as the values of parameters. The most important of them were set as follows: population size: 200; probability of mutation: .05; maximal number of generations: 100; training set size: 100 instances (50 images per class, randomly selected from the training subset of the MNIST database); tournament size: 5 (Goldberg et al., 1991), and $\alpha = .5$ (see Section 4.1).

In each generation, half of the population was retained unchanged, whereas the other fifty percent underwent modifications. The GP runs used the standard tournament selection based on scalar fitness function, whereas GPPO runs followed the selection procedure described in Section 4.1. Then, the offspring were created by means of the crossover operator, which randomly selects subexpressions (corresponding to subtrees in the graphical representation shown in Fig. 1) in the two parent solutions and exchanges them. The mutation operator applied to a solution randomly selects a subexpression and replaces it by other subexpression generated at random. In these operations the so-called *strong typing* principle must be obeyed (Koza, 1994).

Special precautions have been taken to prevent overfitting of hypotheses to the training data. In the GP case, the scalar fitness function was extended by additional penalty term implementing parsimony pressure. Particularly, solutions growing over 100 terms were linearly penalized with the evaluation decreasing to 0 when the threshold of 200 terms is reached. In the GPPO approach, a solution composed of 100 or more terms was always outranked, no matter how it performed on the training data.

5.4.4 PRESENTATION OF RESULTS

Table 1 presents the comparison of the best solutions (see Section 4.3) obtained in GP and GPPO runs. Table rows reflect consecutive stages of the evolution process (selected generations). Each row summarizes the comparison of 135 paired GP and GPPO runs (see Section 5.4.3). The description includes:

- the number of pairs of GP and GPPO runs (per total of 135) for which the best solution[2] evolved in GPPO yielded strictly better accuracy of classification on the training set than the best one obtained from 'plain' GP ('#GPPO BETTER'),
- the average increase of accuracy of classification of GPPO in comparison to GP ('AVERAGE INCREASE'),
- the false reject probability of Wilcoxon matched pairs signed rank test ('FALSE POSITIVE PROBABILITY'); the test takes into account the relative magnitude of differences in GP and GPPO accuracy.

[2] For both GP and GPPO, the term 'best' in this context refers to the best solution found in the evolution process, with respect to the *scalar* evaluation function, i.e. the accuracy of classification (see Section 4.3).

Table 2 presents the summary of the performance of the same solutions as in Table 1 when evaluated on an independent test set. The test set for each task contains 1600 objects, i.e. 800 images for both positive and negative classes, selected randomly from the testing part of the MNIST database. Note that the training (fitness) set and testing set are independent in a strong sense, i.e. contain digits written by another people (LeCun et al., 1995).

The tables do not refer directly to the (average) accuracy of classification, as it would not make much sense due to the heterogeneity of particular experiments (different pairs of decision classes). However, to give the reader an idea about the absolute performances of hypotheses elaborated by both algorithms, we provide the average accuracy of classification at the end of evolutionary runs (training and testing set, respectively): 90.3±6.0% and 85.2±10.2% for GP, 92.2±4.9% and 87.7±7.4% for GPPO (standard deviations included).

Table 1. Comparison of the best solutions evolved in GP and GPPO runs with respect to the accuracy of classification on the training set.

GENERATION	#GPPO BETTER	AVERAGE INCRASE [%]	FALSE POSITIVE PROBABILITY
20	74/135	0.55	.8681
40	76/135	0.93	.2880
60	89/135	1.67	.0085
80	88/135	1.46	.0096
100	105/135	1.97	.0002

Table 2. Comparison of the best solutions evolved in GP and GPPO runs with respect to the accuracy of classification on the test set.

GENERATION	#GPPO BETTER	AVERAGE INCRASE [%]	FALSE POSITIVE PROBABILITY
20	69/135	-0.02	.6234
40	82/135	1.68	.1093
60	86/135	1.92	.0461
80	89/135	1.69	.0325
100	92/135	2.63	.0061