Information Sciences 436–437 (2018) 352–366

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Semantic tournament selection for genetic programming based on statistical analysis of error vectors

Thi Huong Chu[a], Quang Uy Nguyen[a,*], Michael O'Neill[b]

[a] Faculty of IT, Le Quy Don Technical University, Hanoi, Vietnam
[b] Natural Computing Research & Applications Group and Lero, School of Business, University College Dublin, Ireland

ARTICLE INFO

Article history:
Received 13 March 2017
Revised 12 January 2018
Accepted 13 January 2018

Keywords:
Genetic programming
Tournament selection
Statistical test
Code bloat
Semantics

ABSTRACT

The selection mechanism plays a very important role in the performance of Genetic Programming (GP). Among several selection techniques, tournament selection is often considered the most popular. Standard tournament selection randomly selects a set of individuals from the population and the individual with the best fitness value is chosen as the winner. However, an opportunity exists to enhance tournament selection as the standard approach ignores finer-grained semantics which can be collected during GP program execution. In the case of symbolic regression problems, the error vectors on the training fitness cases can be used in a more detailed quantitative comparison. In this paper we introduce the use of a statistical test into GP tournament selection that utilizes information from the individual's error vector, and three variants of the selection strategy are proposed. We tested these methods on twenty five regression problems and their noisy variants. The experimental results demonstrate the benefit of the proposed methods in reducing GP code growth and improving the generalisation behaviour of GP solutions when compared to standard tournament selection, a similar selection technique and a state of the art bloat control approach.

1. Introduction

Genetic Programming (GP) is a biologically inspired method of using a computer to evolve solutions, in the form of computer programs, for a problem [24,37]. To solve a problem using a GP system, a population of individuals is first initialised. The population is then evolved, under fitness based selection, through a number of generations by applying genetic operators. The evolutionary process terminates when a desired solution is found or when the maximum number of generations is exceeded.

There are several factors that can affect the performance of GP for a given problem. These factors include the size of the population, the fitness evaluation of individuals, the selection mechanisms for reproduction and the genetic operators for modifying individuals. Amongst these, selection plays a critical role in GP performance [4]. To date, there have been many selection schemes proposed [23] and the most widely used selection in GP is tournament selection [11].

Tournament selection compares the fitness values of sampled individuals. The individual with the best fitness is then selected as the winner. This implementation is simple and its effectiveness has been widely evidenced [11]. However, the

* Corresponding author.
E-mail addresses: huongktqs@lqdtu.edu.vn (T.H. Chu), quanguyhn@lqdtu.edu.vn (Q.U. Nguyen), m.oneill@ucd.ie (M. O'Neill).

https://doi.org/10.1016/j.ins.2018.01.030
0020-0255/

standard approach only uses the fitness value while ignoring information from the error vectors of individuals in all fitness cases. Consequently, some information that is potentially useful for GP search may be lost. Recent research has shown that significant benefit could be gained by using semantic information of GP individuals (e.g., [21,22,28,31,35]). The genetic search operators of crossover and mutation can be modified to improve the semantic locality of search [9,30,34]. In addition, the preservation of semantic diversity is a desirable feature of an evolving GP population to avoid local optima [5,12], thus, it is also attractive to examine whether using the error vectors of individuals on the fitness cases during selection can improve GP performance.

In our preliminary research [6], we have proposed two forms of semantic tournament selection that are based on statistical analysis of the error vectors of individuals. The experimental results on a set of GP benchmark problems showed the benefit of the proposed techniques [6]. In this paper, we extend this research with the main contributions of this paper being:

- We introduce the use of statistical analysis of GP error vectors to create novel forms of tournament selection. Based on a Wilcoxon signed rank test, three variants of tournament selection are proposed to exploit semantic diversity and to explore the potential of the approach to control program bloat.
- The performance of the selection strategies are examined on a large set of regression problems employing the original problems and noisy variants. We observe that the new selection techniques help to reduce the code growth and improve the generalization ability of the evolved solutions when compared to standard tournament selection and a state of the art method for controlling code bloat in GP.
- The simplicity of the design of the proposed selection strategies allows for further improvements. In this paper, the addition of a state of the art crossover operator is observed to further enhance performance.

In the next section, we present the background of the paper. Section 3 reviews the related work on improving tournament selection in GP. Three proposed tournament selection strategies are presented in Section 4. Section 5 presents the experimental settings adopted in the paper. Section 6 analyses and compares the performance of the proposed selection strategies with standard tournament selection. The approach is further enhanced through it's coupling to a state of the art crossover strategy in Section 7. Section 8 investigates the ability of the proposed techniques on noisy datasets. Finally, Section 9 concludes the paper and highlights some future work.

2. Background

This section presents some important concepts used in the proposed selection strategies, including the semantics of a GP individual, the error vector of an individual, and the Wilcoxon signed rank test.

In GP, it is common to define the semantics of a program simply as its behaviour with respect to a set of input values [27,31]. Formally, the semantics of a program is defined as follows:

Definition 2.1. Let $K = (k_1, k_2, \ldots k_N)$ be the fitness cases of the problem. The *program semantics* $S(P)$ of a program P is the vector of output values obtained by running P on all fitness cases.

$$S(P) = (P(k_1), P(k_2), \ldots, P(k_N)), \text{ for i} = 1, 2, \ldots, \text{N}.$$

This definition is valid for problems where a set of fitness cases is defined. The error vector of an individual is calculated by comparing the semantic vector with the target output of the problem. More precisely, the error vector of an individual is defined as:

Definition 2.2. Let $S = (s_1, s_2, \ldots s_N)$ be the semantics of an individual P and $Y = (y_1, y_2, \ldots y_N)$ be the target output of the problem on N fitness cases. The *error vector* $E(P)$ of a program P is a vector of N elements calculated as follows.

$E(P) = (|s_1 - y_1|, |s_2 - y_2|, \ldots, |s_N - y_N|)$.

In this study, the error vectors of individuals competing in a tournament are compared using a Wilcoxon signed rank test. The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples to assess whether their population mean ranks differ [20]. This test is used as an alternative to the paired Student's t-test when the population cannot be assumed to be normally distributed. Let N be the sample size of the test and $x_{1,i}$ and $x_{2,i}$ denote the ith pair sample. Let H_0: be the hypothesis that difference between the pairs follows a symmetric distribution around zero and H_1: be the hypothesis that difference between the pairs does not follow a symmetric distribution around zero. The test is performed as follows:

1. For $i = 1, \ldots, N$, calculate $|x_{2,i} - x_{1,i}|$, and $sgn(x_{2,i} - x_{1,i})$, where sgn is the sign function:

$$sgn(x) := \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases} \quad (1)$$

2. Exclude pairs with $|x_{2,i} - x_{1,i}| = 0$. Let N_r be the reduced sample size.
3. Order the remaining N_r pairs from smallest absolute difference to largest absolute difference, $|x_{2,i} - x_{1,i}|$.