (with standard deviation of Gaussian = 0.1). These settings are not considered 'optimal' in any sense but were based on experience with modelling other data sets of similar size. The run took approximately 15 minutes on a PC with a dual core processor running at 2.2GHz with 3.5GB of RAM.

VII. Results

The model that performed best on the holdout validation data was chosen. This model has coefficients of determination (i.e. proportion of the variation in the response explained by the model) of R^2(training) = 0.83, R^2(holdout) = 0.78 and R^2(test) = 0.78. In [11] the results are reported in terms of MAE (mean absolute error) for two test tests referred to in the paper as Validation set 1 (339 compounds) and Validation set 2 (110 compounds) that comprise the whole test set used here. In terms of MAE, the evolved GPTIPS model has MAE(training) = 0.3292, MAE(holdout) = 0.3573 and MAE(test) = 0.3518.

The authors of [11] report the results of a number of individual models, built using various descriptor packages and modelling techniques. Some of these models consider the "applicability domain" (AD) of the compounds (i.e. whether the compounds lie in the region of descriptor space deemed to

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$y = -2.092 - 0.7548\,x_{911} + 0.7548\,x_{1558} - 0.8997\,\tanh(\tanh(x_{1426})) + 0.09443\,(x_{911} - x_{1558})\,x_{654} - 0.1481\,x_{1552} + 0.1481\,x_{391} - 0.2489\,x_{1429} - 0.2489\,\sin(x_{967} - x_{709}) + 0.7143\,x_{1245} - 0.5978\,x_{1662} - 0.5978\,\tanh(x_{1429} + x_{525}) + 0.7802\,x_{1426} + 0.7802\,x_{1563}$$

Fig 2. Graphical rendering of evolved symbolic *T. pyriformis* toxicity model.

be suitable for generating a prediction) whereas others do not employ AD considerations. In general, models that consider AD give more accurate predictions but only the results of the non AD models using the DRAGON descriptors are repeated here. The first DRAGON descriptor based model is a support vector machine (SVM; [13]) regression that yields MAE(Validation set 1) = 0.37 and MAE(Validation set 2) = 0.42. This corresponds to an MAE(test) = 0.38. The second DRAGON based model is a *k*- nearest neighbour (*k*-NN) approach that achieves MAE(Validation set 1) = 0.29, MAE(Validation set 2) = 0.43 corresponding to MAE(test) = 0.32. Hence it can be seen that the evolved GPTIPS model lies between the SVM and the *k*-NN approaches, i.e. GPTIPS can achieve predictive performance of the order of the current state of the art empirical modelling methodologies.

GPTIPS was used to mathematically simplify and export the evolved model as a PNG graphics file. This is shown in Fig 2.

It can be seen that the evolved model is reasonably compact, consists of both linear terms and low order non-linear transformations of the inputs and has selected a small number of descriptors from the 1664 available.

VIII. CONCLUSIONS

In this article we have introduced the multigene symbolic regression capabilities of GPTIPS and demonstrated it with an application in which a predictive symbolic QSAR model of *T. pyriformis* aqueous toxicity was evolved. It was demonstrated that the evolved model is compact and offers similar high performance to recently published QSAR models of the same data. The point of this article is not to assert that multigene symbolic regression (using low order non-linear transforms of the inputs) is better or worse than other methods, but that it is an alternative and complementary approach to existing empirical modelling and data analysis techniques. It is also an approach that is facilitated by the free GPTIPS toolbox for MATLAB, a program that is used widely in academia and industry.

REFERENCES

[1] Koza JR. Genetic programming: on the programming of computers by means of natural selection. The MIT Press, USA, 1992.
[2] Alfaro-Cid E, Esparcia-Alcázar AI, Moya P, Femenia-Ferrer B, Sharman K & Merelo JJ. Modeling pheromone dispensers using genetic programming. In Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Vol. 5484/2009, 635-644, 2009.
[3] Kordon, A.K. Future Trends in Soft Computing Industrial Applications, Proceedings of the 2006 IEEE Congress on Evolutionary Computation, 7854-7861, 2006
[4] Searson, D. GPTIPS: Genetic Programming & Symbolic Regression for MATLAB, http://gptips.sourceforge.net, 2009.
[5] Hinchliffe MP, Willis MJ, Hiden H, Tham MT, McKay B & Barton, GW. Modelling chemical process systems using a multi-gene genetic programming algorithm. In Genetic Programming: Proceedings of the First Annual Conference (late breaking papers), 56-65. The MIT Press, USA, 1996.
[6] Searson DP, Willis MJ & Montague GA. Co-evolution of non-linear PLS model components, Journal of Chemometrics, 2, 592-603, 2007.
[7] Poli R, Langdon WB, and McPhee NF. A field guide to genetic programming. Published via http://lulu.com and freely available at http://www.gp-field-guide.org.uk, 2008.
[8] Luke S & Panait L. Lexicographic parsimony pressure. Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2002), 2002.
[9] Sykora VJ & Leahy DE. The Chemical Descriptor Library CDL: A Generic, Open Source Software Library for Chemical Informatics, J. Chem. Inf. Model., 48, 1931-1942, 2008.
[10] Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E & Willighagen E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. J. Chem. Inf. Comput. Sci., 43, 493 – 500, 2003.
[11] Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Oberg T, Dao P, Cherkasov A & Tetko IV. Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis . J. Chem. Inf. Model.,48, 766 -784, 2008.
[12] Schultz TW; Yarbrough JW & Woldemeskel M. Toxicity to Tetrahymena and abiotic thiol reactivity of aromatic isothiocyanates. Cell Biol. Toxicol.,21, 181–189, 2005.
[13] Vapnik, VN. The nature of statistical learning theory, second edition, Springer-Verlag, New York, 2000.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)