Tournament Selection Based on Statistical Test in Genetic Programming

Thi Huong Chu[1], Quang Uy Nguyen[1(✉)], and Michael O'Neill[2]

[1] Faculty of IT, Le Quy Don Technical University, Hanoi, Vietnam
huongktqs@gmail.com, quanguyhn@gmail.com
[2] Natural Computing Research and Applications Group, UCD Business, University College Dublin, Dublin, Ireland
m.oneill@ucd.ie

Abstract. Selection plays a critical role in the performance of evolutionary algorithms. Tournament selection is often considered the most popular techniques among several selection methods. Standard tournament selection randomly selects several individuals from the population and the individual with the best fitness value is chosen as the winner. In the context of Genetic Programming, this approach ignores the error value on the fitness cases of the problem emphasising relative fitness quality rather than detailed quantitative comparison. Subsequently, potentially useful information from the error vector may be lost. In this paper, we introduce the use of a statistical test into selection that utilizes information from the individual's error vector. Two variants of tournament selection are proposed, and tested on Genetic Programming for symbolic regression problems. On the benchmark problems examined we observe a benefit of the proposed methods in reducing code growth and generalisation error.

Keywords: Genetic Programming · Tournament selection · Statistical test

1 Introduction

There are several factors that can effect the performance of Genetic Programming (GP) for given problems. These factors include the size of a population, the fitness evaluation of individuals, the selection mechanisms for reproduction, the encoding and genetic operations for modifying individuals. Amongst these, selection plays a critical role in GP performance [4]. To date, there have been many selection schemes proposed, and popular selection schemes in GP include fitness proportionate selection, ranking selection, and tournament selection [9]. Among these, the most widely used selection in GP is tournament selection [6].

Tournament selection is based on comparing 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 evidenced by a number of research [6]. However, the standard implementation used only fitness

J. Handl et al. (Eds.): PPSN XIV 2016, LNCS 9921, pp. 303–312, 2016.
DOI: 10.1007/978-3-319-45823-6_28

value while ignoring information from the error 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., [10,12]). Thus, it is attractive to examine whether using the error value of individuals on the fitness cases for selection can improve GP performance.

In this paper, the error vector of individuals is used in tournament selection. Two individuals are compared using a statistical test using their error vector. If the statistical test (using a Wilcoxon signed rank test in this case) shows that there is a significant difference, the individual with the better fitness is selected and tested against the others. This process is repeated for all individuals in the tournament sample with the winner selected based on the statistical test. We test the proposed selection technique on a set of benchmark regression problems, and observe that the proposed method helps to reduce GP code growth and generalisation error.

The remainder of this paper is organized as follows. In the next section, we briefly review the related work on improving tournament selection in GP. The two proposed tournament selection methods are presented in Sect. 3. Section 4 presents the experimental settings adopted, with the results presented and discussed in Sect. 5. Finally, Sect. 6 concludes the paper and highlights some future work.

2 Related Work

This section presents a brief review of previous research on tournament selection in GP. Tournament selection is the most popular selection operator in GP [17]. In standard tournament selection, a number of individuals (tournament size) are randomly selected from the population. These individuals are compared together and the winner (in terms of better fitness) is selected to go to the mating pool. This process is then repeated N times where N is the population size [4]. The advantage of tournament selection is that it allows the adjustment of the selection pressure by tuning the tournament size. Moreover, this method does not require a comparison of the fitness between all individuals that helps to save a large amount of processing time [19].

As the standard tournament selection consists of two steps: sampling and selecting. There is a large number of research focusing on different sampling and selecting strategies in tournament selection. Xie et al. [20] analysed the performance of no-replacement tournament selection in GP. In the no-replacement strategy, no individual can be sampled multiple times within the same tournament. Another problem in tournament selection is not-sampled problem, in which some individuals are not sampled at all if a too small tournament size is used. This problem was discussed by Xie et al. in [18]. Later, Gathercole et al. [7] analyzed the selection frequency of each individual and the likelihoods of not-selected and not-sampled individuals in tournament selection of different tournament sizes. Sokolov and Whitley proposed unbiased tournament selection [15] where all individuals have a fair chance to participate into the tournament.

Overall, previous research has shown that sampling strategies have minor impact to GP performance. Consequently, researchers paid more attention to the second step in tournament selection: selection. Baeck [3] introduced the selection probability of an individual of rank j in one tournament for a minimization task, with an implicit assumption that the population is wholly diverse. Blickle and Thiele [4] extended the selection probability model in [3] to describe the selection probability of individuals with the same fitness f_j. They defined the worst individual to be ranked 1^{st} and introduced the cumulative fitness distribution, which denotes the number of individuals with fitness value f_j or worse.

In this paper, we propose a method for ranking individuals in tournament selection that is based on the use of a statistical test. To the best of our knowledge, this techniques has not been studied in GP. The detailed description of our method will be presented in Sect. 3.

3 Methods

This section describes two new tournament selection techniques. The first technique is called *Statistics-TS1*. Similar to the standard tournament selection, a number of individuals are randomly selected and compared. The winner is then chosen to go to the mating pool. However, instead of using the fitness value for comparing between individuals, a statistical test was applied to the error vector of these individuals. For a pair of individuals, if the test shows that they are different, then the individual with better fitness value is the winner. Conversely, if the test confirms that two individuals are not different, a random individual is selected from the pair. After that, the winner individual is tested against other individuals and the process is repeated for all individuals in the tournament size. The detailed description of Statistics-TS1 is presented in Algorithm 1.

Algorithm 1. Statistics test tournament selection 1

Input: Tour size, Population.
Output: The winner individual.
$A \longleftarrow RandomIndividual()$;
for $i \leftarrow 1$ **to** $TourSize$ **do**
  $B \longleftarrow RandomIndividual()$;
  $sample1 \longleftarrow Error(A)$;
  $sample2 \longleftarrow Error(B)$;
  $p - value \longleftarrow Testing(sample1, sample2)$;
  if $p - value < alpha$ **then**
    $A \longleftarrow GetBetterFitness(A, B)$;
  else
    $A \longleftarrow GetRandom(A, B)$;
  end
end
$TheWinnerIndividual \longleftarrow$ A ;

Table 1. Problems for testing statistical tournament selection

Abbreviation	Name	Attributes	Training	Testing
A. **Benchmarking Problems**				
F1	korns-11	5	20	20
F2	korns-12	5	20	20
F3	korns-14	5	20	20
F4	vladislavleva-1	2	20	2025
F5	vladislavleva-2	1	100	221
F6	vladislavleva-4	5	500	500
F7	vladislavleva-5	3	300	2640
F8	vladislavleva-6	2	30	93636
F9	vladislavleva-7	2	300	1000
F10	vladislavleva-8	2	50	1089
B. **UCI Problems**				
F11	airfoil_self_noise	5	800	703
F12	casp	9	100	100
F13	Slump_test_Compressive	7	50	53
F14	slump_test_FLOW	7	50	53
F15	slump_test_SLUMP	7	50	53
F16	winequality-red	11	800	799
F17	winequality-white	11	1000	1000
F18	wpbc	31	100	98

In Algorithm 1, the function $RandomIndividual()$ returns a random individual from the GP population. Function $Error(A)$ calculates the vector error of individual A and function $Testing(sample1, sample2)$ performs a Wilcoxon signed rank test on two samples. Two last functions, $GetBetterFitness(A, B)$ and $GetRandom(A, B)$ aims at finding the better fitness individual among A and B or returning a random individual between two, respectively. Finally, *alpha* is the critical value used to decide if the null hypothesis is rejected or accepted. If the output of the test $(p - value)$ is smaller than the critical value, then the null hypothesis is rejected. This means that two individuals are significantly different and the better individual is selected as the winner. If the test can not reject the null hypothesis, then a random individual is selected from the pair.

The second tournament selection is called *Statistics-TS2*. Statistics-TS2 is similar to Statistics-TS1 but aims at reducing code grow in the GP population. In Statistics-TS2, if the statistical test can not reject the null hypothesis, then the individual with smaller size is selected from the pair. In other words, if two individuals involved in the test are not statistically different, then the smaller individual will be the winner.

Table 2. Evolutionary parameter values.

Parameter	Value
Population size	500
Generations	100
Selection	Tournament
Tournament size	7
Crossover probability	0.9
Mutation probability	0.1
Initial Max depth	6
Max depth	17
Max depth of mutation tree	15
Raw fitness	mean absolute error on all fitness cases
Trials per treatment	30 independent runs for each value

4 Experimental Settings

In order to measure the impact of the two new tournament selection to GP performance, we tested them on eighteen multivariate regression problems. Among these, ten problems are benchmark problems [16] and eight problems were taken from UCI machine learning dataset [2]. The tested problems are presented in Table 1.

The GP parameters used for our experiments are shown in Table 2. The terminal set for each problem includes N variables corresponding to the number of attributes of that problem. The function set include eight functions (+, -, *, /, sin, cos, log, exp) that are popularly used in GP. The raw fitness is the mean of absolute error on all fitness cases. The elitism technique was also used in which the best individual in the current generation is copied to the next generation. In the new tournament selection schemes, two critical values (0.05 and 0.1) were used for the Wilcoxon signed rank test to decide if the null hypothesis is rejected. For each problem and each parameter setting, 30 runs were performed.

5 Results and Discussion

This section analyses the performance of two new tournament selection methods and compares them with the standard tournament selection (Standard-TS). There metrics used for the comparison are: training error, testing error and solution size.

The first metric is the mean best fitness on the training data and this is presented in Table 3. This table shows that two new selection methods did not help to improve the performance of GP on the training data. By contrast, the training error of standard tournament selection is often better than that of statistical test based tournament selections. This result is not very surprising since the statistical based tournament selection techniques impose less pressure on the improving training error compared to standard tournament selection. Comparing between

Table 3. The mean best fitness on training data. If the result of Statistics-TS1 and Statistics-TS2 is significantly worse ($p - value < 0.05$) than the result of standard-TS, than its value is printed bold and italic faced.

Problems	Standard-TS	Statistics-TS1		Statistics-TS2	
		alpha = 0.05	alpha = 0.1	alpha = 0.05	alpha = 0.1
A. **Benchmarking Problems**					
F1	1.44	***2.33***	***2.24***	***3.59***	***2.84***
F2	0.24	***0.35***	***0.33***	***0.58***	***0.48***
F3	4.71	6.09	5.50	***6.74***	***6.83***
F4	0.01	0.01	0.01	***0.03***	***0.02***
F5	0.04	0.04	0.04	***0.06***	0.05
F6	0.12	0.12	0.12	0.12	0.12
F7	0.10	0.10	0.09	0.09	0.10
F8	0.37	0.49	***0.62***	***1.08***	***1.05***
F9	1.32	***1.53***	***1.53***	***1.81***	***1.60***
F10	0.42	0.45	0.41	***0.53***	***0.48***
B. **UCI Problems**					
F11	8.17	9.54	8.73	9.00	8.77
F12	3.48	***3.90***	***3.92***	***4.19***	***4.00***
F13	3.35	***4.79***	***4.62***	***7.20***	***6.29***
F14	8.05	***10.02***	***9.82***	***12.22***	***11.90***
F15	4.31	***5.95***	***5.53***	***7.28***	***6.90***
F16	0.49	0.50	***0.50***	***0.52***	***0.50***
F17	0.61	***0.62***	***0.63***	***0.64***	***0.63***
F18	25.04	***30.18***	***28.95***	***32.02***	***31.78***

Statistics-TS1 and Statistics-TS2, the table shows that Statistics-TS2 is often slightly worse than Statistics-TS1 on the training data.

We also conducted a statistical test to compare the training error of standard tournament selection with two new selection methods using a Wilcoxon signed rank test with the confident level of 95 %. If the test shows that the training error of statistical based tournament selection techniques is significantly worse than that value of standard tournament selection, this value is printed bold and italic faced in Table 3. It can be seen that, on most problem, the training error of statistical based selection is significantly worse compared to standard-TS.

The second metric used to compare the performance of the tested tournament techniques is their ability to generalize beyond the training data. In each run, the best solution was selected and evaluated on an unseen data set (the testing set). The testing error of the best individual was then recorded and the median of these values across 30 runs was calculated and presented in Table 4. This table shows

that the testing error of two new tournament selection methods is often better than the value of standard tournament selection. This is very encouraging since the result in Table 3 shows that the training error of statistical based selection is often worse compared to standard tournament selection. The result on the testing error demonstrates that, using statistical test to only select the winner individual for the mating pool when the individual is statistically better than others help to improve the generalization of GP.

Table 4. The Median of test error. If the result of Statistics-TS1 and Statistics-TS2 is significantly better than the result of standard-TS, than their value is printed bold faced. Conversely, if their result is significantly worse than standard-TS, this value is printed bold and italic faced.

Problems	Standard-TS	Statistics-TS1		Statistics-TS2	
		alpha = 0.05	alpha = 0.1	alpha = 0.05	alpha = 0.1
A. **Benchmarking Problems**					
F1	10.75	**6.80**	**6.58**	**5.28**	**4.09**
F2	1.02	**0.89**	**0.90**	**0.82**	**0.82**
F3	38.70	**13.98**	**15.13**	**15.27**	**13.85**
F4	0.93	0.39	0.74	0.81	0.79
F5	0.03	0.04	0.05	0.07	0.05
F6	0.13	0.13	0.13	0.13	0.13
F7	0.22	0.23	0.24	**0.20**	0.18
F8	1.63	1.37	1.90	1.97	2.02
F9	1.86	***2.19***	***2.07***	***2.53***	2.25
F10	1.75	1.76	1.74	1.67	**1.44**
B. **UCI Problems**					
F11	24.85	20.99	27.06	28.23	31.57
F12	4.86	4.79	4.91	**4.65**	**4.64**
F13	7.49	6.86	6.80	8.40	8.03
F14	18.23	15.83	16.01	**13.11**	**14.51**
F15	8.97	8.31	**8.10**	8.57	**8.01**
F16	0.55	0.54	0.56	0.55	0.55
F17	0.66	0.65	0.66	0.66	0.65
F18	40.69	38.19	39.91	**37.03**	**37.06**

The statistical test on the testing error using a Wilcoxon signed rank test with the confident level of 95 % shows that two new tournament selection techniques are more frequently better than standard-TS on the testing error. Precisely, Statistics-TS1 is significantly better than standard-TS on three and four problems with $alpha = 0.05$ and $alpha = 0.1$ respectively while standard-TS is significantly

better than Statistics-TS1 on one problem, $F9$. The testing error of Statistics-TS2 is significantly better than standard-TS on seven and eight problems with $alpha = 0.05$ and $alpha = 0.1$ respectively while standard-TS is significantly better than Statistics-TS2 on only one problem ($F9$) with $alpha = 0.05$. Comparing between Statistics-TS1 and Statistics-TS1, the statistical test shows that Statistics-TS2 is often slightly better than Statistics-TS1 on the unseen data.

Table 5. The average of solutions size of three selection methods. If the solutions found by Statistics-TS1 and Statistics-TS2 are more complex than those found by standard-TS, their value is printed bold and italic faced.

Problems	Standard-TS	Statistics-TS1		Statistics-TS2	
		alpha = 0.05	alpha = 0.1	alpha = 0.05	alpha = 0.1
A. **Benchmarking Problems**					
F1	295.1	263.2	258.7	97.4	118.0
F2	172.9	161.0	145.5	33.9	36.9
F3	260.8	***278.3***	***283.4***	90.0	89.2
F4	175.2	***176.1***	163.4	40.6	48.1
F5	207.8	***213.4***	***235.1***	55.6	61.3
F6	100.8	97.8	84.2	36.4	44.1
F7	120.6	***135.5***	***141.1***	57.0	57.2
F8	176.2	135.1	134.9	55.9	37.8
F9	143.1	***154.8***	***152.6***	50.4	71.2
F10	156.9	***157.2***	***166.4***	47.4	32.6
B. **UCI Problems**					
F11	264.7	***313.5***	***296.9***	185.6	211.6
F12	218.4	164.1	171.6	28.8	45.8
F13	216.4	143.0	152.9	22.8	31.5
F14	185.3	135.1	146.7	20.1	25.1
F15	212.1	164.7	152.1	24.2	30.4
F16	121.4	120.0	121.2	47.1	55.8
F17	147.8	125.6	145.7	40.1	46.1
F18	326.9	97.0	173.5	6.4	11.3

The last metric used to analyze the efficiency of statistics based tournament selection techniques is the size of their solutions. We recored the size of the best fitness individual in each runs. These values are then averaged over 30 runs and are presented in Table 5. In Table 5, when the solutions found of Statistics-TS1 and Statistics-TS2 are more complex than those obtained by standard-TS, their result is printed in bold and italic faced. It can be observed from this that the two new tournament selection techniques often help to find the solution of smaller size.

Apparently, the size of the solutions found by Statistics-TS1 is smaller than that of Standard-TS on most problem. Sometimes, Statistics-TS1 finds solutions that are more complex than the standard tournament selection and this happens on seven out of eighteen problems with $alpha = 0.05$ and on six out of eighteen problems with $alpha = 0.1$. For Statistics-TS2, the size of its solutions is much smaller than that of Standard-TS. It can be seen that the size of the solution obtained by Statistics-TS2 is often equal to one third of the solutions of Standard-TS on most problem. Overall, the results in this section show that statistical based tournament selection methods help GP to find simpler solution and generalize better on unseen data. This result is promising since finding simple solutions that achieve good performance on unseen data is the main objective of GP systems.

6 Conclusions and Future Work

In this paper, we introduced the idea of using a statistical test as part of selection step that utilizes information from fitness case error vectors of GP individuals. We proposed two variations of tournament selection that used statistical tests to select the winner for the mating pool. The effectiveness of the approach was examined on eighteen symbolic regression problems. In the experimental results we observe that the proposed techniques helped GP to reduce code growth and generalisation error.

There are a number of research areas for future work, which arise from this paper. First, we would like to study the approach to improve the performance of the statistical based tournament selection techniques on the training data. This may help the new techniques to perform better on a wider range of problems. One possible approach that can improve the performance of Statistics-TS1 and Statistics-TS2 is to combine them with local search techniques such as Soft Brood Selection [1]. Another approach is to implement these techniques with recent semantic based crossovers [11,13]. Second, at the theoretical level, it is still unclear while Statistics-TS1 and Statistics-TS2 perform well on unseen data though their performance on the training data is not as good as standard tournament selection. One possible reason is that they help to reduce code growth resulting in more parsimonious solutions. It is interesting to compare and analyze these tournament techniques with code bloat control methods like multi-objective GP [8] and operator equalisation [14]. Finally, a potential limitation of the proposed approach is the overuse of statistical tests without consideration for the increased probability of a significant difference being detected by chance [5]. Future research will include an exploration of the impact of different statistical tests and assessment of their suitability.

Acknowledgment. This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2014.09. MON acknowledges the support of Science Foundation Ireland grant 13/IA/1850.

References

1. Altenberg, L.: The evolution of evolvability in genetic programming. In: Advances in Genetic Programming, pp. 47–74. MIT Press (1994)

2. Bache, K., Lichman, M.: UCI machine learning repository (2013). http://archive.ics.uci.edu/ml
3. Bäck, T.: Selective pressure in evolutionary algorithms: a characterization of selection mechanisms. In: Proceedings of the First IEEE Conference on Evolutionary Computation, pp. 57–62. IEEE Press, Piscataway (1994)
4. Blickle, T., Thiele, L.: A comparison of selection schemes used in evolutionary algorithms. Evol. Comput. **4**(4), 361–394 (1996)
5. Cumming, G.: Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. Routledge, New York (2012)
6. Fang, Y., Li, J.: A review of tournament selection in genetic programming. In: Cai, Z., Hu, C., Kang, Z., Liu, Y. (eds.) ISICA 2010. LNCS, vol. 6382, pp. 181–192. Springer, Heidelberg (2010)
7. Gathercole, C.: An investigation of supervised learning in genetic programming. Ph.D. thesis. University of Edinburgh (1998)
8. Jong, E.D.D., Pollack, J.B.: Multi-objective methods for tree size control. Genet. Program. Evolvable Mach. **4**(3), 211–233 (2003)
9. Kim, J.J., Zhang, B.T.: Effects of selection schemes in genetic programming for time series prediction. Proc. Congr. Evol. Comput. **1**, 252–258 (1999)
10. Nguyen, Q.U., Nguyen, X.H., O'Neill, M., McKay, R.I., Galvan-Lopez, E.: Semantically-based crossover in genetic programming: application to real-valued symbolic regression. Genet. Program. Evolvable Mach. **12**(2), 91–119 (2011)
11. Nguyen, Q.U., Pham, T.A., Nguyen, X.H., McDermott, J.: Subtree semantic geometric crossover for genetic programming. Genet. Program. Evolvable Mach. **17**(1), 25–53 (2016)
12. Pawlak, T.P., Wieloch, B., Krawiec, K.: Review and comparative analysis of geometric semantic crossovers. Genet. Program. Evolvable Mach. **16**(3), 351–386 (2015)
13. Pawlak, T.P., Wieloch, B., Krawiec, K.: Semantic backpropagation for designing search operators in genetic programming. IEEE Trans. Evol. Comput. **19**(3), 326–340 (2015)
14. Silva, S., Dignum, S., Vanneschi, L.: Operator equalisation for bloat free genetic programming and a survey of bloat control methods. Genet. Program. Evolvable Mach. **13**(2), 197–238 (2012)
15. Sokolov, A., Whitley, D.: Unbiased tournament selection. In: Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation, pp. 1131–1138. ACM, New York (2005)
16. White, D.R., McDermott, J., Castelli, M., Manzoni, L., Goldman, B.W., Kronberger, G., Jaskowski, W., O'Reilly, U.M., Luke, S.: Better GP benchmarks: community survey results and proposals. Genet. Program. Evolvable Mach. **14**(1), 3–29 (2013)
17. Xie, H., Zhang, M.: Parent selection pressure auto-tuning for tournament selection in genetic programming. IEEE Trans. Evol. Comput. **17**(1), 1–19 (2013)
18. Xie, H., Zhang, M., Andreae, P., Johnston, M.: Is the not-sampled issue in tournament selection critical? In: IEEE World Congress on Computational Intelligence, pp. 3710–3717, June 2008
19. Xie, H., Zhang, M., Andreae, P.: Automatic selection pressure control in genetic programming. In: Yang, B., Chen, Y. (eds.) 6th International Conference on Intelligent System Design and Applications, pp. 435–440. IEEE (2006)
20. Xie, H., Zhang, M., Andreae, P., Johnson, M.: An analysis of multi-sampled issue and no-replacement tournament selection. In: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation, GECCO 2008, pp. 1323–1330. ACM, New York (2008)