

The Impact of Chromosome Lineage Upon Genetic Program Modeling

by Gary D. Boetticher and Kim Kaminsky

ABSTRACT—One of the challenges in data mining is to provide sufficient coverage of the search space in order to produce an acceptable model. Traditionally, genetic programs (GP) consider all chromosomes within a population for breeding purposes. Considering the enormous size of the search space, it is imperative to focus breeding efforts in genetic programs in order to attain a better solution in less time. This research examines the lineage of genetic programs in order to identify any breeding patterns. Five separate experiments have been conducted where chromosomes are grouped into five classes. Lineage patterns are assessed for the best-, middle-, and worst-class parental chromosomes. Based upon the results, a new genetic programming process is proposed.



Gary D. Boetticher

SOLVING LARGE PROBLEMS USING GENETIC PROGRAMS (GPs) consumes excessive amounts of computer resources. Though genetic programs may successfully evolve solutions to complex problems, their use can be cost-prohibitive.

What is desired is a more efficient approach to exploring the search space. This may be accomplished qualitatively by focusing the search efforts or, quantitatively, by increasing the number of searches.

This research explores the qualitative approach by examining the breeding patterns of a GP. Key questions focused upon are:

- Does chromosome lineage information provide any insight into the effectiveness of solving problems?
- If so, how could these insights be utilized to make better breeding decisions?

Goals of the project

Gaining a better understanding about the lineage of a chromo-

Original GP			Lineage-Based GP		
Fitness	Final r2	Gen.	Fitness	Final r2	Gen.
591.8	0.8734	29.1	740.9	0.9315	28.5
210.9	0.7244	50.0	346.5	0.8069	48.6

Table 1. Original vs. Lineage Approach

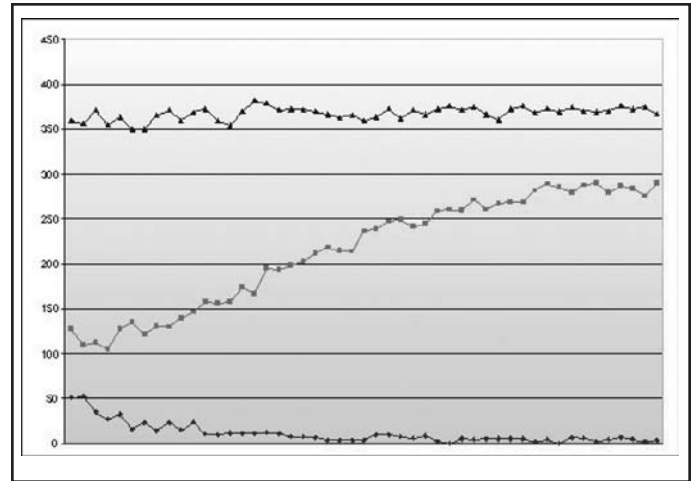


Figure 1. Results from an Initial Experiment

some, in terms of how fitness values propagate over generations, is beneficial in several ways.¹ Greater emphasis could be placed on those chromosomes that produce better offspring. Secondly, the utility of such a discovery could focus the search efforts, thus reducing training time, and requiring fewer computing resources. All of these benefits are immensely important when applying GPs to large, complex, noisy problem spaces.

To explore the role chromosome lineage plays in the breeding process, five initial experiments have been conducted using synthetic datasets. Chromosomes are clustered into different classes (e.g., best, middle, and lower classes). Each of these classes has been tracked over a generation to determine whether certain classes are prone to producing good (or poor) solutions.

Based upon the results of the initial set of experiments, an alternative breeding approach is proposed that focuses on those chromosomes with a solid pedigree. A second set of experiments examines this novel approach along with a traditional approach to determine the merit of focusing on a certain portion of a GP population.

Results

Figure 1 depicts the results from a typical experiment. The x-axis represents the number of GP generations (1 through 50) per trial. The y-axis shows the average fitness values for the best-class, middle-class, and worst-class groups. The top line in the 300-400 range represents average fitness values of the offspring for the best-class parents. The middle line is the average fitness values of the offspring for the middle class parent chromosomes. The bottom line shows the average fitness values of the

offspring for the worst class parents. Figure 1 shows a clear distinction between best-, middle-, and worst-class groups. At no time do any of the group averages intersect. A t-test reveals these differences as statistically significant.

These results in Fig. 1 are representative of all experiments. The next step applies the observed breeding habits to two equations. Table I shows the results after applying the lineage technique in 20 trials.

These results show clearly that a lineage-based GP produces better models in terms of fitness and correlation in fewer generations.

References

¹G. Boetticher and K. Kaminsky, "The Impact of Chromosome Lineage upon Genetic Program Modeling," *ISSO Annual Report*, 2004. 122-27.

Publications

Boetticher, G. and K. Kaminsky. "The Assessment and Application of Lineage Information in Genetic Programs for Producing Better Models," IEEE Information Reuse and Integration Conference 2006 (IRI), Big Island of Hawaii, September 2006.

Presentations

Boetticher, G. and K. Kaminsky. "The Assessment and Application of Lineage Information in Genetic Programs for Producing Better Models," IEEE Information Reuse and Integration Conference 2006 (IRI), Big Island of Hawaii, September 2006.



SITES—Departmental offices for the College of Science and Computer Engineering at UHCL are located in the Bayou Building. The nearby Delta Building also houses experiments for NASA-JSC projects.