Universal Information Distance for Genetic Programming

Marco Gaudesi
Politecnico di Torino
Corso Duca degli Abruzzi, 24
10129 Torino, Italy
marco.gaudesi@polito.it

Giovanni Squillero
Politecnico di Torino
Corso Duca degli Abruzzi, 24
10129 Torino, Italy
giovanni.squillero@polito.it

Alberto Tonda
INRA UMR 782, MALICES
1 Avenue Lucien Brétignères
78850 Thiverval-Grignon,
France
alberto.tonda@grignon.inra.fr

ABSTRACT

This paper presents a genotype-level distance metric for Genetic Programming (GP) based on the *symmetric difference* concept: first, the *information* contained in individuals is expressed as a set of symbols (the content of each node, its position inside the tree, and recurring parent-child structures); then, the difference between two individuals is computed considering the number of elements belonging to one, but not both, of their symbol sets.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

Algorithms; Measurements; Genetic Programming; Diversity Preservation; Distance Metric; Fitness Sharing; Experimental Analysis; Individual Encoding; Symbolic Regression

1. INTRODUCTION

In an Evolutionary Algorithms (EA) a reliable *distance metric* between individuals can be used to promote diversity inside the population's gene pool, to avoid the overexploitation of niches in the fitness landscape, to balance exploration and exploitation, and – broadly speaking – to prevent *premature convergences*.

Distances can be computed at *genotype*, *phenotype* or *fitness* level. "Genotype" is the internal representation of candidate solutions; "fitness" is a set of values that encode the goodness of an individual for the specific purpose of the problem. "Phenotype" is much harder to characterize: in biology, the phenotype is the sum of all the observable characteristics of an organism that result from the interaction of its genotype with the environment; but in evolutionary computation, there is no proper *environment*, just its indirect effects modeled by the fitness function.

GECCO'14, July 12-16, 2014, Vancouver, BC, Canada.
ACM 978-1-4503-2881-4/14/07
http://dx.doi.org/10.1145/2598394.2598440.

The proposed approach computes the *symmetric difference* [1] between the global information contained in two individuals; the global information itself is evaluated resorting to corresponding recurring structures in the trees, a concept similar to the n-grams [8] used in natural language analysis.

2. BACKGROUND

2.1 Fitness Sharing in Genetic Programming

The symbolic regression problem [7] is commonly used as a paradigmatic illustration when introducing GP: the goal is to automatically extract free-form equations that correlate data from a given experimental dataset. Candidate solutions are formulas, encoded as trees with terminal nodes corresponding to constants and variables of the problem, while intermediate nodes encode mathematical functions.

Fitness sharing is an established method to enforce diversity inside the population of an EA [5, 6], and it relies upon the definition of a distance measurement between individuals. The general idea of fitness sharing is to artificially decrease the fitness of individuals in crowded areas of the search space. The fitness f_i of an individual I_i is modified in a fitness $f_i' = f/m_i$, where m_i is dependent upon the number of individuals in a given radius σ_s from individual I_i, and their distance from the individual itself. In particular,

$$m_i = \sum_{j=0}^{N} sh(I_i, I_j) \tag{1}$$

where N is the number of individuals in the population, and $sh(I_i, I_j)$ is defined as

$$sh(I_i, I_j) = \begin{cases} 1 - (\frac{d(I_i,I_j)}{\sigma_s})^{\alpha} & d(I_i, I_j) < \sigma_s \\ 0 & d(I_i, I_j) \geq \sigma_s \end{cases} \tag{2}$$

where σ_s is a user-defined radius, and $d(I_i, I_j)$ is a distance measure applicable to the individuals' representation. α is a constant parameter that regulates the shape of the sharing function. It has been experimentally demonstrated that fitness sharing can lead to important improvements in GP, permitting significantly smaller populations to achieve similar results to larger populations using raw fitness [3].

2.2 Symmetric Difference

In set theory, the symmetric difference [1] of two sets A and B is defined as

$$A \triangle B = A \cup B - A \cap B \tag{3}$$

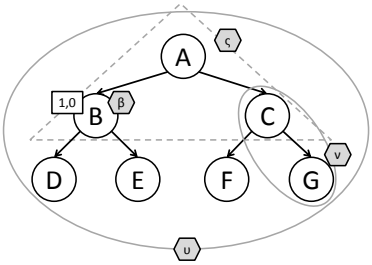

Figure 1: Example of the symbols adopted: node and position (β), node and child (ν), node and children (ξ), node and grandchildren (υ)

In words, the symmetric difference contains all elements which are in either of the sets and not in their intersection. The symmetric difference is usually denoted with the symbol "$\triangle$". The symmetric difference exhibits useful properties for a distance: it is commutative and the empty set is neutral.

3. PROPOSED APPROACH

A generic genotypic *Universal Information Distance* (UID) for individuals in GP is proposed. Considering two individuals I_i and I_j, the UID is defined as

$$UID(I_i, I_j) = \frac{|S(I_i) \triangle S(I_j)|}{|S(I_i)| + |S(I_j)|} \quad (4)$$

where $S(I)$ represents a set of symbols associated individual I, $\triangle$ is the symmetric difference as defined in Equation 3, and operator $|S|$ denotes the cardinality of set S.

The proposed genotype-level distance stems from a previous paper presenting a similar metric for Linear Genetic Programming (LGP) [2]. In that work, following the idea that recurring structures might possess meaning, nodes and *n-grams* of nodes are adopted as symbols to characterize an individual. An n-gram is a group of n consecutive items from a longer sequence. Four new different kinds of symbols are used to characterize a GP individual:

Node and position symbols encode the content of a node and its (x, y) position inside the GP tree, where x is the level of the node, and y its order inside the level;

Node and child symbols encode node A and one of its children B, without considering its relative position;

Node and children symbols encode the content of node A and all its children, also taking into account their position with respect to the parent node;

Node and grandchildren symbols encode the content of node A, all its children, and all its children's children, taking into account their position with respect to the parent node.

Given a specific individual, all symbols belonging to each category are computed and stored inside its symbol set. For a summary of the symbols, see Figure 1.

Symbols, assigned to an individual when it is created, can be later used to compute a *distance* between different individuals in the population. First, a symmetric difference is performed on their symbol sets, and then the cardinality of the resulting set is used to quantify the distance.

4. EXPERIMENTAL RESULTS

In order to validate the proposed approach, the minimal GP engine *TinyGP* [4] was modified to include the UID. As testbench, the very same example of symbolic regression reported in the book was used. Each experiment evolved 1000 individuals for 100 generations, and was repeated ten times. As a baseline, TinyGP was run unmodified. Then, when using fitness sharing, three different radius where used: 0.1, 0.15 and 0.2 (the UID is a normalized value). For the sake of efficiency, symbols described in Section 3 are computed resorting to the DJB[1] hash function. Finally, to further speed up calculations, m_i was simply set to the number of individuals within the given radius divided by a constant k. Five different values for k were tested.

Table 1: Percentage of fitness improvements using fitness sharing with different parameters.

radius	$k = 1,000$	$k = 500$	$k = 100$	$k = 50$	$k = 10$
0.10	43.72	58.73	84.28	86.33	25.61
0.15	77.02	63.41	56.77	62.99	40.79
0.20	75.78	39.02	11.30	11.80	10.24

Table 1 reports the percentage of fitness improvements using fitness sharing with the different parameters. It may be noted that results using the diversity preservation mechanism are definitely superior, and that the UID can be exploited to evaluate similarity between the individuals.

5. CONCLUSIONS

This paper proposes a new distance metric for Genetic Programming, based on the normalized symmetric difference between the *information* contained in the genome, represented as symbols based upon the node's content and recurring structure in the binary trees. Experiments demonstrate its potential and usefulness for implementing fitness-sharing in a classical symbolic regression problem.

6. REFERENCES

[1] Symmetric Difference. In E. J. Borowski and J. M. Borwein, editors, *The HarperCollins Dictionary of Mathematics*. HarperCollins, 1991.

[2] M. Gaudesi, G. Squillero, and A. Tonda. An Efficient Distance Metric for Linear Genetic Programming. pages 925–932. GECCO '13, ACM, 2013.

[3] R. I. McKay. Fitness Sharing in Genetic Programming. In *GECCO*, pages 435–442, 2000.

[4] R. Poli, W. W. B. Langdon, N. F. McPhee, and J. R. Koza. *A field guide to genetic programming*. Lulu. com, 2008.

[5] C. D. Rosin and R. K. Belew. New methods for competitive coevolution. *Evolutionary Computation*, 5(1):1–29, 1997.

[6] B. Sareni and L. Krahenbuhl. Fitness sharing and niching methods revisited. *Evolutionary Computation, IEEE Transactions on*, 2(3):97–106, 1998.

[7] M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

[8] C. Suen. N-gram statistics for natural language understanding and text processing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):164–172, 1979.

[1] http://cr.yp.to/djb.html