Improving Generalisation Performance Through Multiobjective Parsimony Enforcement

Yaniv Bernstein, Xiaodong Li, Vic Ciesielski, and Andy Song

School of Computer Science and Information Technology
RMIT University, VIC 3001, Melbourne, Australia

Abstract. This paper describes POPE-GP, a system that makes use of the NSGA-II multiobjective evolutionary algorithm as an alternative, parameter-free technique for eliminating program bloat. We test it on a classification problem and find that while vastly reducing program size, the technique does improve generalisation performance.

Program Bloat, the phenomenon of ever-increasing program size during a GP run, is a recognised and widespread problem. Traditional techniques to combat program bloat are program size limitations or parsimony pressure (penalty functions). These techniques suffer from a number of problems, in particular their reliance on parameters whose optimal values it is difficult to *a priori* determine. In this work we study the performance of POPE-GP, a new algorithm that uses the NSGA-II multiobjective algorithm as the basis for parsimony enforcement. We are especially interested in finding out if small solutions generalise better than large solutions. To achieve this, we compare the performance of POPE-GP on a real-world classification problem with that of a GP with more traditional parsimony control and a GP with no control at all, paying particular attention to the performance of solutions on the unseen testing set as a measure of generalisation performance.

The Pseudo-Objective Parsimony Enforcement GP (POPE-GP) uses the NSGA-II multiobjective optimisation algorithm [1] as a base for its operation. The two objectives are defined as being the actual objective of the GP run (the fitness) and the size of the program. Once these objectives have been defined, the NSGA-II algorithm attempts to find the Pareto Front for these two objectives. We compared the generalisation performance of classifier programs generated by the POPE-GP algorithm with those generated by a standard GP with a depth limit of eight and one with no limits at all. We used the Wisconsin Breast Cancer Database[1], which has been widely used as a testbed for classification. We divided the 699 instances in the data set randomly into training and testing sets, so that 70% (479 instances) of the data made up the training set and the remaining 30% (204 instances) constituted the testing set. We used the RMITGP[2] GP programming library with strongly-typed GP. The fitness of an individual was taken to be the gross classification error – ie. the number of instances in the training set that are misclassified.

[1] *http://www.ics.uci.edu/∼mlearn/MLSummary.html*
[2] *http://yallara.cs.rmit.edu.au/∼dylanm/rmitgp.html*

K. Deb et al. (Eds.): GECCO 2004, LNCS 3103, pp. 702–703, 2004.

Table 1. (a) End-of-run average values for the algorithms tested. (b) Mean classification accuracy on the testing set.

Algorithm	AvDepth	AvFitness	AvSize	BestDepth	BestFitness	BestSize
POPE-GP (500)	6.71	0.9586	18.77	9.40	0.9865	31.50
POPE-GP (50)	5.40	0.9467	13.12	8.00	0.9811	23.20
Depth-Limited (500)	7.99	0.9388	300.79	8.00	0.9845	282.72
Depth-Limited (50)	7.97	0.9175	270.27	7.86	0.9753	261.06
No Parsimony Pressure (500)	33.99	0.9658	1266.01	21.30	0.9858	691.56

(a)

	POPE 500	POPE 50	DL 500	DL 50	No Pressure
Mean Accuracy (%)	95.971	95.932	95.463	94.537	95.151
Standard Deviation	1.065	0.954	1.466	2.442	1.268

(b)

The classification accuracies (Table 1) on the testing data lend support to the hypothesis that enforcing parsimony does lead to improved generalisation performance. In particular the two POPE-GP algorithms (50 and 500 individuals) clearly outperformed all other algorithms in terms of generalisation performance. The most interesting result was the excellent generalisation performance of POPE-GP with 50 individuals. In fact, the classification accuracy of programs generated by this algorithm on the test data was statistically indistinguishable from the POPE-GP with 500 individuals, and clearly superior to the programs generated by other algorithms. This is an impressive result considering that the algorithm makes only 10% of the evaluations made by the POPE-GP with 500 individuals. Running a standard depth-limited GP with the reduced population produced poor results. This is a vindication of the hypothesis that parsimonious solutions tend to generalise better and of the approach of using multiobjective techniques for parsimony enforcement.

The excellent performance of the POPE-GP with 50 individuals warrants further investigation and analysis, but we do have some clues to why it performed as it did. Firstly, the average size of the best performing individual in this algorithm was substantially smaller — by almost 30% — than the best individuals generated by POPE-GP with 500 individuals. Also, their classification error on the testing data was significantly lower than those produced by the 500 individual algorithm. In other words, the programs were more general and 'fit' the training data less tightly, using only the strongest predictors in the underlying data for classification purposes.

References

1. Deb, K., Pratap, A., Agarwal, S. and Meyarivan,T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197,(2002)