ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 193-198

Genetic Algorithm with Crossover Based on Confidence Interval as an Alternative to Traditional Nonlinear Regression Methods

D. Ortiz Boyer*, C. Hervás Martínez*, J. Muñoz Pérez**

*Department of Computer Science, University of Córdoba
C2 Building, Rabanales Campus, 14071 Córdoba, Spain
** Department of Languages and Computer Science, University of Málaga,
Technology Complex, Teatinos Campus, 29071 Málaga, Spain

Abstract.
Most processes in the real world are controlled by nonlinear models. This explains the interest of the scientific community in the development of new methods to estimate the parameters of nonlinear models that allow the modelling of such processes. In this article we propose a new method for the estimation of parameters for nonlinear problems using genetic algorithms with real encoding (RCGA). In these genetic algorithms we use a crossover operator based on confidence intervals, which uses information from the best individuals in the population. For the resolution of this kind of problems, this operator is remarkably robust and efficient when compared with other crossover operators used in RCGAs.

1 Introduction

A usual method for the estimation of parameters for both linear and nonlinear models is the Least Squares method. It tries to find the parameters that minimize the sum of squared errors (SSE), also called sum of squared residuals. This method has to start from an initial solution, a starting point for the parameters. From this point the nonlinear regression (NLR) method is applied iteratively, trying to reduce the squared error sum until the improvement is small enough, that is, until the method converges. The choice of the method for minimization and the estimation of the starting point are very important, as both the convergence of the method and the finding of a global (not local) solution depend on them. The most widely used NLR method is Marquard's. It shows a good performance in practice, which makes it the reference method

This work has been supported by the Project ALI98-0676-CO-02 of the Spanish Comisión Interministerial de Ciencia y Tecnología

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 193-198

for NLR. Nevertheless, it can converge to a local solution, as it allows significant changes to the parameters only if they result in an improvement of the SSE.

The problems that affect classic NLR methods can be avoided by using Real-Coded Genetic Algorithms. Genetic algorithms are stochastic search algorithms based on principles of natural selection and recombination. They attempt to find the optimal solution to the problem at hand by manipulating a population of candidate solutions. The population is evaluated and the best solutions are selected to reproduce and mate to form the next generation. Over a number of generations, good traits dominate the population, resulting in an increase in the quality of the solutions.

Unlike other NLR methods, that work with a single solution, RCGAs use a population of estimations. The probability of finding an optimal global solution is thus increased. Genetic Algorithms used for the optimization of parameters whose values belong to a continuous domain use a real coding for the genes of the individuals [1]. So, each gene represents a parameter of the regression model.

The crossover operator plays a central role in GAs. In fact it may be considered to be one of the algorithm's defining characteristics, and it is one of the components to be borne in mind to improve the GA behaviour [3]. In this context, it is fundamental the capacity of crossover operators to resolve the balance between the exploration an exploitation of the search space, which is associated to the intervals determined by the extremes of domain of the genes and by the corresponding alleles of the parents chosen for crossing [2].

In this work we propose an operator based on confidence intervals (CI) whose balance between exploration and exploitation is very adequate for this kind of problems. We compare its performance and robustness with different crossover methods used in RCGAs. In order to validate the methods based on AGs, we have used NLR problems taken from the Statistical Reference Datasets Project (STRDP) developed by the staff of the Statistical Engineering Division and the Mathematical and Computational Sciences Division within the Information Technology Laboratory of the National Institute of Standards and Technology, that can be consulted on http://www.nist.gov/itl/div898/strn/nls.

2 Crossover operators based on confidence intervals

These operators are associated with the capacity of interpolation (exploitation), associated with the belonging of a population parameter to a confidence interval and of extrapolation (exploitation) derived from its not belonging to that interval. Let β be the set of the n individuals of the population and let $\beta^* \subset \beta$ be the set of the best k of them, according to their aptitude. Also, let $\beta^f = (\beta_0^f, \beta_1^f, \ldots, \beta_i^f, \ldots \beta_p^f) \in \beta$ be a chromosome selected for crossing. If we consider that the population of each one of the genes of the chromosomes of β^*

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 193-198

is distributed in a normal manner, we can define three new individuals: those formed by the lower ends (CILL), upper ends (CIUL) and means (CIM) of the confidence intervals of each gene. They are defined as follows:

$$\begin{aligned} CILL &= (CILL_0, CILL_1, \ldots, CILL_i, \ldots CILL_p) \\ CIUL &= (CIUL_0, CIUL_1, \ldots, CIUL_i, \ldots CIUL_p) \\ CIM &= (CIM_0, CIM_1, \ldots, CIM_i, \ldots CIM_p) \end{aligned} \quad (1)$$

where

$$CILL_i = \overline{\beta}_i - t_{k-1}(\alpha/2)\frac{\overline{S}_{\beta_i}}{\sqrt{k}}; \quad CIUL_i = \overline{\beta}_i + t_{k-1}(\alpha/2)\frac{\overline{S}_{\beta_i}}{\sqrt{k}}; \quad CIM_i = \overline{\beta}_i$$

being $i = 1, \ldots, k$, $\overline{\beta_i}$ the mean of each gene, $\overline{s}_{\beta_i}$ the typical quasi deviation of the k individuals of a population, $t_{k-1}(\alpha/2)$ a value obtained in Student´s t distribution tables with $k-1$ degrees of freedom and α the probability of a gene not belonging to its confidence interval.

The individuals CILL and CIUL divide each gene's domain, D_i, into three subintervals I_1, I_2 e I_3, such that

$$D_i \equiv I_1 \cup I_2 \cup I_3; \quad I_1 \equiv [a_i, CILL_i]; \quad I_2 \equiv (CILL_i, CIUL_i); \quad I_3 \equiv [CIUL_i, b_i]$$

being a_i y b_i the lower and upper limits of the domain D_i. The interval I_2 is a confidence interval built from the best k individuals in the population under the hypothesis that they are distributed following a normal distribution, and that there is a probability $1-\alpha$ of their genes' values belonging to that interval (the exploitation interval). There is, therefore, a probability α of the genes' values belonging to intervals I_1 and I_3 (exploration intervals). If we consider $\alpha = 0.5$ we shall have an equilibrium between exploring and exploiting. In addition, if k increases, for a fixed α, the amplitude of the interval will diminish and we shall have a greater exploitation since the intervals will be shorter but more selective. Analogously, if $\overline{S}_{\beta_i}$ diminishes, which is usual as the GA converges, the amplitude of the interval diminishes and we shall be able to exploit more, as the intervals will be more selective. Thus, it is necessary to assign values to both α and k that set the adequate balance between exploration and exploitation for each kind of problem.

The crossover operator proposed in this article will create, from the individual $\beta^f \in \beta$, the individuals CILL, CIUL and CIM, and their aptitudes, a single offspring β^s in the following way:

- If $\beta_i^f \in I_1$ then, if the fitness of β^f is higher than CILL then $\beta_i^s = r(\beta_i^f - CILL_i) + \beta_i^f$, else $\beta_i^s = r(CILL_i - \beta_i^f) + CILL_i$.
- If $\beta_i^f \in I_2$ then, if the fitness of β^f is higher than CIM then $\beta_i^s = r(\beta_i^f - CIM_i) + \beta_i^f$, else $\beta_i^s = r(CIM_i - \beta_i^f) + CIM_i$; figure 1b.
- If $\beta_i^f \in I_3$ then, if the fitness of β^f is higher than CIUL then $\beta_i^s = r(\beta_i^f - CIUL_i) + \beta_i^f$, else $\beta_i^s = r(CIUL_i - \beta_i^f) + CIUL_i$.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 193-198

Where r is a random number belonging to [0,1].

From this definition it is clear that the genes of the offspring always take values toward the best parent: β^f and one of CILL, CIUL or CIM, depending on the interval to which β^f belongs. If β^f is far from the other parent, the offspring will probably suffer an important change, and vice versa. Logically, the first circumstance will appear mainly in the first stages of the evolution, and the second one in the last ones.

The idea for this crossover operator is taken from the real world, where many species evolve inside highly hierarchical groups, in which only a elite of individuals mate with the rest of the population, transmitting their genes to the offspring.

3 Experiment

The RCGA used has a constant population size of 100 individuals, from which 10 percent are mutated using a non-uniform mutation with parameter b equal to 5. The probability of mutation of a gene in the selected individual is 0.5. A 60 percent of the population will be subjected to crossover using the following operators: confidence interval crossover ($k = 5, 1 - \alpha = 0.90$), arithmetical crossover ($\lambda = 0.5$), BLX-α crossover ($\alpha = 0.5$), discrete crossover, linear crossover, linear BGA crossover ($rang = 0.5$), flat crossover and Wright's heuristic crossover. A uniform selection method is used to choose the individuals for crossover and mutation, and a non-elitist tournament selection with tournament size 2 will be used to choose the individuals for the next generation. The RCGA will evolve during 5000 generations and 10 randomly independent runs will be performed for each type of RCGA.

To evaluate the performance two NLR problems have been used: BoxBOD and Thurber. Both are catalogued on the STRDP as of high complexity. The first one requires the fitting to f_1 of a set of 6 observations (fig. 1c) and the second fits to f_2 a set of 37 observations (fig. 1d), where b_i are the parameters of the models.

$$f_1 = b_1(1 - e^{-b_2 x}) + \epsilon \qquad f_2 = \frac{b_1 + b_2 x + b_3 x^2 + b_4 x^3}{1 + b_5 x + b_6 x_2 + b_7 x^3} + \epsilon$$

The SSE surfaces of the BoxBOD (fig. 1a) and Thurber problems are multimodal functions with high epistasis among their parameters. To establish the differences in performance among the different crossover operators, big domains have been used. Moreover, the different domains are not centred around the optimum value, as this could favour the convergence of some crossover methods. So, for the BoxBOD problem, $D_{b_1} \equiv [10, 1000]$ and $D_{b_2} \equiv [0.01, 1]$. And for the Thurber problem $D_{b_1} \equiv D_{b_2} \equiv [100, 10000]$, $D_{b_3} \equiv [10, 1000]$, $D_{b_4} \equiv [1, 100]$, $D_{b_5} \equiv D_{b_6} \equiv [0.01, 1]$ and $D_{b_7} \equiv [0.001, 0.1]$.

As can be seen on table 1, for the BoxBOD problem all the crossover operators converge to the optimal solution, except the arithmetical, discrete and linear, although their differences are minimal. As for the speed of convergence (fig.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 193-198

10^3	Opt. S.	C. Int.	Aritmet.	BLX-α	Discrete	Linear	L. BGA	Flat	Wright
BoxBOD	1.168008	*1.168008*	1.168010	*1.168008*	1.168241	1.168010	*1.168008*	*1.168008*	*1.168008*
Thurber	5.642708	*5.655013*	5.708023	5.811555	7.685453	6.270148	6.016960	5.678330	6.754761

Table 1: $SSE.10^3$ for BoxBOD and Thurber where the second column shows the optimal minimum and the following ones the average values obtained in 10 runs for each type of crossover.

1e), the confidence interval crossover is the best one, although the convergence speed is high. The difference among the different crossover operators is not great. The cause is that, even though this problem is listed as complex due to its epistasis and its multimodal character (fig. 1a), the fact that it only has two parameters makes it very easy to tackle by any kind of RCGA, even using wide domains. On the other hand, NLR methods appear vulnerable to this problem.

For the Thurber problem, the best results are obtained with the confidence interval crossover. Its SSE (table 1) is very close to the optimum and the convergence speed (fig. 1f) is the highest. From the results it can be deduced that this problem is very complex for a RCGA owing to its epistasis, multimodality and to the fact that the number of dimensions of the search space is high. This makes that a slight increment in the domain of the parameters affects decisively the complexity of the problem, often making it unfit for classic NLR methods.

4 Conclusion

In this article it has been proved that RCGAs based on confidence interval crossover operators are an alternative to classic NLR methods. Specially in problems where there is no a-priori information that allows to choose an initial solution close to the optimum. The expansion of the search space makes the RCGAs, which perform a statistical parallel search, with crossover based on confidence intervals which balances the exploration and exploitation of the search, the best method for the modelling of this kind of problems.

References

[1] D. E. Goldberg. Real-coded genetic algorithms, virtual alphabets, and blocking. *Complex Systems*, (5):139–167, 1991.

[2] F. Herrera, M. Lozano, and J. L. Verdegay. Tackling real-coded genetic algorithms: Operators and tools for behavioural analysis. *Artificial Inteligence Review*, (12):265–319, 1998. Kluwer Academic Publisherr. Printed in Netherlands.

[3] G. E. Liepins and M. D. Vose. Characterizing crossover in genetic algorithms. *Annals of Mathematics and Artificial Intelligence*, (5):27–34, 1992.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 193-198

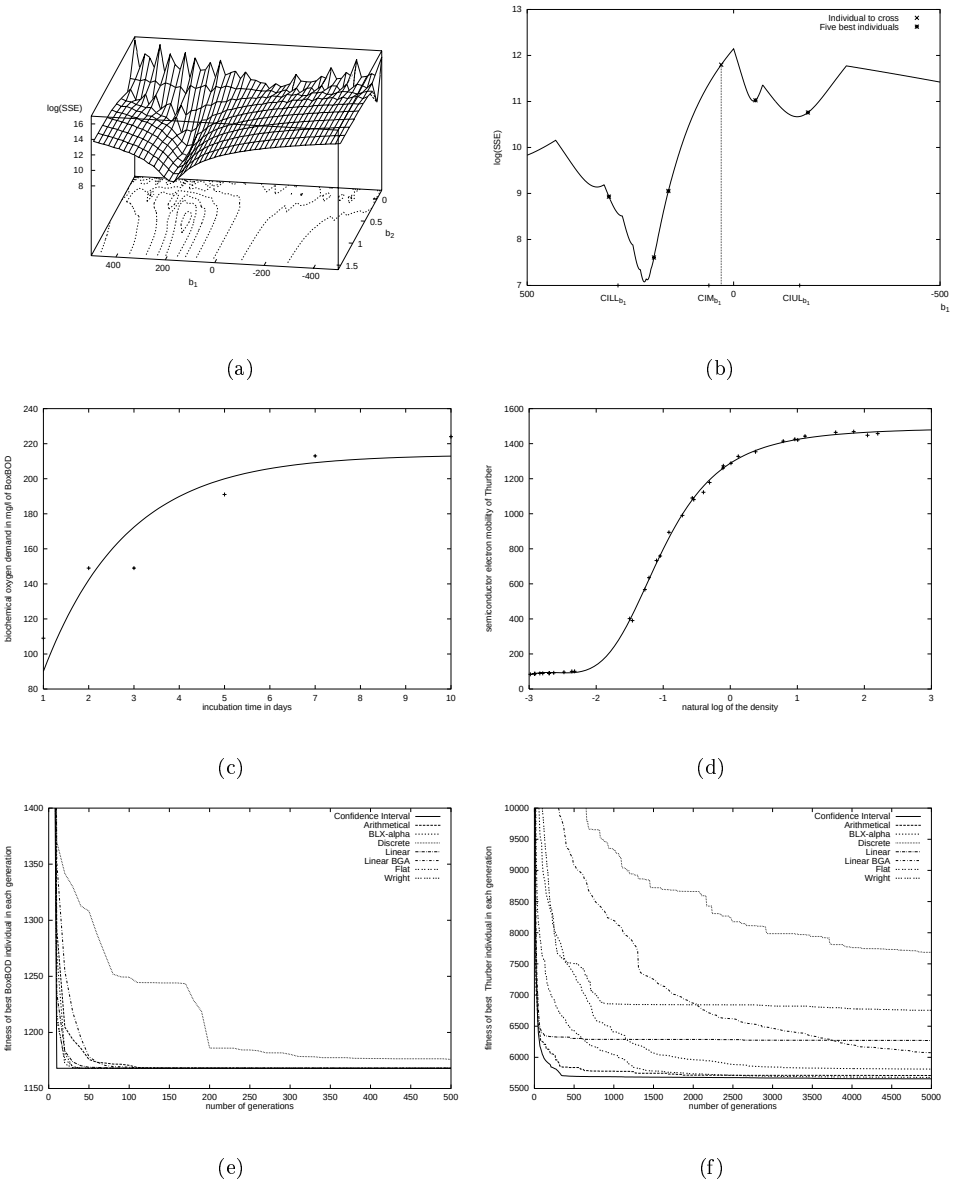

Figure 1: (a) log(SSE) surface for BoxBOD where b_1 and b_2 represent the model's parameters; (b) Cut of BoxBOD parallel to b_1 showing graphically the philosophy of confidence interval crossover; (c) NLR model of BoxBOD; (d) NLR model of Thurber; (e) Average aptitude of the best individual in 10 runs for BoxBOD using different crossovers; (f) Idem for Thurber.