Application of Crossover Operators Based on Confidence Interval in Modeling Problems Using Real-Coding Genetic Algorithms*

Rafael del Castillo Gomariz, César Hervás Martínez, Sebastián Ventura Soto, and Domingo Ortiz Boyer

Department of Computing and Numerical Analysis, University of Córdoba
14071 Córdoba, Spain
grupo@ayrna.org

Abstract. In this work we develop and compare multi-parent crossover operators based on the extraction of characteristics from the best individuals in the population (average, median, standard deviation and quantiles). These statistics evolve in parallel with the algorithm. The proposed operators are used in combination with a real-coded genetic algorithm for the evolution of polynomial functions to solve microbial growth problems. Their performance is compared to other crossover operators for real-coded genetic algorithms. Both the prediction errors made in the modelling of systems and the objectivity and speed in the identification of models show the viability of this type of models that mix base functions with evolutionary computation.

1 Introduction

Nowadays, the modelling of systems is one of the most interesting problems in many scientific branches. The resolution of this problem has been classically approached by using regression techniques in order to minimize an error function, over a model type previously established by the researcher. Most often the functional model to apply is non-linear and it usually presents a high dimensionality, making the process considerably more complicate, as there is scarce additional information, or none at all.

The most common approximation functions are linear and generalized linear models, flattened hyperplanes, response surfaces, artificial neural networks, Fourier series, wave functions, decision trees and flattened kernel functions. All of them provide explicit models for the relationship between the predictive variables x and the response variable y [2] (only one in our case).

In this work we present an methodology for the estimation of Response Surface (RS) polynomial models through Real-Coded Genetic Algorithms (RCGA) using specific real-coded crossover operators: BLX-α [3], an adaptation of BLX-α, and the recently developed multi-parent crossovers CIXL and CIXL2 [4].

In Section 2 we introduce multiparent crossover operators. Section 3 presents a general approach to the evolution of RS models with RCGAs and a crossover operator

* This work has been funded by the spanish Ministry of Science and Technology, MCyT, through Project TIC 2002-04026-C02, and by FEDER funds.

R. Conejo et al. (Eds.): CAEPIA-TTIA 2003, LNAI 3040, pp. 177–186, 2004.

adapted from BLX-α. Section 4 shows the results of a equality of means test over two factors: grade of the initial RS and type of crossover used. Conclusions are drawn and presented in Section 5.

2 Crossover Algorithms Based on Confidence Intervals

In the resolution of RS polynomial models, the multi-parent crossover operators give to the RCGA the additional value of being able to use information from several individuals to create a new one, with a better fitness if possible. We present in this section a type of multi-parent crossover algorithm based on the location and dispersion characteristics of the genes from the best individuals in the population. These characteristics will be used to build virtual parents that inherit the traits associated to the previous estimators.

The aforementioned idea leads to the definition of two crossover operators based on Confidence Intervals using the norms L_2 (CIXL2) and L_1 (CIXL1), whose equilibrium between exploration and exploitation seems to be very suitable for this type of problems. Their performance in model identification problems has been made clear in [6], where they have been applied to non-linear regression problems taken from "the Statistical Reference Datasets Projects (STRDP)" which can be consulted in http://www.nist.gov/itl/div898/strn/nls.

2.1 Intervals Associated to Median and Mean as Location Parameters of the Genes

Let β be the set of the n individuals in the population and let $\beta^* \subset \beta$ be the set formed by the best n individuals (the ones with highest fitness). If we consider that the genes β_i of the chromosomes β^* are independent random variables following a continuous distribution function H (β_i), with a location parameter μ_{β_i}, then we have the model $\beta_i = \mu_{\beta_i} + e_i$, being e_i a random variable, for each i=1,... p.

If we suppose, for each i, that the best n individuals actually form a simple random sample ($\beta_{i1}, \beta_{i2}, \ldots \beta_{in}$) of the β_i distribution then the model takes the form

$$\beta_{ij} = \mu_{\beta_i} + e_{ij}, \text{ for } j=1,..., n \quad (1)$$

Now, from the model proposed in (1), if we consider the norm L_1, given by $\|\beta_i\|_1 = \sum_{j=1}^{n} |\beta_{ij}|$, and we look for an estimator of μ_{β_i} associated with the negative gradient method, that is, S1 (μ_{β_i}) = -dD_1 (μ_{β_i})/dμ_{β_i}, where the dispersion function induced by the norm L_1 is D_1 (μ_{β_i})=$\sum_{j=1}^{n} |\beta_{ij} - \mu_{\beta_i}|$, and we define H as the distribution function of the β_i, then we have that the negative gradient estimator of the location

parameter through the norm L_1 is the median of the β_i distribution. That is, $\mu_{\beta_i} = M_{\beta_i}$ being its distribution binomial with parameters *n* and ½. From this distribution we can already build confidence intervals for the location parameter, populational median, whose estimator is the sample median M_{β_i} of the genes of the *n* better individuals, for a generic sample of size *n*, with a confidence coefficient *1-α*. In this case we apply the Neyman method for calculating confidence intervals and we have that

$$I_{1-\alpha}(\mu_{\beta_i}) = [\beta_{i(k+1)}, \beta_{i\,(n-k)}], \tag{2}$$

being $\beta_i(k+1)$ and $\beta_i(n-k)$ the values of the genes associated to the position *k*+1 and *n-k* when the sample has been sorted, and where the *k* value is determined from the underlying binomial distribution.

If we take into consideration the norm L_2, defined as $\|\beta_i\|_2 = \sum_{j=1}^{n} \beta_{ij}^2$, it can be proved that the negative gradient estimator of the location parameter through the norm L_2 is the average of the distribution β_i. Assuming that the distribution of the genes $H(\beta_i)$ is normal, the confidence interval is calculated as:

$$I_{1-\alpha}(\mu_{\beta i}) = [\bar{\beta}_i - t_{n-1,\alpha/2} \times \bar{S}_{\beta_i} / \sqrt{n}\,;\, \bar{\beta}_i + t_{n-1,\alpha/2} \times \bar{S}_{\beta_i} / \sqrt{n}\,] \tag{3}$$

where t_n*-1* is a Student t distribution with *n-1* degrees of freedom.

From the previous confidence intervals we build 3 virtual parents: one formed by all the lower limits (*CILL)*, other formed by all the upper limits (*CIUL)* and a third one (*CIM)* formed by the average (if using CIXL2) or median (if using CIXL1) values of the confidence intervals of each gene. The *CILL* and *CIUL* individuals divide the domain of each gene, D_i, into 3 subintervals I_i^L, I_i^{IC} and I_i^R, such that $D_i \equiv I_i^L \cup I_i^{IC} \cup I_i^R$, $I_i^L \equiv [a_i, CILL_i]$, $I_i^{IC} \equiv (CILL_i, CIUL_i)$ and $I_i^R \equiv [CIUL_i, b_i]$, being a_i and b_i the lower and upper limits of the domain D_i

The crossover operators will create a single offspring β^s from the individual $\beta^f \in \boldsymbol{\beta}$, the individuals *CILL, CIUL* and *CIM*, and their fitnesses, in the following way:

- If $\beta_i^f \in I_i^L$ then, if the fitness of β^f is bigger than that of *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 $\beta_i^f \in I_i^{IC}$ then, if the fitness of β^f is bigger than that of *CIM* ,then $\beta_i^s = r(\beta_i^f - CIM_i) + \beta_i^f$, else $\beta_i^s = r(CIM_i - \beta_i^f) + CIM_i$;
- If $\beta_i^f \in I_i^R$ then, if the fitness of β^f is bigger than that of *CIUL*, then $\beta_i^s = r(\beta_i^f - CIUL_i) + \beta_i^f$, else $\beta_i^s = r(CIUL_i - \beta_i^f) + CIUL_i$,

where *r* is a random number in the interval [0,1].

3 Estimation and Design of Polynomial Base Functions with RCGA

In general, the modelling of a system whose equation is known is a problem of conventional regression. In this type of problems there is a functional relationship between a series of independent variables x_i and a dependent variable y, in the form:

$$y = f(\beta_0, \beta_1, ..., \beta_m, x_1, ..., x_n) \tag{4}$$

where β_i are the coefficients to be adjusted in order to minimize the sum of squared residuals. This optimization problem can be solved with a classical algorithm, or with a genetic algorithm. If we opt for the second option we would codify the individual as a set of genes, each one representing a coefficient.

3.1 Response Surface Models and Genetic Algorithms

Response surface models explain a large variety of phenomena. The expression that defines them is a grade G polynomial in each variable [5][8]. Therefore they are functions following the form

$$f(x_1, x_2, ..., x_n) = c_0 + \sum_{i=1}^{n} c_i x_i + ... + \sum_{\substack{i_1, i_2 ... i_G = 1 \\ i_k \le i_{k+1}}}^{n} c_{i_1 i_2 ... i_G} x_{i_1} x_{i_2} ... x_{i_G} \tag{5}$$

where G is the grade of the model, x_i are the independent variables, n is the number of independent variables and β_i are the coefficients.

If we want to model the structure of a phenomenon by using the aforementioned model, we will codify individuals with as many genes as the coefficients in the model that we pretend to develop. This number of coefficients depends on the number of variables and the grade of the model in question but, as we already mentioned, the interpretability of the models is a very desirable characteristic in every type of modeling, and that leads us to seek simple models. The codification of an individual uses one gene for each coefficient of the model. However, this gene has two well-differentiated parts. On one hand, there is an allele which indicates the presence or absence of the corresponding term (monomial) in the model; and on the other hand there is another allele to codify the value of the coefficient in question.

Figure 1 shows an individual that represents a grade 2 RS with 3 variables, adapted to this method. To obtain expressions with the minimum number of terms, we include one term in the fitness function that rewards the smaller (i.e. simpler) models. In this way, our problem turns into a problem with two objectives: on one hand, it is convenient that the error is minimum, but, on the other hand, it is also interesting to obtain models with a small number of coefficients.

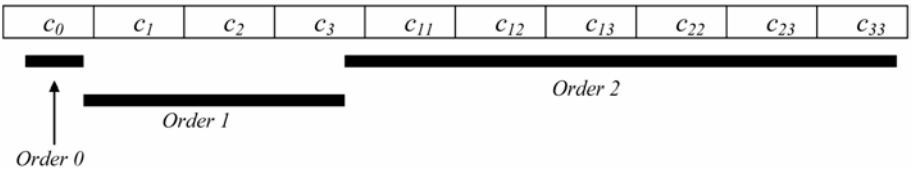

Fig. 1. Individual that represents a response surface of order 2 with 3 independent variables.

Since the number of objectives is very reduced, we did not consider a multi-objective algorithm and to simplify we chose a fitness function that calculates a linear combination of them, weighing up their importance with a coefficient.

3.2 Microbian Growth Models

In this work we develop a model for the growth prediction of the altering microorganism *Leuconostoc mesenteroides ssp. mesenteroides* [9], which has been frequently isolated as a responsible for the alteration of different types of meat products. The available data have been 210 signal-time curves of Leuconostoc Mesenteroides growth under different conditions of temperature T (10.5, 14, 17.5, 21 and 24º C), pH (5.5, 6, 6.5, 7 and 7.5), sodium chloride concentration NaCl (0.25, 1.75, 3.25, 4.75 and 6.25%) and sodium nitrite concentration NaNO2 (0, 50, 100, 150 and 200 ppm). These 210 curves correspond to 30 different experimental conditions chosen according to a Composite Central Design[1] of experiments. From each one of these 30 conditions, 7 experiment replicas were conducted. Five of the experimental results sets were chosen at random to form the training set, and the remaining two formed the generalization set. Thus the training set is composed of 150 curves and the generalization (or test) one by 60.

Next, these resultant values of absorbancy, considered throughout the time, were adjusted by means of an exponential Baranyi and Roberts-type model [1] with the help of the DMFit 1.0 program (József Baranyi, Institute of Food Research, Norwich Research Park, Norwich NR4 7UA, UK). The results were the training and generalization values of the kinetic growth parameters lnlag, grate and yend (the logarithm of the adaptation phase, the growth rate and the maximum density) of the microorganism for each experimental condition.

3.3 Genetic Algorithm

Table 1 summarizes the parameters used for the genetic algorithm. The fitness function presents two terms; the first one represents the error term (according to the minimization of the squared residuals sum) and the second one represents the complexity term of the model (according to the minimization of the number of coefficients).

The first term is a transformation of the standard error of prediction (%SEP), an adimensional coefficient of the form:

$$SEP = \frac{100}{\bar{y}} \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}, \tag{6}$$

where y_i represents the value of the function in this point, $\hat{y}_i$ is the estimated value and, $\bar{y}$ the mean value of all the y_i. The second term modulates linearly the number of terms in the expression, growing as the number n_T of terms decreases. In this way, the fitness expression would be:

[1] This type of experiments design focuses the sampling on the central values of the experimental variables.

$$A = (1-\alpha)(1-\frac{SEP}{K}) + \alpha\left(1-\frac{n_T - n_{Tm}}{n_{TM} - n_{Tm}}\right) \quad (7)$$

where the coefficients n_{Tm} and n_{TM} represent respectively the minimum and maximum number of coefficients that the model can represent, and the constant *K*, whose values ($1 \leq K < \infty$) are obtained heuristically, modulates the SEP value to weight the differences among patterns in order to get equilibrium between the two objectives.

Table 1. Parameters used in the genetic algorithms for the modelling of the response surfaces.

GENERAL ASPECTS OF THE ALGORITHM			
Population size		500 individuals	
	Duplication	p_d=0.2	Tournament selection
	Crossover	p_c=0.6	Tournament selection BLX-α *(α=0.5)* CIXL1 and CIXL2 (*1-α= 0.7, n=5*)
	Mutation	p_m=0.2	Random selection Non uniform mutation (parameter *b= 5*)
Stop criterium		500 generations	

This nondecreasing fitness function takes a maximum value of 1, which could be possible only if the standard error of prediction was null and the model had t_m terms.

The number of genes forming each individual of the population will depend on the grade of the chosen RS for the model. The crossover operators used in the genetic algorithm have been the BLX-α crossover [3], three adaptations of this crossover to this problem, and the multi-parent crossovers CIXL1 and CIXL2 [4]. The mutation operator has been the Non Uniform. These operators, specific for the real coding, have been adapted to be able to work with the double codification previously explained.

All the algorithms have been implemented in Java using Sun Microsystems Java Development Kit version 1.3.1, and the JCLEC class library for evolutionary computation [10]. The analysis of variance for the comparison of means has been performed using the statistics software SPSS version 11.0.

3.4 Adaptation of the BLXα Crossover

We have designed an adaptation of the BLX-α operator (even though it could have been any other arity-2 crossover designed for RCGA). Let $\beta^1 = \{(s_1^1, c_1^1), ..., (s_i^1, c_i^1), ..., (s_p^1, c_p^1)\}$ and $\beta^2 = \{(s_1^2, c_1^2), ..., (s_i^2, c_i^2), ..., (s_p^2, c_p^2)\}$ be two parents chosen for crossover, with *p* genes each one and representing two RS models with *p* coefficients. Each gene corresponds to a monomial in the corresponding RS, and each allele represents, respectively a selector that indicates the presence or absence of the monomial in the model and the value of the coefficient associated to the term. These two

parents will generate two offsprings $\beta^{h1} = \{(s_1^{h1}, c_1^{h1}),...,(s_i^{h1}, c_i^{h1}),...,(s_p^{h1}, c_p^{h1})\}$ and $\beta^{h2} = \{(s_1^{h2}, c_1^{h2}),...,(s_i^{h2}, c_i^{h2}),...,(s_p^{h2}, c_p^{h2})\}$. The genes of the best parent will be inherited more likely than those of the other, so that each gene $\left(s_i^{h1}, c_i^{h1}\right)$ and $\left(s_i^{h2}, c_i^{h2}\right)$ will have the following values:

If round $\left(s_i^1\right)$ = round $\left(s_i^2\right)$

then $s_i^{h1} \leftarrow s_i^{h2} \leftarrow$ round $\left(s_i^1\right)$

(c_i^{h1}, c_i^{h2}) ← application of BLXα over the alleles (c_i^1 and c_i^2)

else *apt1* ← fitness of parent β^1 ; *apt2* ← fitness of parent β^2

n_1 and n_2 ← two random integers with a probability *apt*1/(*apt*1+*apt*2) of taking value 1 and probability *apt2*/(*apt*1+*apt*2) of taking value 2

$(s_i^{h1}, c_i^{h1}) \leftarrow (s_i^{n_1}, c_i^{n_1})$; $(s_i^{h1}, c_i^{h1}) \leftarrow (s_i^{n_2}, c_i^{n_2})$

EndIf

That is, the generated offspring will inherit the terms existing in both parents and the BLXα crossover will be applied to the coefficients of these terms. When the terms exist only in one of the parents, the more fit this parent is compared to the other, the more possibilities of passing to the offspring they will have.

4 Results

We have searched for an optimal topology as well as the coefficients of the model using RS of grades 2 to 5, so as to check if our methodology is able to find models for different topologies previously used. It means that the size of the weight space to estimate increases exponentially with the grade of the starting polynomial. For each one of the 3 growth parameters we have analyzed if there are significative differences in the mean values of the generalization SEP according to the grade of the polynomial (RS2 to RS5) and according to the four types of crossover operators used (BLXα, BLXαAD1, CIXL1 and CIXL2). An test for equality of means has been performed, taking into account both intrapopulation variances and interpopulation variances. For each cell of the ANOVAII model, 30 runs were performed[2], using the parameters discussed in the previous section, and we can affirm with a significance level of 99% that for the three performed analysis, one for each parameter of the growth curves:

1. There are significative differences in the variances associated to each cell (Sig=0.000). There are significative differences in the averages: according to the interaction between the grade of the starting polynomial and the type of crossover used (Sig=0.000), according to the polynomial grade (Sig=0.000) and according to the crossover type (Sig=0.000). There are not significant differences when starting with RS2 or RS3, but they appear when starting with RS4 or RS5.

[2] Currently, more tests are being executed to confirm the results.

Table 2. Statistic results summary (Mean, Standard Deviation) in the three experiments.

Parameter/SR Crossover	lnlag / SR3		grate / SR2		yend / SR2	
	Mean	StdDev	Mean	StdDev	Mean	StdDev
BLX-alfa	7.92	0.52	15.51	4.01	15.71	0.51
BLXAD1	8.15	1.51	15.62	2.19	15.87	0.69
CIXL1	7.25	0.55	11.79	2.11	15.58	0.39
CIXL2	12.55	2.59	21.17	6.67	21.32	8.88

$$\text{lnlag} = 1.8585 - 0.2366(T) - 0.0938(pH) + 0.273(NaCl) + 0.1029(NaNO_2) + 0.0374(pH)^2 - 0.1294(pH)(NaCl) - 0.0569(pH)(NaNO_2) - 0.0923(pH)(NaCl)(NaNO_2) \quad (8)$$

$$\text{Grate} = 0.1802 + 0.0718(T) + 0.0250(T)^2 + 0.0206(NaCl)^2 - 1.9315(T)^2(NaNO_2) - 10.4226(pH)^2(NaNO_2) + 0.0071(pH)(NaCl)^2 - 0.0102(NaCl)^3 + 12.3471(NaCl)^2(NaNO_2) \quad (9)$$

$$\text{yend} = -0.6844 + 0.1522(T) - 0.2222(NaCl) - 0.2437(NaNO_2) + 0.0591(pH)(NaCl) - 0.0427(NaCl)^2 + 0.1510(T)^2(pH) + 4.3559(T)^2(NaCl) + 0.0301(pH)(NaNO_2) + 0.0186(pH)^3 + 5.2791(pH)^2(NaCl) - 0.017(pH)(NaCl)(NaNO_2) - 9.6772(NaCl)(NaNO_2)^2 \quad (10)$$

2. In the models where the dependent variable is *lnlag*:
- The RS of grade 3 produces the best total results (for the six crossovers) in mean (8.97). The means and variances are shown in Table3.
- There are no significative differences in mean between the crossovers BLX and CIXL1. But there are between them and the crossover CIXL2. The CIXL1 crossover is the one that produces the best results in mean starting with a grade 2 polynomial and especially grade 3. These differences are significant if we eliminate the results of the crossovers CIXL2.
- We concluded using polynomials of grade 3 and the CIXL1 crossover. In this way the statistic results of the 30 proofs are shown in Table 3. The best model chosen according to %SEP and smaller number of parameters is shown in equation 8.

3. In the models where the dependent variable is *grate*:
- The RS of grade 2 produces the best total results (for the six crossovers) in mean (16.02). The means and variances are shown in Table 3.
- There are no differences in mean between the crossovers BLX and CIXL1. But there are between them and the crossover CIXL2. The CIXL1 crossover is the one that produces the best results in mean starting with a grade 2 polynomial, being these differences significant if we eliminate the results of the crossover CIXL2
- We concluded using polynomials of grade 2 and CIXL1 crossover. In this way, the statistic results of the 30 proofs are shown in Table 3. The best model chosen according to %SEP and smaller number of parameters is shown in equation 9.

Table 3. Data of the best models obtained in the three experiments with CIXL1.

	Lnlag	grate	yend
%SEP Training	7.17	8.51	-11.88
%SEP Test	7.05	9.22	-13.49
Num. coefficients	9	9	13
Fitness	0.69	0.65	0.48

4. In the models where the dependent variable is *yend*:
- The RS of grade 2 produces the best total results (for the six crossovers) in mean (17.12). The means and variances of the %SEP for RS2 and the 6 crossovers are shown in Table 3.
- There are no differences in mean between the crossovers BLX and CIXL1. But there are between them and the crossover CIXL2. The crossover CIXL1 is the one that produces the best results in mean starting with a grade 2, but these differences are not significant even though we eliminate the results of the crossover CIXL2.
- We concluded using polynomials of grade 2 and CIXL1 crossover. In this way, the statistic results of the 30 proofs are shown in Table 3. The best model chosen according to %SEP and smaller number of parameters is shown in equation 30.

5 Conclusions

We have experimentally demonstrated how, starting from overdimensionated response surface models, our methodology finds the model fitting the phenomenon's response. A fitness function has been proposed that considers the quadratic relative errors and weights up the simplicity of the model as well. This makes the expressions evolve until they present a minimum size, improving their interpretability through the decrease of the number of terms in the polynomial function and their capacity of generalization. A specialized genetic algorithm has been implemented, using a double codification and specifically adapted operators. These operators, CIXL1 and CIXL2, make possible to extract the adaptive statistic characteristics of the best individuals and to use them to lead the search in the most effective way. In particular we checked that with the CIXL1 crossover operator we obtained better results, proposing its use in this type of algorithm. This procedure represents an advantage over the use of statistic tests to eliminate coefficients and identify the model exactly, much more tedious and, in some cases, biased by the subjective appreciations of the investigator. We also checked that with this algorithm it is possible to reach better results than the ones obtained with non-linear regression.

References

1. Baranyi, J., Roberts, T. A.: 1994. A dynamic approach to predicting bacterial growth in food. Int. J. Food Microbiol., Vol. 23 (1994) 277-294
2. Denison, D., Holmes, C., Mallick, B., Smith, A.: Bayesian Methods for Non-linear Classification and Regression. John Wiley & Sons, Chichester (2002)
3. Eshelman, L.J., Schaffer, J.D.: Real-coded genetic algorithms and interval-schemata. Foundations of Genetic Algorithms, Vol. 21, Morgan Kaufmann (1993) 187-202
4. Hervás, C., Ortiz, B., García, N.: Theoretical Análisis of the Confidence Interval Based Crossover for Real-Coded Genetic Algorithms. Parallel Problem Solving from Nature PPSN VII, 2439, Springer-Verlag, Granada (2002) 153-161
5. Myers Raymond, H.M., Montgomery, D. C.: Response Surface Methodology: Process and Product Optimisation Using Designed Experiments. Second Edition, John Wiley & Sons, New York (2002)

6. Ortiz, D., Hervás, C., Muñoz, J.: Genetic algorithm with crossover based on confidence intervals as an alternative to least squares estimation for non-linear models. Metheuristic International Congress, Porto (2001)
7. Ortiz, D.: Operadores de cruce basado en intervalos de confianza en algoritmos genéticos con codificación real, Tesis Doctoral, Málaga (2001)
8. Rawlings, J.O., Pantula, S.G., Dickey, D.: Applied regression analysis: A research tool. Springer-Verlag, New York (1998)
9. Rodríguez Pérez, R.: Elaboración de modelos predictivos de crecimiento microbiano de lactobacilus plantarum.Tesis Doctoral. Departamento de Bromatología y Tecnología de los alimentos. Universidad de Córdoba (2003)
10. Ventura, S., Ortíz, D., Hervás, C. JCLEC. Una librería de clases Java para Computación Evolutiva. Congreso Español de Algoritmos Evolutivos y Bioinspirados (2002)