Randomness and Computation
Joint Workshop "New Horizons in Computing" and "Statistical Mechanical Approach to Probabilistic Information Processing" (18-21 July, 2005, Sendai, Japan)

A Novel Diversity Measure of Genetic Programming

Naoki Mori [1]
Department of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, Sakai 599-8531, Osaka, Japan

1 Introduction

The Genetic Programming (GP)[1] is one of the novel evolutionary computation which is applied to complicated structure problems. The most important feature of GP is that individuals in GP are represented by the tree structure. The expression ability of tree structure is very flexible, and this is why GP is expected to solve many difficulties of Genetic Algorithms (GAs).

However, the analysis of GP dynamics is difficult because of tree structure[2]. To solve this problem, I propose a novel diversity measure in genetic programming by means of subtree entropy. I also propose a tree simplification method which removes redundancy parts from an individual genotype. To show an advantage of our methods, the computational experiments are carried out taking a symbolic regression problem as an example.

2 GP setting

2.1 Problem and fitness

The problem is finding the target function $\cos 2X$ by given 20 random points. X range is $[-\pi, \pi]$. 20 points are generated randomly in each partition of 20 even intervals, and fixed in different runs. There exist 3 different main expressions of solution, $1 - 2\sin^2 X$, $\sin(\frac{\pi}{2} - 2X)$ and $\sin(\frac{\pi}{2} + 2X)$. The raw fitness(RF) is calculated as sum of 20 absolute errors. Modified fitness is represented by $f = \frac{1}{1+RF}$.

2.2 GP setting

I use following operators and operands for individual expression.

Operator: +, -, *, %, sin **Operand** X, 1, 0(only for simplify)

"%" is the protected divide; ex. 1%0 → 1.

The number of generations is set to 200 and the population size is set to 500. I use subtree mutation with mutation rate 0.1 and simple crossover with crossover rate 0.9.

3 Simplification

In this paper, I utilize the operation called "Simplification" to remove redundancy tree parts from an individual genotype. Though there are many kinds of way to achieve this operation, I proposed following numerical way.

Check all the subtrees if they are identical to X, 1 or 0. Generally, we can utilize any nodes for these candidate. If a subtree is equal to 3 basic candidate nodes(X, 1, 0), then replace the whole subtree under this node to a correspond node.

Simplification is applied recursively until there is no change in tree. Even if one node is changed, simplification is restarted from root node. (In the real implementation, simplification goes on the same depth nodes in order to reduce the computational time. The nodes order of check is breadth first order.)

[1] E-mail: mori@cs.osakafu-u.ac.jp

4 Entropy measure

One of the good measure for diversity is entropy. Though the standard deviation (S.D.) is another candidate, S.D. become maximum in set like {0,0,0,100,100,100} (there are only 2 groups, one is consist of smallest value, another is consist of maximum value). To solve this problem, I proposed a novel index called *subtree entropy*. To obtain this value, the template of subtree is set first. Next, check the all subtree which has the structure of this template. The subtree entropy is calculated by this subtree set. In this paper, I only utilize the 3 nodes template which has one root and two leaves (like this shape $\wedge$).

5 Result

To compare the effects of proposed methods, I made 3 groups. First group contains only runs which found optimum solutions between generation 20 $\sim$ 29. Second group contains only runs which found optimum solutions between generation 50 $\sim$ 69. Last group contains only runs which failed to find optimum solutions. Fig. 1 shows the result of subtree entropy of original search results. Fig. 2 shows the result of subtree entropy of after applied simplification for original search results. Only Fig. 2 shows the result that subtree entropy of better searches reduce rapidly. Fig. 3 show the Welch's t test result between opt 20-29 group and fail group of original search results. Fig. 4 show the Welch's t test result between opt 20-29 group and fail group of after applied simplification. Fig. 3 and Fig. 4 show the result that statistical difference becomes clear only cases of applied simplification.

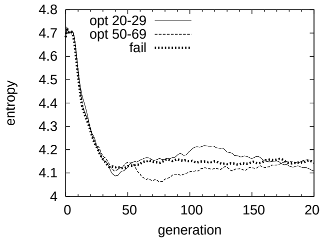

Fig. 1 Subtree entropy (original)

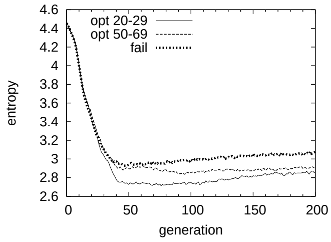

Fig. 2 Subtree entropy (simplification)

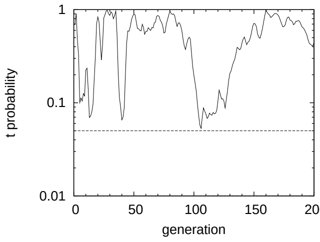

Fig. 3 t test result (original)

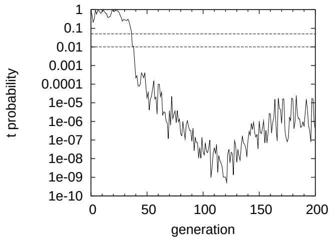

Fig. 4 t test result (simplification)

6 Conclusions

In this paper, I proposed new diversity measure of GP called subtree entropy. I also proposed tree simplification method. Simulation results show the effectiveness of proposed methods. Deeper analysis of subtree entropy and applying the proposed method to another cases are subjects of further study.

References

[1] J. R. Koza, Martin A. Keane et al.: Genetic Programming IV, Kluwer Academic Publisher s, 2003.

[2] E. Burke, S. Gustafson and G. Kendall: Survey and analysis of diversity measures in genetic programming, Proceedings of the Genetic and Evolutionary Computation Conference, 2002.