How Statistics Can Help in Limiting the Number of Fitness Cases in Genetic Programming

Mario Giacobini, Marco Tomassini, Leonardo Vanneschi
Institut d'Informatique, University of Lausanne, 1015 Lausanne, Switzerland
{Mario.Giacobini,Marco.Tomassini,Leonardo.Vanneschi}@iis.unil.ch

For most real world applications of GP it is well known that fitness evaluation is the most time consuming operation. It would thus be interesting to establish criteria that can help in limiting the time spent in this phase as much as possible without compromising results in terms of quality. One is thus confronted with two problems: how to select a sufficient number of fitness cases and how to choose those fitness cases in such a way that they are effective in driving the learning process towards a solution. Here we approach the former problem from a standard statistical and information-theoretical viewpoint.

Let us consider a GP problem where the target function is defined on N fitness cases. It can be shown that the mean distance of all the individuals of a population from the target function is normally distributed ($\bar{x} \sim N(\mu, \sigma)$). A standard result for the confidence interval gives [2]:

$$P\left(\bar{x} - t_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} + t_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) \geq 1 - \alpha,$$

where $1-\alpha$ is the confidence with which we can expect the mean μ to be contained in the given interval. The $t_{\alpha/2}$ is the Student cumulative distribution such that the mean deviates from its true value in the interval $(-t_{\alpha/2}, t_{\alpha/2})$. The standard deviation σ is unknown but can be estimated by the sample variance S. If we set $K = 2t_{\alpha/2}(\sigma/\sqrt{n})$, the length of the confidence interval, we get a function relating the number of fitness cases n that must be used in order for the mean fitness to be estimated to be in the confidence interval K with a given probability $1 - \alpha$.

The target function $g : \{x_1, ..., x_N\} \rightarrow \{y_1, ..., y_M\}$ can be seen, from another viewpoint as a random variable, and it is thus possible to calculate its entropy:

$$H(g(x)) = -\frac{1}{ln(N)} \sum_{j=1}^{M} p_j \ln(p_j),$$

where $p_j = P(g(x) = y_j)$ for $j \in \{1, ..., M\}$. Such a measure indicates the quantity of information needed to determine the function itself, i.e., the minimal number of fitness cases needed for a reliable reconstruction of the target function g.

To test the validity of our assumptions we have studied two simple problems: a seven variables boolean function, and a step function. The aim is to show the statistical behavior of the GP evolutions when the number of fitness cases is decreased. For such a purpose standard GP has been run 50 times for each percentage of fitness cases, randomly chosen with uniform probability. For both target functions we observe a similar statistical behavior. When the number of fitness cases is such that the level of confidence is 0.99 we observe a normal convergence behavior, while with a number of fitness cases lower than such a value we get oscillating curves and the length of the confidence interval drastically increases. Such a result is consistent with the entropy which is found to be close to that value of n. For the boolean function the minimal n is about 18 with respect to 128 fitness cases, while for the step function we get 27 instead of the full 100 cases.

Our results are of a statistical nature and thus they do not depend on the particular problem. Some previous works have tackled the problem of limiting the number of fitness cases heuristically (e.g. [1]). Knowing that the number of fitness cases can be significantly reduced for statistical reasons can be useful for selecting a reduced but sufficient number of significant fitness cases.

References

[1] C. Gathercole and P. Ross. Tackling the boolean even N parity problem with genetic programming and limited-error fitness. In John R. Koza, Kalyanmoy Deb, Marco Dorigo, David B. Fogel, Max Garzon, Hitoshi Iba, and Rick L. Riolo, editors, *Genetic Programming 1997: Proceedings of the Second Annual Conference*, pages 119–127, San Francisco, CA, USA, 1997. Morgan Kaufmann.

[2] S. M. Ross. *Introduction to Probability and Statistics for Engineers and scientists.* Academic Press, New York, 2000.