Guiding Function Set Selection in Genetic Programming based on Fitness Landscape Analysis

Nguyen Quang Uy
Faculty of Information Technology
Military Technical Academy
Hanoi, Vietnam
quanguyhn@gmail.com

Truong Cong Doan
Faculty of Information Technology
Hanoi Open University
Hanoi, Vietnam
truongcongdoan@gmail.com

Nguyen Xuan Hoai
IT Research and Development Center
Hanoi University
Hanoi, Vietnam
nxhoai@gmail.com

Michael O'Neill
Natural Computing Research & Application Group
University College Dublin
Belfield, Dublin 4, Ireland
m.oneil@ucd.ie

ABSTRACT

This paper attempts to provide a guideline for function set selection based on fitness landscape analysis. We used two well-known techniques, autocorrelation function and information content, to analysize the fitness landscape of each function set. We tested these methods on a large number of real-valued symbolic regression problems and the experimental results showed that there is a strong relationship between autocorrelation function value and the performance of a function set. Therefore, autocorrelation function can be used as a good indicator for selecting an appropriate function set for a problem.

Categories and Subject Descriptors

I.2.8 [**Problem Solving, Control Methods, and Search**]: Heuristic methods

General Terms

Algorithms

Keywords

Genetic Programming, Function Set, Fitness Landscape

1. INTRODUCTION

One of the early but important step in applying Genetic Programming (GP) to a problem is to select a set of functions and terminals. While the set of terminals usually arise naturally from the definition of the problem, the choice of the function set is much less obvious. In this paper, we propose a method for choosing a function set among the possible ones with a quantitative indicator that is based on examining the fitness landscape generated by each function set. This research aims to give some first answers to the following question: Is there any correlation between the characteristics of the fitness landscape generated by a function set and the performance of GP?

GECCO'13 Companion, July 6–10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

2. METHODS

This section presents two methods used to characterise fitness landscapes in this paper. The first method is autocorrelation function [4] and the second one is information content [3].

2.1 Autocorrelation Function

Using an autocorrelation function to study fitness landscapes was first proposed in [4]. For a given fitness landscape with f as the fitness function, a starting point s_0 is randomly selected. Using a mutation operator to create a neighbouring point s_1 of s_0. Repeat this process N times to get a random walk of $N+1$ steps $F = \{f(s_i)\}_{i=0}^{N}$. Then the autocorrelation function of this random walk is defined as follows:

$$\rho(h) = \frac{R(h)}{s_f^2} \tag{1}$$

where h is the distance between two points in the random walk and was set at 1 in this paper. s_f^2 is variance of the sequence and $R(h)$ is the autocovariance function of sequence F.

2.2 Information Content

The second method for characterising fitness landscapes used is based on the concept of information content [3]. A random walk of $N+1$ steps, $F = \{f(s_i)\}_{i=0}^{N}$ is first conducted. A sequence driven from F, $S(\epsilon) = s_1, s_2, ..., s_N$, to represent this random walk for each ϵ, is generated, where $s_i = \Psi(i, \epsilon)$, and

$$\Psi(i,\epsilon) = \begin{cases} -1 & \text{if } f_i - f_{i-1} < -\epsilon; \\ 0 & \text{if } |f_i - f_{i-1}| \leq \epsilon; \\ 1 & \text{if } f_i - f_{i-1} > \epsilon; \end{cases} \tag{2}$$

After that, the information content that characterises the

Table 1: Function Sets.

Sets	Functions
$F1$	+, -, *, /, sin, cos, exp, log, sqrt, sqr
$F2$	+, -, *, /, sin, cos, exp, log
$F3$	+, -, *, /, sin, cos, sqrt, sqr
$F4$	+, -, *, /, sin, cos
$F5$	+, -, *, /, exp, log
$F6$	+, -, *, /, sqrt, sqr
$F7$	+, -, *, /
$F8$	sin, cos, exp, log

ruggedness of a fitness landscape is defined as follows (ϵ was set at 1 in this paper):

$$H(\epsilon) = \sum_{p \neq q} P_{[pq]} log_6 P_{[pq]} \quad (3)$$

3. EXPERIMENTAL SETTINGS

To investigate the possible relationship between the fitness landscape generated by a function set and the performance of GP when using this function set, we examined the fitness landscapes of eight function sets for the symbolic regression problem. These function sets are detailed in Table 1. We tested GP using these function sets on twelve different real-valued target functions taken from [1].

We divided our experiments into two sets. The first set was to investigate the characteristics of the fitness landscape generated by the tested function sets. A random walk of 1000 steps were created using the subtree mutation for each function set on each problem. For each problem, 200 random walks were conducted, making a total of 200,000 fitness evaluations for each problem.

The second set of experiments aim to measure the impact of function sets on GP performance. The evolutionary parameter settings are similar to [2]. For each function set and each problem, 100 runs were conducted. The fitness function is the mean of absolute error. A classical performance metric, mean of best fitness, is used. The results are presented and discussed in the following section.

4. RESULTS AND DISCUSSION

We first calculated the Pearson's correlation coefficient of the information content value and the mean best fitness as well as of the autocorrelation function value and the mean best fitness. These coefficients are shown in Table 2. It can be seen from this table the (inverse) correlation of autocorrelation function value to the mean best fitness is often (in 8 out of 12 problems) stronger than the correlation of information content value to the mean best fitness. Therefore, from here, we shall only focus on analyzing the relationship between autocorrelation function value and the GP performance.

In order to highlight the correlation between autocorrelation function value and the mean best fitness we count the number of problems where the greatest value and the second greatest value of the autocorrelation function corresponding to the function sets that resulted in the best and second best GP performance. The results are that [1], on 7 out of 12 problems, the function set that has the greatest of autocorrelation value is also the function set that fostered the best GP performance. In the case that the best function set in terms of fitness landscape did not bring the best performance, it usually led to the second best performance and this situation happened in 4 out of 12 target functions. Overall, on 11 out of 12 problems (91.3%), the function set that generates the smoothest fitness landscape helped GP to achieve the best or second best performance.

[1]The results are not shown here due to limited space

Table 2: Correlation coefficient

Functions	Sextic	Septic	Notic	R1	R2	R3
Infor Cont	0.52	0.67	0.64	0.50	0.46	0.42
Auto Funcs	-0.40	-0.28	-0.55	-0.20	-0.74	-0.95
Functions	Ng-5	Ng-6	Ng-7	Kei-1	Kei-4	Kei-9
Infor Cont	0.29	0.41	0.42	0.29	0.09	-0.35
Auto Funcs	-0.40	-0.58	-0.46	0.45	-0.42	0.10

5. CONCLUSIONS AND FUTURE WORK

In this paper we study the fitness landscape of different function sets in Genetic Programming (GP). The results show that there is a strong correlation between autocorrelation function value and the performance of GP. Therefore, it suggests that the smoothness fitness landscape generated by a function set measured by the autocorrelation function could be used as a quantitative indicator to select a good function set among the set of candidates for a given problem.

Acknowledgment

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2012.04.

6. REFERENCES

[1] K. Krawiec and T. Pawlak. Locally geometric semantic crossover: a study on the roles of semantics and homology in recombination operators. *Genetic Programming and Evolvable Machines*, 14(1):31–63, 2013.

[2] N. Q. Uy, M. O'Neill, N. X. Hoai, B. McKay, and E. G. Lopez. Semantic similarity based crossover in GP: The case for real-valued function regression. In P. Collet, editor, *Evolution Artificielle, 9th International Conference*, Lecture Notes in Computer Science, pages 13–24, October 2009.

[3] V. K. Vassilev, T. C. Fogarty, and J. F. Miller. Information characteristics and the structure of landscapes. *Evolutionary Computation*, 8(1):31–60, Spring 2000.

[4] E. D. Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biological Cybernetics*, 63:325–336, 1990.