UNIVERSIDADE DA CORUÑA

Correlación Epistática: un estudio de simulación

María Teresa Iglesias Otero (totero@udc.es), Manuel Antonio Presedo Quindimil (mpresedo@udc.es)
Departamento de Matemáticas. Universidade da Coruña.

La *epistasis* [1] es una medida estadística de la dificultad que una función presenta para ser optimizada mediante un algoritmo genético. *Grosso modo*, la epistasis mide la distancia de la función que se desea optimizar a la clase de *funciones de primer orden*: funciones libres de interacciones entre los *genes* (variables binarias, en el caso clásico). Entre las medidas empleadas para clasificar la dificultad de las funciones, la epistasis resulta interesante por su bajo coste computacional, aunque posee un inconveniente: no utiliza información sobre la dinámica de los algoritmos genéticos.
En este trabajo se considera la *correlación epistática,* que calculamos para varias funciones estándar. Realizamos un estudio de simulación donde analizamos la correlación epistática como medida de la dificultad que presenta una función a ser optimizada por un algoritmo genético.

❑ Introducción

La *epistasis* es una medida estadística de la dificultad que una función presenta para ser optimizada por un algoritmo genético (A.G.). Analizar las causas que dificultan el hallazgo del óptimo de una función empleando para ello algoritmos genéticos es un punto clave en el desarrollo de la teoría de este tipo de algoritmos.
Las funciones engañosas (*deceptive functions* en terminología de los AA.GG.) fueron las primeras candidatas a ser consideradas como funciones difíciles. Sin embargo, en 1992, Grefenstette ([2]) muestra un ejemplo de función, ---caso extremo de engaño (*fully deceptive function*)--- que no presenta dificultad alguna para ser procesada por un A.G.
Paralelamente, Davidor ([1]) introduce la noción de *epistasis* que deriva de su homónimo en Genética. En este contexto, un gen se dice epistático a otro gen cuando el primero enmascara el efecto (fenotípico) del segundo. Adaptado este concepto al campo de los AA.GG., la epistasis mide la cantidad de linealidad de la función (generalmente codificada en binario) $f\colon \Omega = \{0,1\}^{\ell} \to \mathbb{R}$, que se desea optimizar.
En la década de los noventa del siglo pasado, la epistasis ha sido profusamente estudiada pues es una medida interesante, comparada con otras, porque su cálculo requiere un coste computacionalmente bajo (lo cual es una condición necesaria para llenar el vacío entre la teoría y la práctica de los AA.GG). Sin embargo, los resultados de esta última década han arrojado luz sobre alguno de sus inconvenientes. Quizá el más importante es que es una medida "estática": no tiene en cuenta la dinámica de un A.G.
Dado que la *epistasis dinámica* es la diferencia entre la función y su valor previsto sobre muestras de poblaciones de un tamaño fijo, en este trabajo calculamos la correlación epistática para varias funciones estándar ([3], [5]), de las que es conocida su epistasis estática o global. Tomamos como medida de la dificultad, de cada una de esas funciones, el número medio de generaciones necesarias por un algoritmo genético clásico para alcanzar el óptimo y analizamos el poder de la *correlación epistática* para predecir el comportamiento del algoritmo. Observamos que, en el caso de las funciones Template, la epistasis dinámica ordena las funciones, según su dificultad, en la misma forma que lo hace la epistasis estática: a mayor grado de dificultad de la función menor correlación. Además, en todas las funciones, alta correlación epistática implica facilidadpara los AA.GG., lo que corrobora otros estudios tales como [3], [4] o [6], por ejemplo.

❑ Epistasis

Fijemos una función f en el espacio de búsqueda $\Omega = \{0,1\}^{\ell}$. Dada una muestra P en Ω, el *valor medio de idoneidad* de la muestra viene dado por
$f(P) = \Sigma s \in P\ f(s)/|P|$, donde $|P|$ denota la cardinalidad de P (se permiten elementos repetidos). El *valor de exceso de idoneidad de una cadena* $s=s_1 \cdots s_{\ell} \in P$ se define como $f(s) - f(P)$. Si denotamos por $P_{i,a}$ el subconjunto de P formado por todas las cadenas que en la posición i tiene el valor $a \in \{0,1\}$, Entonces

el i-ésimo valor medio *alélico* es $A_{i,P}(s_i) = \frac{\sum_{t \in P_{i,s_i}} f(t)}{|P_{i,s_i}|}$. Y el *valor de exceso génico* es $E_P(s) = \sum_{i=1}^{\ell} [A_{i,P}(s_i) - f(P)]$.
Una predicción de la idoneidad de una cadena s se puede obtener como $A_P(s) = f(P) + E_P(s)$. La diferencia $\varepsilon_P(s) = f(s) - A_P(s)$ puede verse como una medida de la epistasis de la cadena s. Reuniendo las definiciones anteriores en una única fórmula, tenemos:

$$\varepsilon_P(s) = f(s) - \sum_{i=1}^{\ell} \frac{1}{|P_{i,s_i}|} \sum_{t \in P_{i,s_i}} f(t) + \frac{\ell - 1}{|P|} \sum_{t \in P} f(t).$$

En el caso especial de que $P = \Omega$ tenemos la **epistasis global** (***estática***):

$$\varepsilon(s) = f(s) - \sum_{i=1}^{\ell} \frac{1}{2^{\ell - 1}} \sum_{t \in \Omega_{i,s_i}} f(t) + \frac{\ell - 1}{2^{\ell}} \sum_{t \in \Omega} f(t).$$

En este trabajo usamos, como estimador de la dificultad que el AG tiene para optimizar una función, la correlación epistática entre los valores, obtenidos en la muestra, de la función a estudio $f(s)$ y los valores previstos $A_P(s)$.

❑ Funciones utilizadas en el estudio

Para este estudio hemos utilizado las nueve funciones binarias del cuadro 1.
Todas ellas son funciones clásicas en las evaluaciones de AA.GG., como las dos primeras mencionadas en [5] y [7] o las conocidas funciones patrón (*Template*). Para una función patrón, T^n, el valor de una cadena binaria de longitud ℓ, $\ell \geq n$, se corresponde con el número de veces que aparece un patrón fijo de bits de longitud n. Por conveniencia, supongamos que el patrón de longitud n es la subcadena de n unos 1^n = 11…11. Entonces, $T^2(1^{\ell}) = \ell - 1$, por ejemplo.
En particular, la función T^1 es la función definida por Ackley en 1987 ([1]), que a su vez es un ejemplo de función "sumadora de unos" (*unitation* en la terminología de los AA.GG.).
Es razonable esperar que al aumentar la longitud del patrón también aumente la epistasis de la función, pues la dificultad para maximizar la función usando un AG aumenta. De hecho, así sucede con la epistasis global (normalizada). En [3] se calcula explícitamente y se obtiene que

$$\varepsilon(T^n) = \begin{cases} 1 - \dfrac{1 + n(\ell - n + 1)^2 + \frac{(\ell - n)}{3}(4 - (\ell - n)^2)}{2^n(3(\ell - n) - 1 + 2^{n - \ell + 1})}, & si\ n \leq \ell \leq 2n \\ 1 - \dfrac{(n^2 + \ell + 2)(\ell - 2n) + \frac{n}{3}(2n^2 + 7) + 2n^2 + 1}{2^n(3(\ell - n) - 1) + (\ell - 2n)^2 + 2(n + 1) - \ell}, & si\ 2n \leq \ell \end{cases}$$

La función de Wilson, F_W, es estándar en la teoría del engaño (*deceptiveness*) ([5], [8]). F_W está definida sobre cadenas binarias de longitud 3m (m >>1) así:

$F_W(x_1, x_2, x_3, \ldots, x_{3m-2}, x_{3m-1}, x_{3m}) = f_W(x_1, x_2, x_3) + f_W(x_4, x_5, x_6) + \ldots + f_W(x_{3m-2}, x_{3m-1}, x_{3m})$

donde,

$f_W(0,0,0) = 2$, $f_W(0,0,1) = f_W(0,1,0) = f_W(1,0,0) = 0$, $f_W(0,1,1) = f_W(1,0,1) = f_W(1,1,0) = 1$ y $f_W(1,1,1) = 3$.

Muchas de las funciones que se encuentran en la literatura para analizar el comportamiento de los algoritmos genéticos son funciones separables (un ejemplo de función separable es F(x, y, z) = G(x) + G(y) + G(z); las funciones F1 o F4 son ejemplo de ello). La comparación exclusiva entre funciones separables puede llevar a conclusiones erróneas. Una forma de introducir interacciones no lineales es considerar funciones escalables no separables, de las que la función de Wilson es un ejemplo característico. Otra función no separable es F_2.

Función	Intervalo de x_i	bits en los que se codifica
$F_1(x_1,x_2,x_3) = x_1^2 + x_2^2 + x_3^2$	[-5.12, 5.11]	30 bits (10 cada x_i)
$F_2(x_1,x_2) = 100(x_1^2 - x_2)^2 + (1 - x_1)^2$	[-2.048, 2.047]	24 bits (12 cada x_i)
F_W: F. *Wilson*	---	30 bits (m = 10)
Template T^1 (*Ackley's Onemax*)	---	6 bits
Template T^2	---	6 bits
Template T^3	---	6 bits
Template T^4	---	6 bits
Template T^5	---	6 bits
Template T^6	---	6 bits

Cuadro 1. Funciones utilizadas en la simulación. (Nótese que las dos primeras corresponden a problemas de minimización y nosotros las hemos cambiado por sus opuestas).

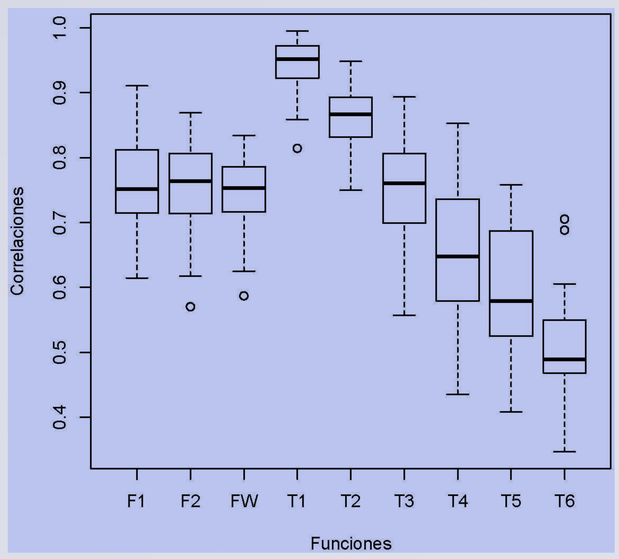
❑ Simulación

Como se ha indicado en la introducción, como indicador de la dificultad de las funciones se ha considerado, **N**, el número medio de generaciones necesarias (en cien ejecuciones) para que un algoritmo genético clásico encuentre el máximo por primera vez. Concretamente, hemos utilizado un SGA (algoritmo genético simple) con muestras de tamaño 20 –suficiente para garantizar la representatividad de cada valor binario en cada posición--, con selección por torneo binario, cruce de un punto con probabilidad 0.7 y mutación con probabilidad 0.02. En cada ejecución se han empleado treinta generaciones.

Función	Epistasis normalizada (estática)	N
$F_1(x_1,x_2,x_3) = x_1^2 + x_2^2 + x_3^2$	0.2105	66.63
$F_2(x_1,x_2) = 100(x_1^2 - x_2)^2 + (1 - x_1)^2$	0.4785	42.94
Fw	0.0739	12.61
Template T^1 (*Ackley's Onemax*)	0.0000	4.47
Template T^2	0.1042	9.60
Template T^3	0.3333	17.62
Template T^4	0.5795	26.27
Template T^5	0.7708	45.25
Template T^6	0.8906	78.94

Cuadro 2. Valores de la epistasis normalizada (estática) de las funciones a estudio

El siguiente gráfico muestra los diagramas de cajas, calculados a partir de 100 correlaciones epistáticas, para cada una de las nueve funciones incluidas en la simulación.
A efectos de comparación, debemos diferenciar el grupo formado por las tres funciones (F1, F2 y FW) de las restantes. Pues, para estas tres primeras funciones, podemos observar como las correlaciones no permiten diferenciar su grado de dificultad.
En cuanto a las funciones *Template*, la correlación epistática las ordena en la misma forma que lo hace la epistasis normalizada (medida estática de la dificultad): a mayor dificultad menor correlación.



❑ Conclusiones

A tenor de los resultados experimentales obtenidos, parece que la epistasis dinámica no es suficiente para medir la dificultad que presentan las funciones a ser optimizadas por un AG. No obstante, considerada sobre funciones definidas por un número limitado de parámetros, como las funciones Template o la función de Wilson, se comporta como lo hace el AG sobre esas funciones. Los resultados experimentales obtenidos aquí con las funciones F1 y F2 quizá se deban a que la epistasis no es capaz de diferenciar funciones de orden superior a 1 y se necesita complementar la información que proporciona la epistasis ordinaria con otras de orden superior.

Referencias

[1] Ackley, D. H. A connectionist machine for genetic hillclimbing. MA: Kluwer Academic, 1987.
[2] Davidor, Y. Epistasis Variance: A Viewpoint on GA-Hardness, In G.J.E. Rawlins (Ed.), Foundations of Genetic Algorithms, 23--35. Morgan Kaufmann, 1991.
[2] Grefenstette, J. J. Deception considered harmful. In L. D. Whitley (Ed.) Proceedings of the second workshop on Foundations of Genetic Algorithms, 75--91. Morgan Kaufmann, 1992.
[3] Iglesias M. T., Verschoren, A., Vidal, C., Computing epistasis of Template functions through Walsh transforms. Computing and Informatics, Vol. 24, 263--279, 2005.
[4] Soule, T. and Foster, J. A. Genetic algorithm hardness measures applied to the maximum clique problem. In T. Baeck (Ed.), Proceedings of the Seventh International Conference on Genetic Algorithms, 81--88. Morgan Kaufmann, 1997.
[5] Rochet, S., Venturini, G., Slimane, M., Kharoubi, E. E. A critical and empirical study of epistasis measures for predicting GA performances: a summary. In J. K. Hao et al, (Ed.), Articial Evolution 97, vol. 1363 of LNCS, 275--286. Springer-Verlag, 1998.
[6] Weinreich, D. M., Lan, Y., Wylie, C. S. and Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? Current Opinion in Genetics & Development, Genetics of system biology, vol. 23, Issue 6, 700--707. Elsevier, 2013.
[7] Whitley, D., Mathias, K., Rana, S., Dzubera, J. Building better test functions. In L. J. Eshelman (Ed.) Proceedings of the Sixth International Conference on Genetic Algorithms, 239-246. Morgan Kaufmann, 1995.
[8] Wilson, S. W. GA_easy does not imply steepest-ascent optimizable. In R.K. Belew and L.B. Booker (Eds) proceedings of the Fourth International Conference on Genetic Algorithms, 85-89. Morgan Kaufmann, 1991.