COMPUTACIÓN BIOLÓGICA

Pedro Isasi[1]

[1]Departamento de Informática
Universidad Carlos III de Madrid
Avda. de la Universidad, 30. 28911 Leganés (Madrid). Spain
email: isasi@ia.uc3m.es

Presentación

Temario

Temario

Temario

Temario

Temario

Temario

Temario

¿Por qué funcionan?

- Los AGs no procesan estrictamente individuos, sino similitudes entre ellos
- Patrones de similitud entre individuos o esquemas
- Dado que cada individuo encaja en muchos patrones a la vez, la eficiencia de la búsqueda se multiplica. Paralelismo Implícito
- Se utiliza como herramienta matemática los esquemas. Teorema del Esquema

Esquemas

- El esquema es una herramienta para estudiar la forma en que una cadena representa a otras cadenas
- Es un patrón de similitud que describe un subconjunto de cadenas con similitudes en ciertas posiciones.
- Consideramos el alfabeto $0, 1, *$:
  - Un esquema recoge una determinada cadena si encada posición del esquema un 1 corresponde a un 1 en la cadena, un 0 a un 0, y el $*$ se corresponde concualquier cosa
  - Por ejemplo, el esquema $*0000$ encaja con las dos cadenas: $00000, 10000$

- Símbolos del alfabeto: 0, 1, #, donde #, símbolo comodín, puede corresponder tanto a 0 como a 1
- Un esquema es una cadena con símbolos fijos y variables que representa a cadenas formadas por símbolos fijos
- El esquema H=01#1# representa las cadenas:

$$\begin{array}{cc} 01010 & 01110 \\ 01111 & 01011 \end{array}$$

- El valor de adecuación de un esquema se calcula como la media de los valores de adecuación de los individuos a los que representa

$$f_H = \frac{f(01010) + f(01011) + f(01110) + f(01111)}{4}$$

Se verifican las siguientes propiedades:

1. Si un esquema contiene k símbolos de indiferencia entonces representa a 2^k cadenas binarias
2. Una cadena binaria de longitud λ encaja en 2^λ esquemas distintos
3. Considerando las cadenas binarias de longitud λ existen en total 3^λ posibles esquemas.
4. Una población de n cadenas binarias de longitud λ contiene entre 2^λ y $n \cdot 2^\lambda$ esquemas distintos
   - 2^λ si todas las cadenas son iguales
   - $n \cdot 2^\lambda$ da una cota superior al número de esquemas distintos en cadenas distintas

- Cada cadena evaluada aporta información parcial acerca del valor de adecuación de todos los esquemas que representan a dicha cadena
- En la reproducción proporcional al valor de adecuación, el número de individuos, m, representados por un determinado esquema, H, en la generación $t+1$ está relacionado con:

$$m(H, t+1) = m(H, t)\frac{f_H(t)}{\bar{f}(t)}$$

donde $f_H(t)$ es el fitness medio del esquema H y $\bar{f}(t)$ el fitness medio de la población

- El número de cadenas que son representadas por un esquema con valor de adecuación superior a la media tienen un crecimiento exponencial

Propiedades de los esquemas

- Algunos esquemas son más específicos que otros:
  - El esquema $011*1**$ está más definido que el esquema $0******$
- Ciertos esquemas abarcan más parte de la longitud total de la cadena que otros:
  - El esquema $1****1*$ abarca una porción de la cadena mayor que el esquema $1*1****$
- Para cuantificar estas ideas se definen dos propiedades de los esquemas:
  - El orden de un esquema
  - La longitud característica

Dado un esquema H se definen:

- Orden de un esquema, $o(H)$: Es el número de posiciones fijadas que contiene dicho esquema
  - Por ejemplo, el orden del esquema $o(011*1**) = 4$ y $o(0******) = 1$
- Longitud característica de un esquema, $\delta(H)$: Es la distancia entre las primera y última posiciones fijadas de la cadena
  - El esquema $H = 011*1**$ tiene una longitud longitud característica de: $\delta(H) = 4$
  - La última posición especificada es 5 y la primera es 1
  - El esquema $H = 0******$ tiene longitud de definición $\delta(H) = 0$

- Un esquema H representa a $2^{\lambda - o(H)}$ cadenas:
  - Cuanto mayor sea el orden del esquema a menos cadenas representará
  - El orden de un esquema da una medida de su especificidad
- La longitud característica da una medida de la compacidad de la información contenida en el esquema
- Un esquema con una sola posición especificada tiene una longitud característica de cero
- Ejemplo: El esquema $H = ****00*** $ tiene $O(H) = 2$ y $\delta(H) = 6 - 5 = 1$

Efecto de la reproducción

- Supóngase que en un instante dado de tiempo t hay m ejemplares de un esquema H: $m(H, t) = m$
- Durante la reproducción, una cadena se copia deacuerdo con su aptitud. Una cadena A_i es seleccionada con probabilidad:

$$p(A_i) = \frac{f_i}{\sum f_j}$$

- Si la población es de tamaño n. Mediante reemplazamientos se espera tener $m(H, t+1)$ representantes del esquema en la siguiente generación, donde:

$$m(H, t+1) = m(H, t) \cdot n \cdot \frac{f(H)}{\sum f_j}$$

- Un esquema particular crece como el porcentaje de la aptitud media del esquema respecto de la aptitud de la población

Efecto de la reproducción

- Si un esquema H permanece por encima de la media una cantidad $c \cdot \bar{f}$, el efecto de la reproducción es:

$$m(H, t+1) = m(H, t)\frac{\bar{f} + c \cdot \bar{f}}{\bar{f}} = (1+c)m(H, t)$$

  Comenzando en $t = 0$:

$$m(H, t) = m(H, 0)(1+c)^t$$

- La reproducción asigna un número exponencialmente creciente (decreciente) de ejemplares a los esquemas por encima (por debajo) de la media.

Efecto del cruce sobre los bloques constructivos

- Los operadores genéticos actúan sobre los esquemas:
  - El efecto del sobrecruzamiento es disminuir el incremento exponencial por una cantidad proporcional a p_c, y depende de la longitud de la cadena, λ, y de la longitud de definición del esquema, $\delta(H)$
  - La probabilidad de que una cadena sobreviva al cruce es:

$$p_c \frac{\delta(H)}{\lambda - 1}$$

  - La longitud de definición es la distancia entre el primer y último símbolo fijo (no #) de un esquema:

$$\begin{array}{ll} 01\#1\# & \delta(H) = 3 \\ \#\#1\#1010 & \delta(H) = 5 \end{array}$$

Ejemplo del efecto del cruce

- Sea el individuo: $I = 0111000$
- Está representado por los esquemas:

$$H_1 = *1****0,, \ \delta(H_1) = 5$$

$$H_2 = ***10**,, \ \delta(H_2) = 1$$

- Si se utiliza para cruce y se elige el locus 4:

I	=	011	1000
H_1	=	*1*	***0
H_2	=	***	10**

- El esquema H_2 seguirá representando al descendiente, pero no así el H_1
- La probabilidad de que H_1 y H_2 sean destruidos es respectivamente $5/6$ y $1/6$

Efecto de la mutación sobre los esquemas

- En el caso de la mutación:
  - Si la probabilidad de mutación es p_m, entonces la probabilidad de que un bit sobreviva es $1 - p_m$
  - La probabilidad de que una cadena sobreviva a las mutaciones es:

$$(1 - p_m)^{O(H)}$$

  - La probabilidad de sobrevivir a la mutación con $p_m << 1$, puede ser aproximado por $(1 - O(H) \cdot p_m)$
- El orden de un esquema, $O(H)$, se define como la longitud de la cadena, λ, menos el número de símbolos comodín

- Agrupando los efectos de los operadores genéticos se obtiene el Teorema del Esquema:

$$m(H, t+1) \leq m(H, t) \frac{f_H(t)}{\bar{f}(t)} \left[1 - p_c \frac{\delta(H)}{\lambda - 1} - O(H) \cdot p_m \right]$$

- El número de cadenas cortas, de bajo orden, crece exponencialmente en las siguientes generaciones si están por encima de la media

Teorema del esquema

- El resultado anterior recibe el nombre de Teorema del esquema o Teorema Fundamental de los algoritmos genéticos:

  La presencia de un esquema H en la población P en el instante t en un Algoritmo Genético evoluciona estadísticamente de modo exponencial según la ecuación anterior.

- Los esquemas de orden bajo adaptados por encima de la media reciben un número exponencialmente creciente de oportunidades en siguientes generaciones
- La presencia de un esquema en una población evoluciona estadísticamente en progresión geométrica, cuyo factor está determinado por el fitness del esquema, su longitud característica y su orden

Hipótesis de los Bloques Constructivos

- Los bloques constructivos son esquemas muy adaptados de baja longitud característica y bajo orden
- Los algoritmos genéticos buscan un rendimiento cercano al óptimo mediante la yuxtaposición de los bloques constructivos.
- Hipótesis de los Bloques Constructivos: Los Algoritmos Genéticos exploran el espacio de búsqueda a través de la yuxtaposición de esquemas aventajados, cortos y de bajo orden. Estos esquemas se denominan bloques constructivos

Paralelismo Implícito

- Una cadena de longitud L incluye 2^L esquemas
- Hay 3^L diferentes esquemas de longitud L
- Una población de tamaño N contiene entre 2^L y $min(N \cdot 2^L, 3^L)$ esquemas
- El AG opera implícitamente sobre un número de esquemas mucho mayor que el tamaño de la población:

$$\begin{array}{cc} L = 6,\ N = 20 & \{64, 729\} \\ L = 40,\ N = 100 & \{10^{12}, 10^{19}\} \end{array}$$

- La estrategia de búsqueda en AG's incrementa exponencialmente a los individuos con mejor "fitness", pero sin dejar de probar nuevas soluciones
- El crecimiento de individuos es exponencial en aquellos que se estiman como mejores frente a aquellos que se estiman peores

Explotación

- La estrategia de búsqueda en AG's incrementa exponencialmente a los individuos con mejor "fitness", pero sin dejar de probar nuevas soluciones
- El crecimiento de individuos es exponencial en aquellos que se estiman como mejores frente a aquellos que se estiman peores

Exploración vs. Explotación

- La adaptación es una tensión entre la exploración (búsqueda de nuevas y mejores soluciones) y la explotación (el uso y la propagación de esas soluciones)
- Se trata de encontrar el recorrido que optimiza la generación de nuevos caminos de exploración con la explotación de caminos que están siendo explorados
- A pesar de la ruptura de los esquemas largos de orden alto por los operadores de cruce y mutación, los algoritmos genéticos procesan inherentemente una gran cantidad de esquemas mientras procesan una cantidad relativamente pequeña de cadenas

Two Armed Bandit

- Problema:

  - Máquina tragaperras con dos brazos: A_1 y A_2
  - Cada máquina da premios según una variable aleatoria gaussiana definida por su media y su varianza, $G(\mu_1, \sigma_1)$ y $G(\mu_2, \sigma_2)$, independientes
  - Una de las dos máquinas da más premios que la otra ($\mu_i > \mu_j$), pero se desconoce cuál

Two Armed Bandit

- Jugador:
  - N monedas para jugar en la máquina
  - No conoce a priori los valores de las medias y varianzas de los pagos; sólamente puede estimarlos a partir de las jugadas que realiza
- El objetivo del jugador es maximizar las ganancias (minimizar las pérdidas) totales obtenido tras las N jugadas
- De las N jugadas n las hace en la máquina A_1 y $N - n$ en la máquina A_2

- Las pérdidas generadas serán la diferencia de premio entre lo que se ha jugado y lo que se habría obtenido si se hubiera jugado en la máquina que da más premio:
  - Si la Máquina A_1 es la que produce más ganancias ($\mu_1 > \mu_2$):

$$P = (N - n)(\mu_1 - \mu_2)$$

  - Si la Máquina A_2 es la que produce más ganancias ($\mu_2 > \mu_1$):

$$P = n(\mu_1 - \mu_2)$$

- Se desconocen los valores de μ_1 y μ_2, por lo que solo se puede estimar cuál es la máquina que da más premio
- Se llama $q_e(n)$ a la probabilidad de error, es decir la probabilidad de hacer n jugadas en la peor máquina
- La probabilidad de hacer n jugadas en la mejor máquina será $1 - q_e(n)$

- Según lo anterior, las pérdidas totales al hacer N jugadas, de las cuales n se realizan en la máquina A_1, pensando que es la óptima será:

$$P(N,n) = q_e(n)(N-n)(\mu_1 - \mu_2) + (1 - q_e(n))n(\mu_1 - \mu_2)$$

- Teniendo en cuenta el teorema central del límite, se puede aproximar $q_e(n)$ mediante la cola de una distribución normal:

$$q_e(n) \approx \frac{1}{\sqrt{2 \cdot \pi}} \frac{e^{-c^2/2}}{c}$$

donde:

$$c = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \sqrt{n}$$

- Si se hace $n = 0$, es decir, se apuesta siempre a la misma máquina, el riesgo es máximo. Se hace explotación de una de las máquinas:
  - Si la elegida es la mejor se obtendrán todo ganancias
  - Pero si la elegida es la peor se obtendrán todo pérdidas
- Si se hace $n = \frac{N}{2}$, se está en la postura más conservadora. Se hace únicamente exploración
- Aparentemente la primera parce la postura más razonable, pero si hay muchas máquinas (N bandit armed problem) entonces la probabilidad de elegir la mejor es muy baja
- Lo mejor es hacer una balance entre exploración y explotación

- Para hacer un balance entre exploración y explotación se sustituye la aproximación de $q_e(n)$ en la ecuación de las pérdidas:

$$P(N,n) = \frac{1}{\sqrt{2\cdot\pi}}\frac{e^{-c^2/2}}{c}(N-n)(\mu_1-\mu_2)+(1-\frac{1}{\sqrt{2\cdot\pi}}\frac{e^{-c^2/2}}{c})n(\mu_1-\mu_2)$$

donde:

$$c = \frac{\mu_1-\mu_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\sqrt{n}$$

- Para saber cuál sería la forma óptima de apostar, habría que obtener el máximo de la expresión anterior, o lo que es lo mismo hacer:

$$\frac{\partial P(N,n)}{\partial n}$$

- Si se hacen una serie de simplificaciones, como que las distribuciones son independientes, etc. el óptimo anterior se encuentra en:

$$n^* \approx b^2 ln\left(\frac{N^2}{8\pi b^4 lnN^2}\right)$$

y:

$$N - n^* \approx \sqrt{8\pi b^4 lnN^2} \cdot e^{n^*/2b^2}$$

donde:

$$b = \frac{\sigma_1}{\mu_1 - \mu_2}$$

- Los valores de μ_1 y μ_2 son desconocidos, pero se pueden estimar mediante las medias observadas hasta el momento de las ganancias de las dos máquinas A_1 y A_2

- Por lo tanto la estrategia óptima es apostar de forma exponencial en la mejor máquina observada hasta el momento, de forma proporcional a la diferencia entre las ganancias observadas hasta el momento en las dos máquinas
- Los algoritmos genéticos siguen esta misma idea de dar un número de intentos exponencialmente a los mejores individuos hasta la fecha
- Los mejores esquemas son explorados de forma exponencial frente a los peores, según el teorema del esquema

- Si se considera una posición arbitraria como fija, existen para ella dos posibles esquemas, p. ej. $H_1 = ****0****$ y $H_1 = ****1****$
- Según el teorema del esquema el AG decide implícitamente entre estos dos esquemas, a partir de su fitness (ganancias observadas)
- En este sentido, el AG está resolviendo un gran número de Two Armed Bandit Problems en paralelo

Los AG NO son optimizadores de funciones

- Los Algoritmos Genéticos NO son optimizadores de funciones, lo que hacen es minimizar (maximizar) las pérdidas (ganancias) durante toda la búsqueda
- Supóngase que se quiere obtener una vacuna para solucionar una terrible enfermedad. Durante el periodo de investigación se diseña una vacuna y se prueba con un número de individuos
- Los efectos de la vacuna puede ser beneficiosos, se curan algunos individuos, o perjudiciales, algunos individuos se mueren
- El problema NO es encontrar la mejor vacuna posible, sino encontrarla con el menor número de personas afectadas
- Los Algoritmos Genéticos hacen una búsqueda óptima, bajo estas circunstancias
- El fitness medio de todos los individuos probados a lo largo de todas las generaciones es óptimo