A Genetic Algorithm for Event Selection Optimization in High Energy Physics

Giovanni Organtini [a,1]

[a] *Dip.to di Fisica - Università di Roma TRE and INFN - Sez. di Roma Via della Vasca Navale, 84 - 00146 Roma (Italy)*

A genetic algorithm approach is employed to write a program that finds the best point in the multi-parameter space of the variables used for the event selection in high energy physics experiments. In this paper I show the working principle of the algorithm and how it helps not only in finding the optimum values for cuts, but also in finding the redundant variables of the selection thus allowing for a reduction of the systematic error.

Key words: Genetic Algorithm, Event Selection, Evolutionary Algorithm.

1 Introduction

Evolutionary strategies have been used in past years for the solution of several complicated problems in various fields [1] due to their ability to solve difficult problems with a large number of degrees of freedom. Recently they have been applied to high energy physics too [2].

They simulate the mechanisms which are believed to be followed by Nature in order to perform natural evolution: **reproduction** and **mutations**.

A genetic algorithm is a way to simulate these two effects using computers. Depending on the problem to which they are applied there are several ways to implement genetic algorithms. In this paper I will show a possible application as a tool for the optimization of a selection algorithm (see also [3]).

In section 2 the working principle is presented. Section 3 is devoted to some consideration about the result of the algorithm.

[1] e-mail:*organtini@roma1.infn.it*

Preprint submitted to Elsevier Preprint *21 May 1997*

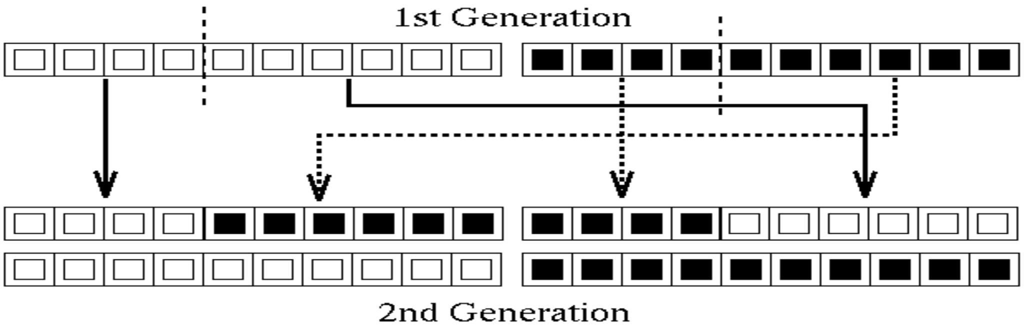

Fig. 1. an example of the reproduction mechanism with the crossing point m=4

2 A Genetic Algorithm for Event Selection

Event selection is based on cuts applied to physical variables in order to reject the maximum possible number of background events while accepting the maximum possible number of signal events.

A selection algorithm can be easily represented in a form that is suitable for the manipulation with a genetic algorithm. For each of the N variables V_i used in the selection we denote with C_i the corresponding cut.

No matter how the operations for the cut are performed and how the variables are defined, once the algorithm is given there is a one to one correspondence between the algorithm itself and an N-tuple of real numbers that represent the cuts C_i to be applied; this N-tuple can be seen as a chromosome whose genes are the cuts.

To allow for evolution, an initial random population of chromosomes is generated. A well defined selection algorithm is defined by each individual in the population; this algorithm can be applied to a given number of signal and background Monte Carlo events. We can define a score for each individual by taking, e.g., the obtained statistical significance from the application of the selection. The population is then sorted according to the score and each individual is propagated to the next generation with a probability proportional to its score.

The first two individuals are taken to generate two offsprings in the following way: let m be an integer random number between 1 and N-1. The two parents are split in two pieces: one containing genes between 1 and m and the other containing the genes between m+1 and N. The pieces are then combined in the way illustrated in fig. 1.

The same reproduction mechanism is applied to all the adjacent couples of survived individuals in the population. After each generation of new individuals, mutations can occur randomly with a given probability, their effect being the random change of one of the gene in the resulting chromosome.

These steps are iterated several times leading to a continuous evolution of the population in terms of score.

3 The Genetic Algorithm output

As a result the genetic algorithm will produce a population of several individuals sorted according to their score. The first individual will be the best one and can be used to define the optimal selection algorithm for the given channel.

This is not the only information that comes from the application of the genetic algorithm. Looking at the distributions of the genes in the population one can note that in some cases the value of a given gene can fluctuate a lot even if the score of the individuals remains the same: this is a clear indication that the cut associated to that gene is useless and may be removed from the selection algorithm. In this way its contribution to the systematic error will disappear while the efficiency of the cut remains the same (or nearly the same).

Conversely there could be genes that have a lot of influence on the score; this indicates that the corresponding cut is very powerful. Sometimes this observation can lead to a re-definition of the selection algorithm in such a way that the variable used in the cut is better studied and the selection can be re-optimized.

The genetic algorithm was applied to several physical analysis as well as to random generated distributions: results in terms of efficiency and signal to noise ratio were always better than those given by physicists performing the optimization by eye.

4 Pros and Cons of the method

In comparison with other optimizing technique, genetic algorithms have the advantage that the definition of the score and the meaning of the genes are completely free, as well as the definition of the effect of a mutation. This allow the definition of several optimization strategies. For example one can use genetic algorithms to train neural networks [4] by defining the genes as the weights to be applied to a given combination of neurons in the corresponding partition layers and the score as the output of the network.

The indications about the usefulness of the cuts used in the selection is a unique feature of the genetic technique, once used in the definition of a typical selection algorithm. Even if in this case the result can be worse in terms of efficiency with respect to neural networks, it allows for a well defined procedure for the evaluation of the systematic error of the measurement. It is well known that using neural networks makes the evaluation of systematic errors difficult, if not impossible.

This flexibility makes the method very powerful, but it makes its results dependent on the ability of the user to define genes and score and does not allow for a formal definition of which is the best choice of variables and strategy.

Moreover it is impossible to estimate *a priori* the number of generations needed for a complete optimization, neither for the value of the mutation probability. These two parameters can only be estimated after each run. The right number of iterations can be estimated looking at the distribution of the average and maximum score with the generations that should show a stable *plateau* after some iterations, while the probability of mutations must be estimated by trying with several values until the final population is heterogeneous enough to allow further evolution (if possible).

5 Conclusion

A genetic algorithm for the optimization of selection algorithms is presented. It produces selection algorithms more efficient than those defined by traditional techniques and allows the study of the usefulness of cuts, leading to a reduction of the systematic errors on the measurements.

References

[1] D.E. Goldberg, “Genetic algorithms in search, optimization and machine learning”, Addison-Wesley.

[2] M. Kunze, “Artificial Intelligence Applications in Particle Physics”, Proc. of AIHENP95, World Scientific.

[3] G. Organtini, “Using Genetics in Particle Physics”,Int. Journal of Mod. Phys. C, (6) 4 (1009) 605-610

[4] H.M.A. Andree *et al.*, “The construction of a feed-forward neural networks by means of genetic algorithms”, Proc. of AIHENP95, World Scientific.