Operadores de cruce por intervalo de confianza en problemas de modelado utilizando algoritmos genéticos con codificación real

R. del Castillo Gomariz, C. Hervás Martínez, S. Ventura Soto, D. Ortiz Boyer

Departamento de Informática y Análisis Numérico. Universidad de Córdoba
14071- Córdoba- España
grupo@ayrna.org

Resumen[1]. En el trabajo presentamos operadores de cruce multipadre basados en la extracción de las características estadísticas tanto de centralización, media y mediana, como de dispersión, desviación estándar y cuantiles, de los mejores individuos de la población, las cuales evolucionan conjuntamente con el algoritmo. Estos operadores se utilizan asociados a un algoritmo genético con codificación real de funciones polinómicas para resolver problemas de crecimiento microbiano, comparando su eficacia con otros operadores de cruce para algoritmos genéticos con codificación real. Tanto los errores de previsión cometidos en el modelado de sistemas, como la robustez del algoritmo, muestran la viabilidad de este tipo de modelos de funciones de base utilizando computación evolutiva.

1 Introducción

El modelado de sistemas es en la actualidad uno de los problemas de mayor interés en numerosas ramas de la ciencia. La resolución de este problema se ha abordado clásicamente usando técnicas de regresión para minimizar una determinada función de error, previo establecimiento por parte del investigador del tipo de modelo a aplicar. En la mayoría de los casos, el modelo funcional a aplicar es no lineal y además suele presentar una alta dimensionalidad, lo que complica considerablemente el proceso, teniendo en cuenta que se dispone de poca o ninguna información adicional.

Las funciones de aproximación más comunes son los modelos lineales y generalizados lineales, hiperplanos alisados, superficies de respuesta, redes neuronales artificiales, series de Fourier, funciones de ondas, árboles de decisión y funciones de núcleo alisadas. Todas ellas, proporcionan modelos explícitos para la relación existente entre las variables predictoras **x** y en nuestro caso una sola variable de respuesta **y**.

[1] Este trabajo ha sido financiado por la CICYT en el proyecto TIC 2002-04026-C02 y con Fondos Feder

En este trabajo presentamos una metodología de estimación de modelos polinómicos de superficie de respuesta mediante Algoritmos Genéticos con Codificación Real (AGCR) donde se utilizan operadores de cruce específicos para codificación real, el cruce BLX-α (Eshelman, L. J. and Schaffer, J. D. 1993) junto con una adaptación del mismo y los cruces multipadre CIXL1 y CIXL2 (Hervás, C., Ortiz, D., García, N. 2002), desarrollados recientemente.

De esta forma en la sección 2 presentamos los operadores de cruce multipadre, que dan paso en la sección 3 a un planteamiento general de los modelos de superficie de respuesta junto con el algoritmo genético de codificación real y un operador de cruce adaptado al BLX estándar. En la sección 4 presentamos los resultados de los contrastes de igualdad de medias en función de dos factores: grado de la superficie de respuesta de partida y tipo de cruce utilizado. Concluimos en la sección 5.

2 Algoritmos de cruce basados en intervalos de confianza

En la resolución de modelos polinómicos de superficie de respuesta, los operadores de cruce multipadres aportan al AGCR un valor añadido al poder utilizar información de varios individuos para formar uno nuevo, a ser posible de mejor aptitud. Presentamos por tanto en esta sección un tipo de algoritmo de cruce multipadre basado en las características de localización y dispersión de los genes de los mejores individuos de la población, que se utilizarán para construir padres virtuales que hereden las características asociadas a los anteriores estimadores.

La idea anterior se concreta en la definición de dos operadores de cruce basados en Intervalos de Confianza usando la norma L_2 (CIXL2) y L_1 (CIXL1), cuyo equilibrio entre exploración y explotación se muestra muy adecuado para este tipo de problemas. Su rendimiento para este tipo de problemas ha sido puesto de manifiesto en (D. Ortiz, C. Hervás y J. Muñoz, 2001a) (D. Ortiz., C. Hervás y J. Muñoz, 2001b), en los que se ha aplicado a problemas de regresión no lineal tomados de "the Statistical Reference Datasets Project (STRDP)" y que pueden ser consultados en http://www.nist.gov/itl/div898/strn/nls.

2.1 Intervalos asociados a la mediana y media como parámetros de localización de los genes

Sea ß el conjunto de los n individuos que forman la población y sea ß*⊂ ß el conjunto formado por los n mejores individuos (en el sentido de tener una mayor aptitud). Si consideramos que los genes β_i de los cromosomas de ß* son variables aleatorias independientes con función de distribución continua $H(\beta_i)$, y con un parámetro de localización de la forma μ_{β_i}. Entonces tenemos el modelo $\beta_i = \mu_{\beta_i} + e_i$, siendo e_i una variable aleatoria, para cada i= 1,...,p.

Si suponemos, para cada i, que los n mejores individuos forman en realidad una muestra aleatoria simple ($\beta_{i1}, \beta_{i2}, \ldots \beta_{in}$) de la distribución de los β_i, entonces el modelo toma la forma

$$\beta_{ij} = \mu_{\beta_i} + e_{ij}, \quad \text{para } j= 1,...,n \tag{1}$$

Ahora, a partir del modelo propuesto en (1), si consideramos la norma L_1, definida en la forma $\|\beta_i\|_1 = \sum_{j=1}^{n} |\beta_{ij}|$, y buscamos un estimador de μ_{β_i} asociado al método del gradiente negativo, esto es $S1(\mu_{\beta_i}) = -dD_1(\mu_{\beta_i})/d\mu_{\beta_i}$, donde la función de dispersión inducida por la norma L_1 es $D_1(\mu_{\beta_i}) = \sum_{j=1}^{n} |\beta_{ij} - \mu_{\beta_i}|$, y definimos H como la función de distribución de los β_i, entonces tenemos que el estimador de gradiente negativo del parámetro de localización mediante la norma L_1 es la mediana de la distribución de β_i [16]. Esto es $\hat{\mu}_{\beta_i} = M_{\beta_i}$ .siendo su distribución binomial de parámetros n y ½. A partir de esta distribución ya podemos construir intervalos de confianza para el parámetro de localización, mediana poblacional, cuyo estimador es la mediana M_{β_i} muestral de los genes de los n mejores individuos, para una muestra genérica de tamaño n, con un coeficiente de confianza 1-α. En este caso aplicamos el método de Neyman, de cálculo de intervalos de confianza tenemos que

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

siendo $\beta_i(k+1)$ y β_i (n-k) los valores de los genes asociados a la posición k+1 y n-k una vez ordenada la muestra, y donde el valor de k se determina a partir de la distribución binomial subyacente.

Si consideramos la norma L_2, definida en la forma $\|\beta_i\|_2 = \sum_{j=1}^{n} \beta_{ij}^2$, se demuestra que el estimador de gradiente negativo del parámetro de localización mediante la norma L_2 es la media de la distribución de β_i. Bajo la hipótesis de que la distribución de los genes $H(\beta_i)$ es normal, el intervalo de confianza se formula como:

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

donde t_n-1 es una distribución t de student con n-1 grados de libertad.

A partir de los intervalos de confianza anteriores construimos tres padres virtuales: el formado por todos los extremos inferiores, CILL, superiores, CIUL, y medias (CIXL2) o medianas (CIXL1), CIM, de los intervalos de confianza de cada gen. Los individuos CILL y CIUL dividen el dominio de cada gen, D_i, en tres subintervalos I_i^L, I_i^{IC} e I_i^R, tal que $D_i \equiv I_i^L \cup I_i^{IC} \cup I_i^R$, $I_i^L \equiv [a_i, CILL_i]$, $I_i^{IC} \equiv (CILL_i, CIUL_i)$ y $I_i^R \equiv [CIUL_i, b_i]$ siendo a_i y b_i los extremos inferiores y superiores del dominio D_i

Los operadores de cruce crearán, a partir del individuo $\beta^f \in \boldsymbol{\beta}$, de los individuos CILL, CIUL y CIM, y de sus aptitudes, un único hijo β^s de la siguiente forma:

- Si $\beta_i^f \in I_i^L$ entonces, si la aptitud de β^f es mayor que la de CILL entonces $\beta_i^s = r\left(\beta_i^f - CILL_i\right) + \beta_i^f$, sino $\beta_i^s = r\left(CILL_i - \beta_i^f\right) + CILL_i$.
- Si $\beta_i^f \in I_i^{IC}$ entonces, si la aptitud de β^f es mayor que la de CIM entonces $\beta_i^s = r\left(\beta_i^f - CIM_i\right) + \beta_i^f$, sino $\beta_i^s = r\left(CIM_i - \beta_i^f\right) + CIM_i$.
- Si $\beta_i^f \in I_i^R$ entonces, si la aptitud de β^f es mayor que la de CIUL entonces $\beta_i^s = r\left(\beta_i^f - CIUL_i\right) + \beta_i^f$, sino $\beta_i^s = r\left(CIUL_i - \beta_i^f\right) + CIUL_i$.

donde *r* es un numero aleatorio perteneciente al intervalo [0,1].

3 Estimación y diseño de funciones de base polinómica con AGCR

En general, el modelado de un sistema cuya ecuación conocemos es un problema de regresión convencional. En este tipo de problemas, existe una relación funcional entre una serie de variables independientes x_i y una variable dependiente y, en la forma:

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

donde β_i son los coeficientes que hay que ajustar, para que minimicen la suma de residuos al cuadrado. Este problema de optimización puede resolverse con un algoritmo clásico, o con un algoritmo genético. Si optamos por lo segundo codificaríamos el individuo como un conjunto de genes, cada uno de los cuales representaría a un coeficiente.

3.1 Modelos de Superficie de Respuesta y Algoritmos Genéticos.

Los modelos de superficie de respuesta son un tipo de modelos que explican una amplia variedad de fenómenos. La expresión que los define se ajusta a un polinomio de grado G en cada una de las variables objeto de estudio (Rawlings, J. O. et al, 1998, Myers, R. H., Montgomery D. C. 2002) Se trata, por tanto, de funciones de la forma:

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

donde G es el grado del modelo, x_i cada una de las variables independientes, n es el número de variables independientes y β_i cada uno de los coeficientes.

Si queremos diseñar la estructura de un fenómeno utilizando el modelo anteriormente expuesto, codificaremos individuos con tantos genes como coeficientes tenga el modelo que pretendamos desarrollar. Este número es función del número de variables disponibles y del grado del modelo en cuestión, pero como ya hemos comentado anteriormente la interpretabilidad de los modelos es una característica muy deseable en cualquier tipo de modelado la cual nos lleva a una mayor

simplicidad. La codificación de un individuo presenta un gen por cada uno de los coeficientes del modelo. Sin embargo, este gen está formado por dos partes bien diferenciadas. Por una parte, hay un alelo que indica la presencia o no del monomio en el modelo a desarrollar; además, existen alelos para codificar el valor del coeficiente en cuestión.

La Figura 1 muestra un individuo que representa un modelo de superficie de respuesta de grado 2 con tres variables, adaptado a esta metodología. Para poder conseguir expresiones con un número mínimo de términos hay que incluir un término en la función de aptitud que premie los modelos con mayor simplicidad. De este modo, nuestro problema se convierte en un problema con dos objetivos: por una parte, es conveniente que el error sea mínimo pero, por otra, es también interesante obtener modelos con un menor número de coeficientes.

Dado que el número de objetivos es muy reducido, hemos optado por incluir una única función de aptitud que realiza una combinación lineal de los mismos, ponderando la importancia de éstos con un coeficiente.

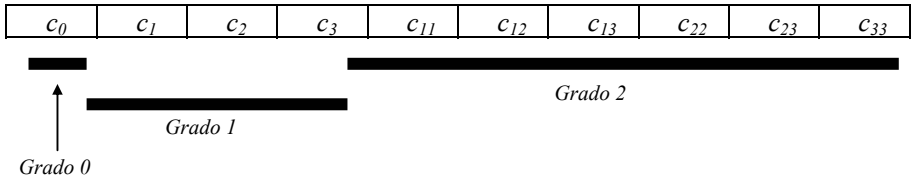

Fig. 1. Individuo que representa a una superficie de respuesta de grado 2 con 3 variables independientes.

3.2 Modelos de crecimiento microbiano

Se ha elaborado un modelo predictivo de crecimiento del microorganismo alterante *Leuconostoc mesenteroides ssp. mesenteroides* de este grupo BAL (Rodríguez R. 2003), el cual ha sido frecuentemente aislado como microorganismo responsable de alteración en diversos tipos de productos cárnicos. Los datos utilizados han sido 210 curvas, señal-tiempo, de crecimiento microbiano de Leuconostoc Mesenteroides bajo diferentes condiciones de temperatura T (10.5, 14, 17.5, 21 y 24ºC), pH (5.5, 6, 6.5, 7 y 7.5), concentraciones de cloruro sódico, NaCl (0.25, 1.75, 3.25, 4.75 y 6.25%) y nitrito sódico, $NaNO_2$ (0, 50, 100, 150 y 200 ppm). Estas 210 curvas se corresponden con 30 condiciones experimentales diferentes según un diseño de experimentos del tipo Diseño Central Compuesto. De estas 30 condiciones se realizaron 7 réplicas para todas las temperaturas de incubación incluidas en el diseño experimental, de ellas cinco, elegidas al azar, se tomaron para formar el conjunto de entrenamiento y las otras dos para formar el conjunto de generalización, de esta forma el conjunto de entrenamiento está formado por 150 curvas y el de generalización o test por 60.

A continuación, estos valores de absorbancia resultantes, considerados a lo largo del tiempo, fueron ajustados mediante un modelo de tipo exponencial de Baranyi y Roberts (1994) con la ayuda del programa DMFit 1.0 (József Baranyi, Institute of Food

Research, Norwich Research Park, Norwich NR4 7UA, UK). Como resultado se obtuvieron los valores de entrenamiento y generalización de los parámetros cinéticos de crecimiento lnlag, grate e yend (el logaritmo de la tasa o velocidad de crecimiento, la fase de adaptación y la densidad máxima) del microorganismo para las condiciones experimentales estudiadas.

3.3 Algoritmo genético.

La Tabla 1 resume los parámetros que se han empleado. La función de aptitud presenta dos términos, el primero representa el término de error (en función de la minimización de la suma de residuos al cuadrado) y el segundo el término de complejidad del modelo (en función de la minimización del número de coeficientes).

El primer término es una transformación del error estándar de predicción (%SEP), un coeficiente adimensional que viene dado por la siguiente expresión:

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

donde y_i representa el valor de la función en ese punto, $\hat{y}_i$ es el valor estimado e $\bar{y}$ el valor medio para los valores de los y_i. El segundo término modula linealmente el número de términos en la expresión, siendo tanto mayor cuanto menor sea el número de términos n_T. De este modo, la expresión de la aptitud sería:

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

donde los coeficientes n_{Tm} y n_{TM} representan, respectivamente, el número de coeficientes mínimo y máximo que puede presentar el modelo, y la constante K permite modular el valor del SEP para que se exalten las diferencias entre patrones de cara a conseguir un equilibrio entre los dos objetivos.

Esta función de aptitud es creciente, tomando un valor máximo de 1 que sólo se daría si el error estándar de predicción fuese nulo y el modelo tuviese n_{Tm} términos.

El número de genes para cada individuo de la población dependerá del grado de la

Tabla 1. Parámetros empleados en los algoritmos genéticos para el modelado de superficies de respuesta.

ASPECTOS GENERALES DEL ALGORITMO			
Tamaño población		500 individuos	
Operadores	*Duplicación*	p_d=0.2	Selección por torneo
	Cruce	p_c=0.6	Selección por torneo BLX-α (α=0.5) CIXL1 y CIXL2 (1-α= 0.7, n=5)
	Mutación	p_m=0.2	Selección aleatoria Mutación No Uniforme (parámetro b= 5)
Criterio parada		500 generaciones	

superficie de respuesta escogida para hacer la búsqueda del modelo. Los operadores de cruce utilizados en el algoritmo genético han sido el cruce BLX-α (Eshelman, L. J. and Schaffer, J. D. 1993), tres adaptaciones a este problema de este cruce y los cruces multipadre CIXL1 y CIXL2 (C. Hervás y col 2002). La mutación utilizada ha sido la No Uniforme. Estos operadores, específicos de la codificación real, han sido adaptados para poder trabajar con la doble codificación mencionada anteriormente.

Todos los algoritmos se han implementado en Java utilizando la versión 1.3.1 del kit de desarrollo Java de Sun Microsystems, y la librería de clases para computación evolutiva JCLEC (Ventura, S., Ortiz, D. y Hervás, C. 2002). El análisis de varianza para la comparación de medias se ha realizado utilizando el software estadístico SPSS 11.0.

3.4 Adaptación del cruce BLXα

Hemos diseñado una adaptación del operador BLXα (aunque podría haber sido cualquier otro cruce de aridad 2 diseñado para AGCR). Supondremos dos padres: $\beta^1 = \{(s_1^1, c_1^1),...,(s_i^1, c_i^1),...,(s_p^1, c_p^1)\}$ y $\beta^2 = \{(s_1^2, c_1^2),...,(s_i^2, c_i^2),...,(s_p^2, c_p^2)\}$ elegidos para ser cruzados, con p genes cada uno y representando a dos modelos de superficie de respuesta con p coeficientes. Cada gen se corresponde a un término (monomio) de la superficie de respuesta y cada alelo representa, respectivamente, a un selector que indica la presencia o ausencia del término en el modelo y el valor del coeficiente asociado al término. Estos dos padres generarán dos hijos $\beta^{h1} = \{(s_1^{h1}, c_1^{h1}),...,(s_i^{h1}, c_i^{h1}),...,(s_p^{h1}, c_p^{h1})\}$ y $\beta^{h2} = \{(s_1^{h2}, c_1^{h2}),...,(s_i^{h2}, c_i^{h2}),...,(s_p^{h2}, c_p^{h2})\}$. Haremos que el material genético del mejor de los dos padres tenga mayor probabilidad de ser heredado por los hijos que el material del otro padre, de manera que cada gen $\left(s_i^{h1}, c_i^{h1}\right)$ y $\left(s_i^{h2}, c_i^{h2}\right)$ tendrá los siguientes valores:

Si $\text{ROUND}\left(s_i^1\right) = \text{ROUND}\left(s_i^2\right)$

Entonces $s_i^{h1} \leftarrow s_i^{h2} \leftarrow \text{ROUND}\left(s_i^1\right)$

$(c_i^{h1}, c_i^{h2}) \leftarrow$ aplicación BLXα sobre los alelos (c_i^1 y c_i^2)

Sino *apt1* ← aptitud del padre β^1

apt2 ← aptitud del padre β^2

n_1 y n_2 ← dos enteros aleatorios con probabilidad *apt*1/(*apt*1+*apt*2) de tomar valor 1 y probabilidad *apt2*/(*apt*1+*apt*2) de tomar valor 2

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

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

FinSi

Esto es, los hijos generados heredaran los términos que existan en ambos padres y a los coeficientes de estos términos se les aplicará el cruce BLXα. En el caso de los términos que existan sólo en alguno de los padres, éstos tendrán más posibilidades de pasar a los hijos cuanto más apto sea este padre respecto del otro.

4 Resultados

Hemos buscado la topología óptima así como los coeficientes del modelo usando superficies de respuesta de grados 2, 3, 4 y 5; para comprobar de esta manera si nuestra metodología era capaz de encontrar el modelo para diferentes topologías utilizadas a priori. Esto significa que el espacio de pesos a estimar aumenta de forma exponencial cuando aumenta el grado del polinomio del que se parte. Para cada uno de los tres parámetros de crecimiento hemos analizado si existen diferencias significativas en los valores medios de SEP de generalización en función del grado del polinomio (SR2 a SR5) y en función de los cuatro tipos de operadores de cruce utilizados (BLXα, BLXαAD1, CIXL1 y CIXL2) hemos realizado un test de igualdad de medias teniendo en cuenta tanto las varianzas dentro de cada población como entre las poblaciones normales consideradas. Para cada celda del modelo ANOVAII se han realizado 10 ejecuciones, utilizando los parámetros expuestos en la sección anterior y podemos afirmar con un nivel de significación del 99% que para los tres análisis efectuados, uno para cada parámetro de las curvas de crecimiento:

1. Existen diferencias significativas en las varianzas asociadas a cada celda (Sig=0.000). Existen diferencias significativas en las medias: en función de la interacción del grado del polinomio de partida y del tipo de cruce utilizado (Sig.= 0.000), en función del grado del polinomio (Sig=0.000) y en función del tipo de cruce (Sig=0.000). No existen diferencias significativas entre empezar con una SR2 o SR3. Si existen diferencias, si empezamos con SR4 o SR5.
2. En los modelos donde la variable dependiente es el *lnlag*:
   - El grado SR3 produce los mejores resultados totales (para los seis cruces) en media (9.00). Las medias y varianzas del %SEPG para SR3 y los seis cruces se muestran en la Tabla 3
   - No existen diferencias en media entre los cruces tipo BLX y CIXL1. Si existen entre estos y el cruce CIXL2. El Cruce CIXL1 es el que produce mejores resultados en media empezando con un polinomio de grado 2 y sobre todo de grado 3. Siendo estas diferencias significativas si eliminamos los resultados del cruce CIXL2
   - Concluimos en utilizar como grado del polinomio el grado 3 y como cruce CIXL1. De esta forma los resultados estadísticos de las 10 pruebas se muestran en la Tabla 3. El mejor modelo elegido en cuanto a %SEP y menor número de parámetros se muestra en la ecuación 8.
3. En los modelos donde la variable dependiente es el *grate* cabe decir que:
   - El grado SR2 produce los mejores resultados totales (para los seis cruces) en media (16.44). Las medias y varianzas del %SEPG para SR2 y los seis cruces se muestran en la Tabla 3
   - No existen diferencias en media entre los cruces tipo BLX y CIXL1. Si existen entre estos y el cruce CIXL2. El Cruce CIXL1 es el que produce mejores resultados en media empezando con un polinomio de grado 2, siendo estas diferencias significativas si eliminamos los resultados de CIXL2
   - Concluimos en utilizar como grado del polinomio el grado 2 y como cruce CIXL1. De esta forma los resultados estadísticos de las 10 pruebas se muestran

en la Tabla 3. El mejor modelo elegido en cuanto a %SEP y menor número de parámetros se muestra en la ecuación 9.

4. En los modelos donde la variable dependiente es el *yend* cabe decir que:
   - El grado SR2 produce los mejores resultados totales (para los seis cruces) en media (16.83). Las medias y varianzas del %SEPG para SR2 y los seis cruces se muestran en la Tabla 3
   - No existen diferencias en media entre los cruces tipo BLX y CIXL1. Si existen entre estos y el cruce CIXL2. El Cruce CIXL1 es el que produce mejores resultados en media empezando con un polinomio de grado 2, pero estas diferencias no son significativas aunque eliminemos los resultados del cruce CIXL2
   - Concluimos en utilizar como grado del polinomio el grado 2 y como cruce CIXL1. De esta forma los resultados estadísticos de las 10 pruebas se muestran en la Tabla 3. El mejor modelo elegido en cuanto a %SEP y menor número de parámetros se muestra en la ecuación 10.

Tabla 2. Resumen resultados estadísticos (Media, Desviación Típica) en los tres experimentos

Parámetro/SR Cruce	lnlag / SR3 Media	Des.T	grate / SR2 Media	Des.T	yend / SR2 Media	Des.T
BLX-alfa	7.89	0.57	15.47	3.88	15.79	0.57
BLXAD1	8.08	1.43	15.67	2.29	15.93	0.65
CIXL1	7.34	0.47	11.99	2.06	15.66	0.43
CIXL2	12.70	2.76	21.28	6.80	21.50	8.90

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

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

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

Tabla 3. Datos de los mejores modelos obtenidos en los tres experimentos con CIXL1

	lnlag	grate	yend
%SEP Entrenamiento	7.23	8.49	-11.97
%SEP Test	7.14	9.13	-13.61
nº coeficientes	9	9	13
Aptitud	0.67	0.62	0.46

5 Conclusiones

Hemos comprobado experimentalmente como, en primer lugar partiendo de modelos de superficies de respuesta sobredimensionados, nuestra metodología

encuentra el modelo al que responde el fenómeno. Se ha propuesto una función de aptitud que, además de considerar los errores cuadráticos relativos, pondera la simplicidad del modelo, lo que conduce a que las expresiones evolucionen hasta presentar un tamaño mínimo, mejorando su interpretabilidad y su capacidad de generalización. Se ha implementado un algoritmo genético real con una doble codificación en el que se usan operadores específicos adaptados, estos operadores CIXL1 y CIXL2, abren la posibilidad de extraer las características estadísticas adaptativas de los mejores individuos y poder utilizarlas para guiar la búsqueda de forma más eficiente y eficaz. En concreto hemos comprobado que con el operador de cruce CIXL1 es con el que se obtienen mejores resultados, proponiendo su utilización en este tipo de algoritmo. Este procedimiento representa una ventaja frente al uso de tests estadísticos para eliminar coeficientes e identificar el modelo con exactitud, mucho más tedioso y, en algunos casos, sesgado por apreciaciones subjetivas del investigador. Hemos comprobado además que con este algoritmo se pueden alcanzar resultados que mejoran a los obtenidos mediante regresión no lineal (donde hay que conocer la forma exacta del modelo).

7 Bibliografía

(1) Rawlings, J. O.; Pantula, S. G.; Dickey, D. *Applied regression analysis: A research tool*; Springer-Verlag: New York , 1998.
(2) Myers Raymond H. M. and Montgomery D. C. Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Second Edition: John Wiley & Sons. New York. 2002
(3) L.J. Eshelman and J.D. Schaffer. Real-coded genetic algorithms and interval-squemata. Whitley, L.D. ed. Foundations of Genetic Algorithms21, 187-202. Morgan Kaufmann, 1993.
(4) D. Ortiz, C. Hervás and J. Muñoz. Genetic algorithm with crossover based on confidence intervals as an alternative to traditional nonlinear regression methods. European Symposium in Artificial Neural Networks, Brujas. 2001.
(5) D. Ortiz, C. Hervás and J. Muñoz. Genetic algorithm with crossover based on confidence intervals as an alternative to least squares estimation for nonlinear models. Metaheuristic International Congress, Oporto. 2001.
(6) D. Ortiz, "Operadores de cruce basado en intervalos de confianza en algoritmos genéticos con codificación real", Tesis Doctoral, Málaga 2001.
(7) R. Rodríguez Pérez. Elaboración de modelos predictivos de crecimiento microbiano de lactobacilus plantarum...Tesis Doctoral. Departamento de Bromatología y Tecnología de los alimentos. Universidad de Córdoba. 2003.
(8) C. Hervás Martínez, D. Ortiz Boyer, N. García Pedrajas, Theoretical Análisis of the Confidence Interval Based Crossover for Real-Coded Genetic Algorithms. Parallel Problem Solving from Mature PPSN VII, 2439, 153-161 Springer-Verlag. Granada. Sep 2002.
(9) Baranyi, J. y Roberts, T. A. 1994. A dynamic approach to predicting bacterial growth in food. Int. J. Food Microbiol., 23, 277-294.
(10) Ventura, S., Ortíz, D. and Hervás, C. JCLEC. Una librería de clases Java para Computación Evolutiva. Primer Congreso Español de Algoritmos Evolutivos y Bioinspirados. 2002.