Diversity in Multipopulation Genetic Programming

Marco Tomassini[1], Leonardo Vanneschi[1], Francisco Fernández[2], and Germán Galeano[2]

[1] Computer Science Institute, University of Lausanne, Lausanne, Switzerland
[2] Computer Science Institute, University of Extremadura, Spain

In the past few years, we have done a systematic experimental investigation of the behavior of multipopulation GP [2] and we have empirically observed that distributing the individuals among several loosely connected islands allows not only to save computation time, due to the fact that the system runs on multiple machines, but also to find better solution quality. These results have often been attributed to better diversity maintenance due to the periodic migration of groups of "good" individuals among the subpopulations. We also believe that this might be the case and we study the evolution of diversity in multi-island GP. All the diversity measures that we use in this paper are based on the concept of *entropy* of a population P, defined as $H(P) = -\sum_{j=1}^{N} F_j \log(F_j)$. If we are considering phenotypic diversity, we define F_j as the fraction n_j/N of individuals in P having a certain fitness j, where N is the total number of fitness values in P. In this case, the entropy measure will be indicated as $H_p(P)$ or simply H_p. To define genotypic diversity, we use two different techniques. The first one consists in partitioning individuals in such a way that only identical individuals belong to the same group. In this case, we have considered F_j as the fraction of trees in the population P having a certain genotype j, where N is the total number of genotypes in P and the entropy measure will be indicated as $H_G(P)$ or simply H_G. The second technique consists in defining a distance measure, able to quantify the genotypic diversity between two trees. In this case, F_j is the fraction of individuals having a given distance j from a fixed tree (called *origin*), where N is the total number of distance values from the origin appearing in P and the entropy measure will be indicated as $H_g(P)$ or simply H_g. The tree distance used is Ekárt's and Németh's definition [1].

Figure 1 depicts the behavior of H_G, H_g and H_p during evolution for the symbolic regression problem, with the classic ploynomial equation $f(x) = x^4 + x^3 + x^2 + x$, an input set composed of 1000 fitness cases and a set of functions equal to F={*,//,+,-}, where // is like / but returns 0 instead of *error* when the divisor is equal to 0. Fitness is the sum of the square errors at each test point. Curves are averages over 100 independent runs for generational GP, crossover rate: 95%, mutation rate: 0.1%, tournament selection of size: 10, ramped half and half initialization, maximum depth of individuals for the creation phase: 6, maximum depth of individuals for crossover: 17, elitism. Genotypic diversity for the panmictic case tends to remain constant over time, and to have higher values than the distributed case. On the contrary, the average phenotypic entropy for the multipopulation case tends to remain higher than in the panmictic case. The oscillating behavior of the multipopulation curves when groups of individuals are sent and received is not surprising: it is due to the sudden change in diversity when new individuals enter a subpopulation. Finally, we remark that the behavior of the two measures used to calculate genotypic diversity (H_G and H_g) is qualitatively equivalent. Analogous results

E. Cantú-Paz et al. (Eds.): GECCO 2003, LNCS 2724, pp. 1812–1813, 2003.

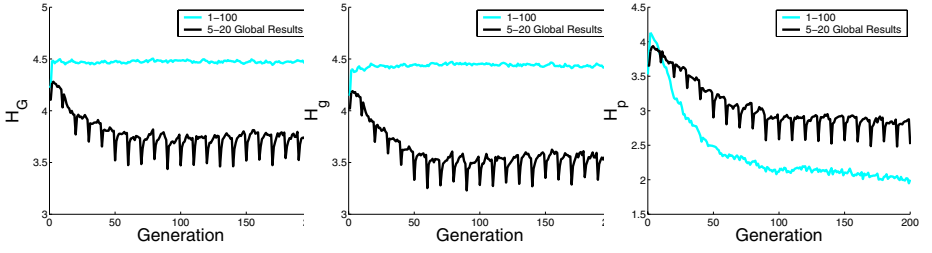

Fig. 1. Symbolic Regression Problem. 100 total individuals. (a): Genotypic diversity calculated using the H_G measure (see the text). Gray curve: panmictic population. Black curve: aggregated subpopulations. (b): Genotypic diversity calculated with the H_g measure (c): Phenotypic diversity (H_p)

Table 1. Numbers represent the average amount of time spent by a solution or a part thereof in the corresponding population. Results refer to the Ant problem with a population size of 1000

Pop 1	Pop 2	Pop 3	Pop 4	Pop 5
20	18.62	21.5	15.62	24.5

(not shown here for lack of space) have been obtained for the Even Parity 4 problem and for the Artificial Ant on the Santa Fe Trail problem.

It is also interesting to study how solutions originate and propagate in the distributed system. Table 1 is a synthesis of a few runs of the Artificial Ant problem. Although data are not statistically significant (too few runs have been done), they do indicate that all the islands participate in the formation of a solution.

By defining genotypic and phenotipic diversity indices and by monitoring their variation over a large number of runs on three standard test problems, we have empirically studied how diversity evolves in the distributed case. While genotypic diversity is not much affected by splitting a single population into multiple ones, phenotypic diversity, which is linked to fitness, remains higher in the multipopulation case for all problems studied here.

We have also studied how solutions arise in the distributed case, and we have seen that all the subpopulations contribute in the building of the right genetic material, which again seems to confirm that smaller communicating populations are more effective than a big panmictic one. In conclusion, using multiple loosely coupled populations is a natural and easy way for maintaining diversity and, to some extent, avoiding premature convergence in GP.

References

1. A. Ekárt and S.Z. Németh. Maintaining the diversity of genetic programs. In J.A. Foster et al., editor, *Genetic Programming, Proceedings of the 5th European Conference, EuroGP 2002*, volume 2278 of *LNCS*, pages 162–171. Springer-Verlag, 2002.
2. F. Fernández, M. Tomassini, and L. Vanneschi. An empirical study of multipopulation genetic programming. *Genetic Programming and Evolvable Machines*, March 2003. Volume 4. Pages 21–51. W. Banzhaf Editor-in-Chief.