Giải thuật di truyền

1. Mở đầu

Giải thuật di truyền (GA) là giải thuật tìm kiếm dựa trên quá trình *chọn lọc tự nhiên*, *di truyền* và *tiến hoá*. Các nguyên lý cơ bản của giải thuật được tác giả Holland đề xuất lần đầu vào năm 1962. Nền tảng toán học của giải thuật GA được tác giả công bố trong cuốn sách "*Sự thích nghi trong các hệ thống tự nhiên và nhân tạo*" xuất bản năm 1975. Giải thuật GA được xem như một phương pháp tìm kiếm có bước chuyển ngẫu nhiên mang tính tổng quát để giải các bài toán tối ưu hoá.

2. Các khái niệm cơ bản của giải thuật di truyền

2.1. Giới thiệu chung

Giải thuật GA thuộc lớp các giải thuật tìm kiếm tiến hoá. Khác với phần lớn các giải thuật khác tìm kiếm theo điểm, giải thuật GA thực hiện tìm kiếm song song trên một tập được gọi là *quần thể* các lời giải có thể. Thông qua việc áp dụng các toán tử gien, giải thuật GA tráo đổi thông tin giữa các cực trị và do đó làm giảm thiểu khả năng kết thúc giải thuật tại một cực trị địa phương. Trong thực tế, giải thuật GA đã được áp dụng thành công trong nhiều lĩnh vực.

Giải thuật GA lần đầu được tác giả Holland giới thiệu vào năm 1962. Giải thuật GA mô phỏng quá trình *tồn tại* của các *cá thể* có *độ phù hợp* tốt nhất thông qua quá trình chọn lọc tự nhiên, sao cho khi giải thuật được thực thi, quần thể các lời giải *tiến hoá* tiến dần tới lời giải mong muốn. Giải thuật GA duy trì một quần thể các lời giải có thể của bài toán tối ưu hoá. Thông thường, các lời giải này được mã hoá dưới dạng một chuỗi các gien. Giá trị của các gien có trong chuỗi được lấy từ một *bảng các ký tự* được định nghĩa trước. Mỗi chuỗi gien được liên kết với một giá trị được gọi là *độ phù hợp*. Độ phù hợp được dùng trong quá trình *chọn lọc*. Cơ chế chọn lọc đảm bảo các cá thể có độ phù hợp tốt hơn có xác suất được lựa chọn cao hơn. Quá trình chọn lọc sao chép các bản sao của các cá thể có độ phù hợp tốt vào một quần thể tạm thời được gọi là *quần thể bố mẹ*. Các cá thể trong quần thể bố mẹ được ghép đôi một cách ngẫu nhiên và tiến hành *lai ghép* tạo ra các cá thể con. Sau khi tiến hành quá trình lai ghép, giải thuật GA mô phỏng một quá trình khác trong tự nhiên là quá trình *đột biến*, trong đó các gien của các cá thể con tự thay đổi giá trị với một xác suất nhỏ.

Tóm lại, có 6 khía cạnh cần được xem xét, trước khi áp dụng giải thuật GA để giải một bài toán, cụ thể là:

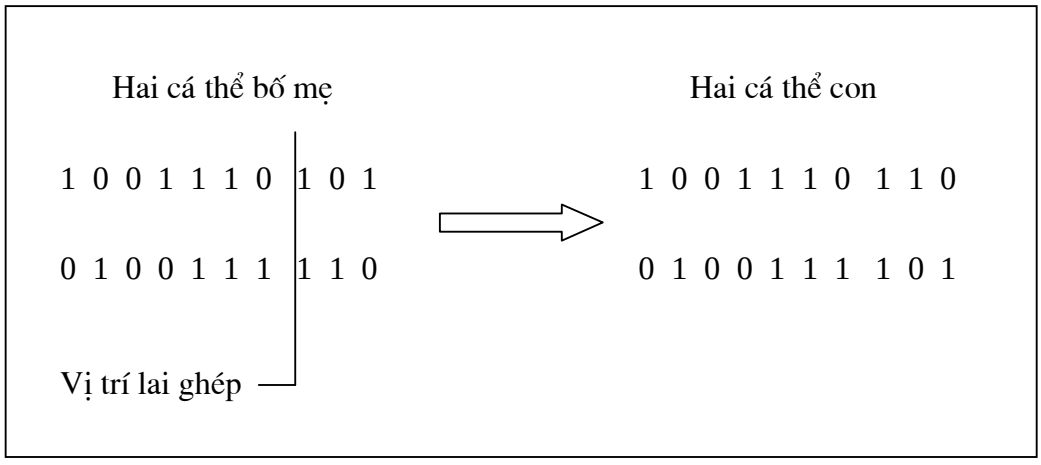
- Mã hoá lời giải thành cá thể dạng chuỗi.

- Hàm xác định giá trị độ phù hợp.
- Sơ đồ chọn lọc các cá thể bố mẹ.
- Toán tử lai ghép.
- Toán tử đột biến.
- *Chiến lược thay thế* hay còn gọi là *toán tử tái tạo*.

Có nhiều lựa chọn khác nhau cho từng vấn đề trên. Phần tiếp theo sẽ đưa ra cách lựa chọn theo J.H. Holland khi thiết kế phiên bản giải thuật GA đầu tiên. Giải thuật này được gọi là *giải thuật di truyền đơn giản* (SGA).

2.2. Giải thuật di truyền đơn giản

J. H. Holland sử dụng mã hoá nhị phân để biểu diễn các cá thể, lý do là phần lớn các bài toán tối ưu hoá đều có thể được mã hoá thành chuỗi nhị phân khá đơn giản [41]. Hàm *mục tiêu*, hàm cần tối ưu, được chọn làm cơ sở để tính độ phù hợp của từng chuỗi cá thể. Giá trị độ phù hợp của từng cá thể sau đó được dùng để tính toán xác suất chọn lọc. Sơ đồ chọn lọc trong giải thuật SGA là sơ đồ *chọn lọc tỷ lệ*. Trong sơ đồ chọn lọc này, cá thể có độ phù hợp f_i có xác suất chọn lựa $p_i = f_i / \sum_{j=1}^{N} f_j$, ở đây N là số cá thể có trong quần thể. Toán tử lai ghép trong giải thuật SGA là *toán tử lai ghép một điểm cắt*. Giả sử chuỗi cá thể có độ dài L (có L bít), toán tử lai ghép được tiến hành qua hai giai đoạn là:



Hình 1. ***Hình ảnh minh hoạ của toán tử lai ghép một điểm cắt.***

- Hai cá thể trong quần thể bố mẹ được chọn một cách ngẫu nhiên với phân bố xác suất đều.

- Sinh một số ngẫu nhiên j trong khoảng [1, L - 1] . Hai cá thể con được tạo ra bằng việc sao chép các ký tự từ 1 đến j và tráo đổi các ký tự từ j + 1 đến L. Quá trình này được minh hoạ như trong hình 1

Điều đáng lưu ý là giải thuật GA không yêu cầu toán tử lai ghép luôn xảy ra đối với hai cá thể bố mẹ được chọn. Sự lai ghép chỉ xảy ra khi số ngẫu nhiên tương ứng với cặp cá thể bố mẹ được sinh ra trong khoảng [0, 1) không lớn hơn một tham số p_c (gọi là *xác suất lai ghép*). Nếu số ngẫu nhiên này lớn hơn p_c, toán tử lai ghép không xảy ra. Khi đó hai cá thể con là bản sao trực tiếp của hai cá thể bố mẹ.

Tiếp theo, J. H. Holland xây dựng toán tử đột biến cho giải thuật SGA. Toán tử này được gọi là *toán tử đột biến chuẩn*. Toán tử đột biến duyệt từng gien của từng cá thể con được sinh ra sau khi tiến hành toán tử lai ghép và tiến hành biến đổi giá trị từ 0 sang 1 hoặc ngược lại với một xác suất p_m được gọi là *xác suất đột biến*. Cuối cùng là chiến lược thay thế hay còn gọi là toán tử tái tạo. Trong giải thuật SGA, quần thể con được sinh ra từ quần thể hiện tại thông qua 3 toán tử là chọn lọc, lai ghép và đột biến thay thế hoàn toàn quần thể hiện tại và trở thành quần thể hiện tại của thế hệ tiếp theo. Sơ đồ tổng thể của giải thuật SGA được thể hiện qua thủ tục GSA() dưới đây

Thủ tục SGA () /* *Bài toán tối ưu* */
{ k = 0;
// *Khởi động quần thể P_0 một cách ngẫu nhiên. Tính giá trị hàm mục tiêu*
// *cho từng cá thể.*
khởi_động (P_k);
tính_hàm_mục_tiêu (P_k);
// *Đặt lời giải của giải thuật bằng cá thể có giá trị hàm mục tiêu tốt nhất.*
X_{best} = tốt_nhất (P_k);
do { // Chuyển đổi giá trị hàm mục tiêu thành giá trị độ phù hợp và
// *tiến hành chọn lọc tạo ra quần thể bố mẹ P_{parent}*
P_{parent} = chọn_lọc (P_k);
// *Tiến hành lai ghép và đột biến tạo ra quần thể cá thể con P_{child}*
P_{child} = đột_biến (lai_ghép (P_{parent}));
// *Thay thế quần thể hiện tại bằng quần thể cá thể con*
k = k + 1;
P_k = P_{child};
tính_hàm_mục_tiêu (P_k);
// *Nếu giá trị hàm mục tiêu obj của cá thể tốt nhất X trong quần*

// *thể* P_k *lớn hơn giá trị hàm mục tiêu của* X_{best} *thì thay thế lời giải*

X = tốt_nhất (P_k);

if (obj (X) > obj (X_{best})) X_{best} = X;

} **while** (k < G); /* *Tiến hành G thế hệ* */

return (X_{best}); /* *Trả về lời giải của giải thuật GA*/

}

Giải thuật di truyền phụ thuộc vào bộ 4 (N, p_c, p_m, G), trong đó N - số cá thể trong quần thể; p_c - xác suất lai ghép; p_m - xác suất đột biến và G - số thế hệ cần tiến hoá, là các tham số điều khiển của giải thuật SGA. Cá thể có giá trị hàm mục tiêu tốt nhất của mọi thế hệ là lời giải cuối cùng của giải thuật SGA. Quần thể đầu tiên được khởi tạo một cách ngẫu nhiên.

3. Cơ chế thực hiện của giải thuật di truyền

Trong phần này chúng ta sẽ tìm hiểu về cơ chế thực hiện của giải thuật di truyền thông qua một bài toán tối ưu số. Không làm mất tính tổng quát, ta giả định bài toán tối ưu là bài toán tìm cực đại của hàm nhiều biến f. Bài toán tìm cực tiểu hàm g chính là bài toán tìm cực đại hàm $f = -g$, hơn nữa ta có thể giả định hàm mục tiêu f có giá trị dương trên miền xác định của nó, nếu không ta có thể cộng thêm một hằng số C dương

Cụ thể bài toán được đặt ra như sau: Tìm cực đại một hàm k biến $f(x_1,.., x_k)$: $\boldsymbol{R^k \rightarrow R}$. Giả sử thêm là mỗi biến x_i có thể nhận giá trị trong miền $D_i = [a_i,b_i] \subseteq \boldsymbol{R}$ và $f(x_1,.., x_k) > 0$ với mọi $x_i \in D_i$. Ta muốn tối ưu hàm f với độ chính xác cho trước: giả sử cần n số lẻ đối với giá trị của các biến

Để đạt được độ chính xác như vậy mỗi miền D_i cần được phân cắt thành $(b_i - a_i) \times 10^n$ miền con bằng nhau, gọi m là số nguyên nhỏ nhất sao cho

$$(b_i - a_i) \times 10^n \leq 2^{m_i} - 1$$

Như vậy mỗi biến x_i được biểu diễn bằng một chuỗi nhị phân có chiều dài m_i. Biểu diễn như trên rõ ràng thoả mãn điều kiện về độ chính xác theo yêu cầu. Công thức sau tính giá trị thập phân của mỗi chuỗi nhị phân biểu diễn biến x_i

$$x_i = a_i + decimal(string_2)\frac{b_i - a_i}{2^{m_i} - 1}$$

trong đó $decimal(string_2)$ cho biết giá trị thập phân của chuỗi nhị phân đó

Bây giờ, mỗi nhiễm sắc thể (là một lời giải) được biểu diễn bằng một chuỗi nhị phân có chiều dài $m = \sum_{i=1}^{k} m_i$, m_1 bit đầu tiên biểu diễn giá trị trong khoảng $[a_1,b_1]$, m_2 bit kế tiếp biểu diễn giá trị trong khoảng $[a_2,b_2]$, …

Để khởi tạo quần thể, chỉ cần đơn giản tạo *pop _size* nhiễm sắc thể ngẫu nhiên theo từng bit

Phần còn lại của giải thuật di truyền rất đơn giản, trong mỗi thế hệ, ta lượng giá từng nhiễm sắc thể (tính giá trị hàm *f* trên các chuỗi biến nhị phân đã được giải mã), chọn quần thể mới thoả mãn phân bố xác suất dựa trên độ thích nghi và thực hiện các phép đột biến và lai để tạo ra các cá thể thế hệ mới. Sau một số thế hệ, khi không còn cải thiện thêm được gì nữa, nhiễm sắc thể tốt nhất sẽ được xem như lời giải của bài toán tối ưu (thường là toàn cục). Thông thường ta cho dừng giải thuật sau một số bước lặp cố định tuỳ ý tuỳ thuộc vào điều kiện tốc độ và tài nguyên máy tính

Đối với tiến trình chọn lọc (chọn quần thể mới thoả phân bố xác suất dựa trên các độ thích nghi), ta dùng bánh xe quay Rulet với các rãnh được định kích thước theo độ thích nghi. Ta xây dựng bánh xe Rulet như sau (giả định rằng các độ thích nghi đều dương)

+ Tính độ thích nghi $eval(v_i)$ của mỗi nhiễm sắc thể v_i ($i = 1,\ldots, pop\_size$)

+ Tìm tổng giá trị thích nghi toàn quần thể: $F = \sum_{i=1}^{pop-size} eval(v_i)$

+ Tính xác suất chọn p_i cho mỗi nhiễm sắc thể v_i, ($i = 1,\ldots, pop\_size$): $p_i = eval(v_i) / F$

+ Tính vị trí xác suất q_i của mỗi nhiễm sắc thể v_i, ($i = 1,\ldots, pop\_size$): $q_i = \sum_{j=1}^{i} p_i$

Tiến trình chọn lọc thực hiện bằng cách quan bánh xe Rulet *pop_size* lần, mỗi lần chọn một nhiễm sắc thể từ quần thể hiện hành vào quần thể mới theo cách sau:

+ Phát sinh ngẫu nhiên một số *r* trong khoảng [0..1]

+ Nếu $r < q_1$ thì chọn nhiễm sắc thể đầu tiên v_1, ngược lại thì chọn nhiễm sắc thể thứ *i*, v_i ($2 \le i \le pop\_size$) sao cho $q_{i-1} < r < q_i$

Hiển nhiên có thể có một số nhiễm sắc thể được chọn nhiều lần, điều này là phù hợp vì các nhiếm sắc thể tốt nhất cần có nhiều bản sao hơn, các nhiễm sắc thể trung bình không thay đổi, các nhiễm sắc thể kém nhất thì chết đi

Bây giờ ta có thể áp dụng phép toán di truyền: kết hợp và lại vào các cá thể trong quần thể mới vừa được chọn từ quần thể cũ như trên. Một trong nhữn tham số của giải thuật là xác suất lai p_c. Xác suất này cho ta số nhiếm sắc thể $pop\_size \times p_c$ mong đợi, các nhiễm sắc thể này được dùng trong tác vụ lai tạo. Ta tiến hành theo cách sau đây:

Đối với mỗi nhiễm sắc thể trong quần thể mới:

+ Phát sinh ngẫu nhiên một số r trong khoảng [0,1]

+ Nếu $r < p_c$, hãy chọn nhiễm sắc thể đó để lai tạo

Bây giờ ta ghép đôi các nhiễm sắc thể đã được chọn một cách ngẫu nhiên: đối với mỗi cặp nhiễm sắc thể được ghép đôi, ta phát sinh ngẫu nhiên một số nguyên *pos* trong khoảng [1, m-1] (m là tổng chiều dài - số bit - của một nhiễm sắc thể). Số *pos* cho biết vị trí của điểm lai. Hai nhiễm sắc thể:

$(b_1 b_2 \ldots b_{pos} b_{pos+1} \ldots b_m)$ và $(c_1 c_2 \ldots c_{pos} c_{pos+1} \ldots c_m)$

được thay bằng một cặp con của chúng:

$(b_1 b_2 \ldots b_{pos} c_{pos+1} \ldots c_m)$ và $(c_1 c_2 \ldots c_{pos} b_{pos+1} \ldots b_m)$

Phép toán kế tiếp là phép đột biến, được thực hiện trên cơ sở từng bit. Một tham số khác của giải thuật là xác suất đột biến p_m, cho ta số bit đột biến $p_m \times m \times pop\_size$ mong đợi. Mỗi bit (trong tất cả các nhiễm sắc thể trong quần thể) có cơ hội bị đột biến như nhau, nghĩa là đổi từ 0 thành 1 hoặc ngược lại. Vì thế ta tiến hành theo cách sau đây:

Đối với mỗi nhiễm sắc thể trong quần thể hiện hành (nghĩa là sau khi lai) và đỗi với mỗi bit trong nhiễm sắc thể:

+ Phát sinh ngẫu nhiên một số r trong khoảng [0,1]

+ Nếu $r<p_m$ hãy đột biến bit đó

Sau quá trình chọn lọc, lai và đột biến, quần thể mới đến lượt lượng giá kế tiếp của nó. Lượng giá này được dùng để xây dựng phân bố xác suất (cho tiến trình chọn lựa kế tiếp), nghĩa là để xây dựng lại bánh xe Rulet với các rãnh được định kích thước theo các giá trị thích nghi hiện hành. Phần còn lại của tiến hoá chỉ là lặp lại chu trình của những bước trên

Toàn bộ tiến trình sẽ được minh hoạ trong một thí dụ cực đại hoá hàm [3]:

$$f(x_1, x_2) = 21.5 + x_1 \sin(4\pi x_1) + x_2 \sin(20\pi x_2)$$

Giả sử kích thước quần thể *pop_size* = 20, các xác suất di truyền tương ứng là $P_c = 0.25$ và $P_m = 0.01$

Giả sử cần tính chính xác đến 4 số lẻ đối với mỗi biến. Miền của biến x_1 có chiều dài 15.1, điều kiện chính xác đòi hỏi đoạn [-3.0, 12.1] cần được chia thành các khoảng có kích thước bằng nhau, ít nhất là 15.1×10000 khoảng, điều này cần 18 bit làm phần đầu tiên của nhiễm sắc thể: $2^{17} \le 151000 \le 2^{18}$

Miền của biến x_2 có chiều dài là 1.7, điều kiện chính xác đòi hỏi đoạn [4.1,5.8] cần được chia thành các khoảng có kích thước bằng nhau là 1.7×10000 khoảng, điều này nghĩa là cần 15 bit làm thành phần cuối của nhiễm sắc thể: $2^{14} \le 17000 \le 2^{15}$

Chiều dài toàn bộ nhiễm sắc thể (vectơ lời giải) là m =18+15 = 33

Để cực đại hoá hàm f bằng giải thuật di truyền ta tạo ra một quần thể có *pop_size* = 20 nhiễm sắc thể. Cả 33 bit trong tất cả các nhiễm sắc thể đều được khởi tạo ngẫu nhiên

Giả sử sau tiến trình khởi tạo ta có quần thể sau đây:

v_1 = (100110100000001111111010011011111)

v_2 = (111000100100110111001010100011010)

v_3 = (000010000011001000001010111011101)

v_4 = (100011000101101001111000001110010)

v_5 = (000111011001010011010111111000101)

v_6 = (000101000010010101001010111111011)

v_7 = (001000100000110101111011011111011)

v_8 = (100001100001110100010110101100111)

v_9 = (011000000101100010110000001111100)

v_{10} = (000001111000110000011010000111011)

v_{11} = (011001111110110101100001101111000)

v_{12} = (110100010111101101000101010000000)

v_{13} = (111011111010001000110000001000110)

v_{14} = (010010011000001010100111100101001)

v_{15} = (111011101101110000100011111011110)

v_{16} = (11001111000001111111000011010010 11)

v_{17} = (01101011111100111101000110111110 1)

v_{18} = (01110100000000111010011111010110 1)

v_{19} = (00010101001111111111000011000110 0)

v_{20} = (10111001011001111001100010111111 0)

Trong giai đoạn lượng giá ta giải mã từng nhiễm sắc thể và tính giá trị hàm thích nghi từ các giá trị (x_1, x_2) mới giải mã, ta có

$eval(v_1) = f(6.084492,5.652242) = 26.019600$

$eval(v_2) = f(10.348434,4.380264) = 7.580015$

$eval(v_3) = f(-2.516603,4.390381) = 19.626329$

$eval(v_4) = f(5.278638,5.593460) = 17.406725$

$eval(v_5) = f(-1.255173,4.734458) = 25.341160$

$eval(v_6) = f(-1.811725,4.391937) = 18.100417$

$eval(v_7) = f(-0.991471,5.680258) = 16.020812$

$eval(v_8) = f(4.910618,4.703018) = 17.959701$

$eval(v_9) = f(0.795406,5.381472) = 16.127799$

$eval(v_{10}) = f(-2.554851,4.793707) = 21.278435$

$eval(v_{11}) = f(3.130078,4.996097) = 23.410669$

$eval(v_{12}) = f(9.356179,4.239457) = 15.011619$

$eval(v_{13}) = f(11.134646,5.378671) = 27.316702$

$eval(v_{14}) = f(1.335944,5.151378) = 19.876294$

$eval(v_{15}) = f(11.089025,5.054515) = 30.060205$

$eval(v_{16}) = f(9.211598,4.993762) = 23.967227$

$eval(v_{17}) = f(3.367514,4.571343) = 13.696165$

$eval(v_{18}) = f(3.843020,5.158226) = 15.414128$

$eval(v_{19}) = f(-1.746635,5.395584) = 20.095903$

$eval(v_{20}) = f(7.935998,4.757338) = 13.666916$

Rõ ràng nhiễm sắc thể v_{15} mạnh nhất và nhiễm sắc thể v_2 yếu nhất

Bây giờ ta xây dựng bánh xe Rulet cho tiến trình chọn lọc. Tổng độ thích nghi của quần thể là:

$$F = \sum_{i=1}^{20} eval(v_i) = 387.776822$$

Xác suất chọn lọc p_i của mỗi nhiễm sắc thể v_i (i = 1,…,20) là:

$p_1 = eval(v_1)/F = 0.067099$ $p_2 = eval(v_2)/F = 0.019547$

$p_3 = eval(v_3)/F = 0.050355$ $p_4 = eval(v_4)/F = 0.044889$

$p_5 = eval(v_5)/F = 0.065350$ $p_6 = eval(v_6)/F = 0.046677$

$p_7 = eval(v_7)/F = 0.041315$ $p_8 = eval(v_8)/F = 0.046315$

$p_9 = eval(v_9)/F = 0.041590$ $p_{10} = eval(v_{10})/F = 0.054873$

$p_{11} = eval(v_{11})/F = 0.060372$ $p_{12} = eval(v_{12})/F = 0.038712$

$p_{13} = eval(v_{13})/F = 0.070444$ $p_{14} = eval(v_{14})/F = 0.051257$

$p_{15} = eval(v_{15})/F = 0.077519$ $p_{16} = eval(v_{16})/F = 0.061549$

$p_{17} = eval(v_{17})/F = 0.035320$ $p_{18} = eval(v_{18})/F = 0.039750$

$p_{19} = eval(v_{19})/F = 0.051823$ $p_{20} = eval(v_{20})/F = 0.035244$

Các vị trí xác suất q_i của mỗi nhiễm sắc thể v_i (i = 1,.., 20) là:

$q_1 = 0.067099$ $q_2 = 0.086647$ $q_3 = 0.137001$

$q_4 = 0.181890$ $q_5 = 0.247240$ $q_6 = 0.293917$

$q_7 = 0.335232$ $q_8 = 0.381546$ $q_9 = 0.423137$

$q_{10} = 0.478009$ $q_{11} = 0.538381$ $q_{12} = 0.577093$

$q_{13} = 0.647537$ $q_{14} = 0.698794$ $q_{15} = 0.776314$

$q_{16} = 0.837863$ $q_{17} = 0.873182$ $q_{18} = 0.812932$

$q_{19} = 0.964756$ $q_{20} = 1.000000$

Bây giờ ta quanh bánh xe Rulet 20 lần, mỗi lần chọn một nhiễm sắc thể cho quần thể mới. Giả sử thứ tự (ngẫu nhiên) của 20 số trong khoảng [0,1] được phát sinh là:

0.513870	0.175741	0.308652	0.534534	0.947628
0.171736	0.702231	0.226431	0.494773	0.424720
0.703899	0.389647	0.277226	0.368071	0.983437
0.005398	0.765682	0.646473	0.767139	0.780237

Số đầu tiên $r = 0.513870$ lớn hơn q_{10} và nhỏ hơn q_{11}, nghĩa là nhiễm sắc thể v_{11} được chọn vào quần thể mới, số thứ 2 $r = 0.175741$ lớn hơn q_3 nhỏ hơn q_4, nghĩa là v_4 được chọn cho quần thể mới,….

Như vậy quần thể mới gồm các nhiễm sắc thể sau:

$v'_1 = v_{11} = (01100111111011010110000110111 1000)$

$v'_2 = v_4 = (10001100010110100111100000011100 10)$

$v'_3 = v_7 = (00100010000011010111101101111 1011)$

$v'_4 = v_{11} = (01100111111011010110000110111 1000)$

$v'_5 = v_{19} = (00010101001111111111000011000 1100)$

$v'_6 = v_4 = (10001100010110100111100000011100 10)$

$v'_7 = v_{15} = (11101110110111000010001111101 1110)$

$v'_8 = v_5 = (00011101100101001101011111100 0101)$

$v'_9 = v_{11} = (01100111111011010110000110111 1000)$

$v'_{10} = v_3 = (00001000001100100000101011101 1101)$

$v'_{11} = v_{15} = (11101110110111000010001111101 1110)$

$v'_{12} = v_9 = (01100000010110001011000000111 1100)$

$v'_{13} = v_6 = (00010100001001010100101011111 1011)$

$v'_{14} = v_8 = (10000110000111010001011010110 0111)$

$v'_{15} = v_{20} = (10111001011001111001100010111 1110)$

$v'_{16} = v_1 = (10011010000000111111101001101 1111)$

$v'_{17} = v_{10} = (00000111100011000001101000011 1011)$

$v'_{18} = v_{13} = (11101111101000100011000000100 0110)$

$v'_{19} = v_{15} = (11101110110111000010001111101 1110)$

$v'_{20} = v_{16} = (11001111000001111110000110100 1011)$

Bây giờ ta sẽ áp dụng phép toán kết hợp, lai cho những cá thể trong quần thể mới (các véc tơ v'_i). Xác suất lai Pc = 0.25 vì thế ta hy vọng 25% nhiễm sắc thể sẽ tham gia lai tạo. Ta tiến hành theo cách sau: đối với mỗi nhiễm sắc thể trong quần thể mới ta phát sinh ngẫu nhiên một số r trong khoảng [0,1], nếu $r < 0.25$, ta chọn một nhiễm sắc thể cho trước để lai tạo

Giả sử thứ tự các số ngẫu nhiên là

0.822951	0.151932	0.625477	0.314685	0.346901
0.917204	0.519760	0.401154	0.606758	0.785402
0.031523	0.869921	0.166525	0.574520	0.758400
0.581893	0.389248	0.200232	0.355635	0.826927

Điều này có nghĩa là các nhiễm sắc thể v'_2, v'_{11}, v'_{13} và v'_{18} đã được chọn để lai tạo. Bây giờ ta cho lai tạo một cách ngẫu nhiên, ví dụ (v'_2, v'_{11}) và (v'_{13}, v'_{18}) được kết cặp. Đối với mỗi cặp trong 2 cặp này, ta phát sinh một số nguyên ngẫu nhiên *pos* thuộc khoảng {1,..,32}. Số *pos* cho biết vị trí của điểm lai tạo.

Cặp nhiễm sắc thể đầu tiên là:

v'_2 = (10001100|10110100111100000011100 10)

v'_{11} = (11101110|10111000010000111111011110)

và giả sử số phát sinh là *pos* = 9, kết quả lai tạo là:

v''_2 = (10001100|10111000010000111111011110)

v''_{11} = (11101110|10110100111100000011100 10)

Cặp nhiễm sắc thể thứ hai là:

v'_{13} = (00010100001001010100|10101111111011)

v'_{18} = (11101111101000100011|00000010001 10)

và giả sử số phát sinh là *pos* = 20, kết quả lai tạo là:

v''_{13} = (00010100001001010100|00000010001 10)

v''_{18} = (11101111101000100011|10101111111011)

Cuối cùng quần thể hiện hành là

v'_1 = (01100111111011010110000110111 1000)

v'_2 = (10001100|10111000010000111111011110)

v'_3 = (00100010000011010111101101111 1011)

v'_4 = (01100111111011010110000110111 1000)

v'_5 = (00010101001111111111000011000 1100)

v'_6 = (10001100010110100111100000011 10010)

v'_7 = (11101110110111000010000111111011110)

v'_8 = (00011101100101001101011111000101)

v'_9 = (01100111111011010110000110111000)

v'_{10} = (00001000001100100000010101110111011101)

v'_{11} = (11101110|1011010011110000011100 10)

v'_{12} = (01100000010110001011000000111110 0)

v'_{13} = (00010100001001010100|0000001000110)

v'_{14} = (10000110000111010001011010110011 1)

v'_{15} = (10111001011001111001100010111111 0)

v'_{16} = (10011010000000111111101001101111 1)

v'_{17} = (00000111100011000001101000011101 1)

v'_{18} = (11101111101000100011|1010111111011)

v'_{19} = (11101110110111000010001111101111 0)

v'_{20} = (11001111000001111110000110100101 1)

Phép toán kế tiếp, đột biến thực hiện trên cơ sở từng bit một. Xác suất đột biến p_m = 0.01, vì thế ta hy vọng 1/100 số bit sẽ qua đột biến. Có 660 bit ($m \times pop\_size$ = 33×20) trong toàn quần thể, ta hy vọng có 6.6 đột biến ở mỗi thế hệ. Mỗi bit có cơ hội đột biến ngang nhau, nên với mỗi bit trong quần thể ta phát sinh một số ngẫu nhiên r trong khoảng [0,1], nếu $r < 0.01$ thì ta đột biến bit này

Điều này có nghĩa ta phát sinh 660 số ngẫu nhiên. Giả sử có 5 trong 660 số này nhỏ hơn 0.01 (bảng dưới)

Vị trí bit	**Số ngẫu nhiên**
112	0.000213
349	0.009945
418	0.009909
429	0.005425
602	0.002835

Bảng sau cho biết nhiễm sắc thể, vị trí của bit bị đột biến tương ứng với 5 vị trí bit trên

Vị trí bit	**Số nhiễm sắc thể**	**Sô bit trong nhiễm sắc thể**

112	4	13
349	11	19
418	13	22
429	13	33
602	19	8

Điều này có nghĩa là 4 nhiễm sắc thể chịu ảnh hưởng của phép đột biến, một trong số này là nhiễm sắc thể số 13 có hai bit bị thay đổi

Quần thể sau phép đột biến được liệt kê dưới đây

v'_1 = (01100111111011010110000110111100 0)

v'_2 = (10001100010111000010000111110111 10)

v'_3 = (00100010000011010111101101111101 1)

v'_4 = (01100111111 **0** 10101100001101111000)

v'_5 = (00010101001111111111000001100011 00)

v'_6 = (10001100010110100111100000111001 0)

v'_7 = (11101110110111000010000111110111 10)

v'_8 = (00011101100101001101011111100010 1)

v'_9 = (01100111111011010110000110111100 0)

v'_{10} = (00001000001100100000101011101110 1)

v'_{11} = (11101110110110100 **0** 11000001110010)

v'_{12} = (01100000010110001011000000111110 0)

v'_{13} = (00010100001001010100 **1** 00001000111 **1**)

v'_{14} = (10000110000111010001011010110011 1)

v'_{15} = (10111001011001111001100010111111 0)

v'_{16} = (10011010000000111111101001101111 1)

v'_{17} = (00000111100011000001101000011101 1)

v'_{18} = (11101111101000100011101011111101 1)

v'_{19} = (11101110 **0** 10111000010001111101111 0)

v'_{20} = (11001111000001111111000011010010 11)

Ta vừa hoàn thành một bước lặp (nghĩa là một thế hệ) của giải thuật di truyền, tiếp theo ta giải mã và tính giá trị của hàm thích nghi:

$eval(v_1) = f(3.130078,4.996097) = 23.410669$

$eval(v_2) = f(5.279082,5.054515) = 18.201083$

$eval(v_3) = f(-2.516603,4.390381) = 19.626329$

$eval(v_4) = f(5.278638,5.593460) = 17.406725$

$eval(v_5) = f(-1.255173,4.734458) = 25.341160$

$eval(v_6) = f(-1.811725,4.391937) = 18.100417$

$eval(v_7) = f(-0.991471,5.680258) = 16.020812$

$eval(v_8) = f(4.910618,4.703018) = 17.959701$

$eval(v_9) = f(0.795406,5.381472) = 16.127799$

$eval(v_{10}) = f(-2.554851,4.793707) = 21.278435$

$eval(v_{11}) = f(3.130078,4.996097) = 23.410669$

$eval(v_{12}) = f(9.356179,4.239457) = 15.011619$

$eval(v_{13}) = f(11.134646,5.378671) = 27.316702$

$eval(v_{14}) = f(1.335944,5.151378) = 19.876294$

$eval(v_{15}) = f(11.089025,5.054515) = 30.060205$

$eval(v_{16}) = f(9.211598,4.993762) = 23.967227$

$eval(v_{17}) = f(3.367514,4.571343) = 13.696165$

$eval(v_{18}) = f(3.843020,5.158226) = 15.414128$

$eval(v_{19}) = f(-1.746635,5.395584) = 20.095903$

$eval(v_{20}) = f(7.935998,4.757338) = 13.666916$

4. Nguyên lý hoạt động của giải thuật di truyền

Nền tảng lý thuyết của giải thuật di truyền dựa trên biểu diễn chuỗi nhị phân và lý thuyết sơ đồ. Một sơ đồ là một chuỗi, dài bằng chuỗi nhiễm sắc thể, các thành phần của nó có thể nhận một trong các giá trị trong tập ký tự biểu diễn gien hoặc một ký tự đại diện '*'. Sơ đồ biểu diễn một không gian con của không gian tìm kiếm. Không gian con này là tập tất cả các chuỗi trong không gian lời giải mà với mọi vị trí

trong chuỗi, giá trị của gien trùng với giá trị của sơ đồ, ký tự đại diện '*' có thể trùng khớp với bất kỳ ký tự biểu diễn gien nào

Thí dụ, các chuỗi và sơ đồ có chiều dài 10.

Sơ đồ (*111100100) sẽ khớp với hai chuỗi:

{(0111100100), (1111100100)}

và sơ đồ (*1*1100100) sẽ khớp với bốn chuỗi:

{(0101100100), (0111100100),(1101100100), (1111100100)}

Đương nhiên sơ đồ (1001110001) chỉ khớp với chính nó và sơ đồ (**********) khớp với tất cả các chuỗi có chiều dài 10. Rõ ràng là mỗi sơ đồ cụ thể có tương ứng 2^r chuỗi với r là số ký tự đại diện '*' có trong sơ đồ. Mặt khác, mỗi chuỗi có chiều dài m sẽ khớp với 2^m sơ đồ.

Thí dụ, xét chuỗi (1001110001), chuỗi này phù hợp với 2^{10} sơ đồ sau:

(1001110001)

(*001110001)

(1*01110001)

:

:

(100111000*)

(**01110001)

(*0*1110001)

:

:

(10011100**)

(***1110001)

:

:

(**********)

Một chuỗi chiều dài m, sẽ có tối đa 3^m sơ đồ. Trong một quần thể kích thước n, có thể có tương ứng từ 2^m đến $n \times 2^m$ sơ đồ khác nhau

Các sơ đồ khác nhau có những đặc trưng khác nhau. Các đặc trưng này thể hiện qua hai thuộc tính quan trọng: bậc và chiều dài xác định

+ Bậc của sơ đồ S (ký hiệu là $o(S)$) là số các vị trí 0 và 1 có trong sơ đồ, đây chính là các vị trí cố định (không phải là những vị trí của ký tự đại diện) trong sơ đồ. Nói cách khác, bậc là chiều dài của chuỗi trừ đi số ký tự đại diện. Bậc xác định đặc trưng của sơ đồ

Thí dụ, ba sơ đồ chiều dài 10

S_1 = (***001*110)

S_2 = (****00**0*)

S_3 = (11101**001)

có bậc tương ứng:

$o(S_1) = 6$, $o(S_2) = 3$, $o(S_3) = 8$

và S_3 là sơ đồ 'đặc hiệu' nhất

Khái niệm bậc của sơ đồ giúp cho việc tính xác suất sống còn của sơ đồ do ảnh hưởng của đột biến

+ Chiều dài xác định của sơ đồ S (ký hiệu $\delta(S)$) là khoảng cách giữa hai vị trí cố định ở đầu và cuối. Nó định nghĩa *"độ nén"* của thông tin trong một sơ đồ

Thí dụ: $\delta(S_1) = 10 - 4 = 6$, $\delta(S_2) = 9 - 5 = 4$, $\delta(S_3) = 10 - 1 = 9$

Như vậy một sơ đồ chỉ có một vị trí cố định duy nhất thì sẽ có chiều dài xác định là 0

Khái niệm chiều dài xác định của sơ đồ giúp tính xác suất sống còn của sơ đồ do ảnh hưởng của phép lai

Như đã thảo luận ở các phần trước, tiến trình mô phỏng tiến hoá của giải thuật di truyền là quá trình lặp gồm có 4 bước

$t \leftarrow t+1$

chọn $P(t)$ *từ* $P(t-1)$

tái kết hợp $P(t)$

lượng giá $P(t)$

Bước 1, $t \leftarrow t+1$ chỉ đơn giản là đếm số thế hệ tiến hoá, bước cuối *lượng giá* $P(t)$ là lượng giá để tính độ thích nghi của các cá thể trong quần thể hiện hành. Hiện

tượng chủ yếu của chu trình tiến hoá xảy ra trong hai bước còn lại: *chọn lọc* và *tái kết hợp*. Ta sẽ bàn về hiệu quả của hai bước này trên một số sơ đồ cần thiết biểu diễn trong một quần thể

4.1. Chọn lọc

Xét bước *chọn lọc*: Giả sử quần thể có kích thước *pop_size* = 20, chiều dài của chuỗi và cũng là chiều dài của các sơ đồ là m = 33 (như thí dụ ở phần trước). Giả sử thêm rằng ở thế hệ thứ t quần thể gồm các chuỗi sau đây:

v_1 = (10011010000000111111110100110111111)

v_2 = (11100010010011011100101010011010)

v_3 = (00001000001100100000010101110111101)

v_4 = (10001100010110100111100000011100010)

v_5 = (00011101100101001101011111000101)

v_6 = (00010100001001010100101011111011)

v_7 = (00100010000011010111101101111011)

v_8 = (10000110000111010001011010110011)

v_9 = (01100000010110001011000000111100)

v_{10} = (00000111100011000001101000011101 1)

v_{11} = (01100111111011010110000110111000 0)

v_{12} = (11010001011110110100010101000000 0)

v_{13} = (11101111101000100011000000100011 0)

v_{14} = (01001001100000101010011110010100 1)

v_{15} = (11101110110111000010001111101111 0)

v_{16} = (11001111000001111110000110100101 1)

v_{17} = (01101011111100111101000110111110 1)

v_{18} = (01110100000000111010011111010110 1)

v_{19} = (00010101001111111111000011000110 0)

v_{20} = (10111001011001111001100010111111 0)

1. Đặt $\xi(S, t)$ là số chuỗi trong quần thể ở thế hệ thứ t phù hợp với sơ đồ S. Thí dụ đối với sơ đồ S_0 = (****111**************************) thì $\xi(S_0, t) = 3$ vì có 3 chuỗi v_{13}, v_{15} và v_{16} phù hợp với sơ đồ S_0. Chú ý rằng bậc của sơ đồ S_0 là $o(S_0) = 3$ và chiều dài của nó $\delta(S_0) = 7\text{-}5 = 2$

2. Gọi $eval(S, t)$ là độ thích nghi của sơ đồ S ở thế hệ t. Giả sử có p chuỗi $\{v_{i1},\ldots,v_{ip}\}$ trong quần thể phù hợp với sơ đồ S vào thời điểm t thì:

$$eval(S,t) = \frac{\sum_{j=1}^{p} eval(v_{ij})}{p}$$

Trong bước *chọn lọc*, một quần thể trung gian được tạo ra gồm $pop\_size = 20$ các chuỗi được chọn ra từ quần thể hiện hành. Các chuỗi được chọn dựa vào độ thích nghi của nó và được chép vào quần thể thế hệ mới. Như ta đã biết trong chương trước, chuỗi v_i có xác suất được chọn là $p_i = eval(v_i)/F(t)$ ($F(t)$ là tổng thích nghi của toàn quần thể vào thời điểm t, $F(t) = \sum_{i=1}^{20} eval(v_i)$)

Sau bước chọn lọc, ta có $\xi(S, t+1)$ chuỗi phù hợp với sơ đồ S do:

1. Với một chuỗi phù hợp với sơ đồ S, trung bình xác suất được chọn của nó là $eval(S, t)/F(t)$ và,

2. ở thế hệ t, số chuỗi phù hợp với sơ đồ S là $\xi(S, t)$ và,

3. Chọn trong pop_size chuỗi

Vậy $\xi(S, t+1) = \xi(S, t) \times pop\_size \times eval(S, t)/F(t)$

Với $\overline{F(t)} = F(t) / pop\_size$ là độ thích nghi trung bình của quần thể, ta có thể viết lại công thức trên thành

$$\xi(S,t+1) = \xi(S,t) \times eval(S,t) / \overline{F(t)} \qquad (4.1)$$

Nói cách khác, số chuỗi trong quần thể tăng bằng với tỷ lệ độ thích nghi của sơ đồ với độ thích nghi trung bình của quần thể. Điều này có nghĩa là sơ đồ "trên trung bình" sẽ nhận được thêm số chuỗi trong quần thể thế hệ kế tiếp, sơ đồ "dưới trung bình" nhận được số chuỗi giảm đi, còn sơ đồ trung bình vẫn giữ nguyên mức

Hệ quả lâu dài của luật trên cũng rõ ràng. Nếu ta cho rằng sơ đồ S vẫn giữ trên trung bình ε% nghĩa là $eval(S,t) = \overline{F(t)} + \varepsilon \times \overline{F(t)}$, thì:

$\xi(S,t) = \xi(S,0) \times (1+\varepsilon)^t$ và $\varepsilon = (eval(S,t) - \overline{F(t)}) / \overline{F(t)}$ ($\varepsilon > 0$ đối với các sơ đồ trên trung bình và $\varepsilon < 0$ đối với các sơ đồ dưới trung bình)

Như vậy ta có thể nói rằng, chẳng những một sơ đồ "trên trung bình" nhận số chuỗi tăng lên trong thế hệ kế tiếp mà nó cũng tiếp tục nhận số chuỗi tăng theo luỹ thừa trong các thế hệ kế tiếp

Ta gọi phương trình 4.1 là phương trình tăng trưởng sinh sản của sơ đồ

Trở lại thí dụ, đối với S_0. Vì có ba chuỗi là v_{13}, v_{15} và v_{16} phù hợp với sơ đồ S_0 nên độ thích nghi $eval(S_0)$ của sơ đồ là:

$eval(S_0,t) = (27.316702+30.060205+23.867227)/3 = 27.081378$

đồng thời độ thích nghi trung bình của toàn quần thể là:

$$\overline{F(t)} = \sum_{i=1}^{20} eval(v_i) / pop\_size = 387.776822 / 20 = 19.388841$$

và tỷ lệ thích nghi của sơ đồ S_0 đối với độ thích nghi trung bình của quần thể là:

$$eval(S_0, t) / \overline{F(t)} = 1.396751$$

Điều này có nghĩa là nếu sơ đồ S_0 trên trung bình thì nó nhận số chuỗi tăng theo luỹ thừa trong các thế hệ kế tiếp, đặc biệt nếu sơ đồ S_0 vẫn giữ trên trung bình do nhân với hằng số 1.396751 thì vào thời điểm t+1 ta sẽ có $3\times1.396751 = 4.19$ chuỗi phù hợp với sơ đồ S_0 và vào thời điểm t+2 ta có $3\times1.396751^2 = 5.88$ chuỗi như vậy

Dễ thấy sơ đồ S_0 xác định một phần hứa hẹn của không gian tìm kiếm và số mẫu đại diện của nó trong quần thể sẽ tăng theo luỹ thừa

Ta kiểm tra dự đoán này vào thí dụ đã nêu ở trên, với sơ đồ S_0. Trong quần thể vào thời điểm t, sơ đồ S_0 có ba chuỗi là v_{13}, v_{15} và v_{16} phù hợp. Trong phần 3 ta đã mô phỏng tiến trình *chọn lọc* sử dụng cùng một quần thể để tạo ra quần thể mới. Quần thể mới gồm các nhiễm sắc thể sau:

$v'_1 = v_{11}$ = (01100111111011010110000110111100 0)

$v'_2 = v_4$ = (10001100010110100111100000111001 0)

$v'_3 = v_7$ = (00100010000011010111101101111101 1)

$v'_4 = v_{11}$ = (01100111111011010110000110111100 0)

$v'_5 = v_{19}$ = (00010101001111111111000011000110 0)

$v'_6 = v_4$ = (10001100010110100111100000111001 0)

$v'_7 = v_{15}$ = (11101110110111000010001111101111 0)

$v'_8 = v_5$ = (00011101100101001101011111100010 1)

$v'_9 = v_{11}$ = (01100111111011010110000110111100 0)

$v'_{10} = v_3$ = (00001000001100100000010101110111011101)

$v'_{11} = v_{15}$ = (11101110110111000010001111101111 0)

$v'_{12} = v_9$ = (01100000010110001011000000011111 00)

$v'_{13} = v_6$ = (00010100001001010100101011111 1011)

$v'_{14} = v_8$ = (10000110000111010001011010110 0111)

$v'_{15} = v_{20}$ = (10111001011001111001100010111111 0)

$v'_{16} = v_1$ = (10011010000000111111101001101 1111)

$v'_{17} = v_{10}$ = (00000111100011000001101000011 1011)

$v'_{18} = v_{13}$ = (11101111101000100011000000100011 0)

$v'_{19} = v_{15}$ = (11101110110111000010001111101111 0)

$v'_{20} = v_{16}$ = (11001111000001111110000110100 1011)

Sơ đồ S_0 bây giờ phù hợp với 5 chuỗi v'_7, v'_{11}, v'_{18}, v'_{19}, v'_{20}

Tuy nhiên nếu chỉ riêng phép chọn thì không giới thiệu một điểm mới nào đáng lưu tâm từ không gian tìm kiếm, chọm lọc chỉ sao chép một số chuỗi để hình thành quần thể trung gian. Vì thế bước thứ hai của chu trình tiến hoá là tái kết hợp, nó có nhiệm vụ giới thiệu những cá thể mới trong quần thể. Điều này được thực hiện bởi các phép toán di truyền *lai* và *đột biến*. Ta lần lượt xem xét tác động của hai phép toán này trên một số sơ đồ trong quần thể

4.2. Phép lai

Bắt đầu với phép lai và xét thí dụ sau: Như ta đã biết một chuỗi trong quần thể, chẳng hạn

v'_{18} = (11101111010001000110000001000110), phù hợp với tối đa 2^{33} sơ đồ, cụ thể là chuỗi trên phù hợp với hai sơ đồ sau:

S_0 = (****111**************************)

S_1= (111**************************10)

Giả sử thêm rằng chuỗi này được chọn để thi hành phép lai (thí dụ trong mục 3?? v'_{18} được lai với v'_{13}), vị trí lai phát sinh tạo *pos* = 20. Rõ ràng sơ đồ S_0 vẫn tồn tại nghĩa là vẫn tồn tại một con phù hợp với S_0. Lý do là vị trí lai này bảo tồn chuỗi ‘111’ trên các vị trí thứ 5, 6 và 7 của chuỗi trong một đứa con

Các chuỗi:

v'_{18} = (11101111101000100011|0000001000110)

v'_{13} = (00010100001001010100|1010111111011)

Sẽ sinh ra

v''_{18} = (11101111101000100011|1010111111011)

v''_{13} = (00010100001001010100|0000001000110)

Tuy nhiên sơ đồ S_1 có thể bị phá vỡ, không có con nào phù hợp với nó, lý do là các vị trí cố định '111' ở đầu mẫu và các vị trí cố định '10' ở cuối được đặt vào con khác

Rõ ràng là chiều dài xác định của sơ đồ đóng vai trò quan trọng trong xác suất bị loại bỏ hay tồn tại của sơ đồ. Chú ý rằng chiều dài xác định của sơ đồ S_0 là $\delta(S_0) = 2$, còn chiều dài xác định của sơ đồ S_1 là $\delta(S_1) = 32$

Và các vị trí lai (m-1 vị trí) có cơ hội được chọn ngang nhau. Điều này có nghĩa là xác suất bị loại của sơ đồ S là: $p_d(S) = \frac{\delta(s)}{m-1}$ và do đó xác suất tồn tại là $p_s(S) = 1 - \frac{\delta(s)}{m-1}$

Cụ thể, các xác suất tồn tại của sơ đồ S_0 và S_1 là:

$p_d(S_0) = 2/32$, $p_s(S_0) = 30/32$; $p_d(S_1) = 32/32 = 1$, $p_s(S_1) = 0$

Có một điều quan trọng cần lưu ý là chỉ có một số nhiễm sắc thể trải qua lai và xác suất lai là p_c . Điều này có nghĩa là xác suất tồn tại của sơ đồ thực sự là:

$$p_s(S) = 1 - p_c \frac{\delta(s)}{m-1}$$

Ta xem lại sơ đồ S_0 vẫn với ví dụ đang xét ($p_c = 0.25$)

$p_s(S_0) = 1\text{-}0.25\times 2/32 = 63/64 = 0.984375$

Cũng chú ý rằng, ngay cả khi đã chọn một vị trí lai trong số các vị trí cố định trong một sơ đồ, sơ đồ vẫn có cơ may tồn tại, thí dụ nếu cả hai chuỗi v'_{18} và v'_{13} đều bắt đầu với '111' và tận cùng là '10' thì sơ đồ S_1 vẫn tồn tại (nhưng xác suất của hiện tượng này rất nhỏ), do đó ta nên hiệu chỉnh công thức xác suất tồn tại của sơ đồ:

$$p_s(S) \geq 1 - p_c \frac{\delta(s)}{m-1}$$

Như vậy tác động kết hợp của chọn lọc và lai cho ta một dạng mới của phương trình tăng trưởng của sơ đồ sinh sản:

$$\xi(S,t+1) = \xi(S,t) \times (eval(S,t)/\overline{F(t)}) \times (1 - p_c \times \frac{\delta(S)}{m-1}) \qquad (4.2)$$

Phương trình 4.2 cho biết kỳ vọng số chuỗi phù hợp với sơ đồ *S* trong thế hệ kế tiếp là hàm của số chuỗi đúng của sơ đồ, về độ thích nghi tương đối của sơ đồ và chiều dài xác định của nó. Rõ ràng là sơ đồ trên trung bình có chiều dài ngắn vẫn có thể có số chuỗi cá thể khớp với nó và tốc độ tăng theo luỹ thừa. Đối với sơ đồ S_0:

$$(eval(S,t)/\overline{F(t)})(1 - p_c \times \frac{\delta(S)}{m-1}) = 1.396751 \times 0.984375 = 1.374927$$

Điều này có nghĩa là sơ đồ ngắn, trên trung bình S_0 vẫn nhận được số chuỗi tăng theo luỹ thừa trong các thế hệ tiếp theo, vào thời điểm t+1 ta có 3×1.374927 = 4.12 chuỗi phù hợp với S_0 và vào thời điểm t+2 ta có $3 \times 1.374927^2 = 5.67$ chuỗi

4.3. Đột biến

Phép toán kế tiếp được bàn đến là đột biến. Phép đột biến thay đổi một vị trí trong nhiễm sắc thể một cách ngẫu nhiên với xác suất P_m. Thay đổi từ 0 thành 1 hoặc ngược lại. Rõ ràng tất cả các vị trí cố định của sơ đồ phải giữ không đổi nếu sơ đồ muốn tồn tại qua đột biến, thí dụ xét chuỗi sau trong quần thể:

v'_{19} = (1110111011011100001000111111011110)

và sơ đồ S_0:

S_0 = (****111**************************)

Giả sử thêm rằng chuỗi v'_{19} tham gia đột biến tại vị trí thứ 8 như đã xảy ra trong mục ??, nên kết quả của nó:

v''_{19} = (1110111001011100001000111111011110)

vẫn phù hợp với sơ đồ S_0. Nếu các vị trí đột biến được chọn là từ 1 đến 4, hay từ 8 đến 33 thì chuỗi kết quả vẫn phù hợp với S_0, chỉ 3 bit (5,6,7) là quan trọng, đột biến ít nhất một trong các bit này sẽ loại bỏ sơ đồ S_0. Rõ ràng số những bit quan trong bằng với bậc của sơ đồ, nghĩa là bằng số các vị trí cố định

Vì xác suất thay đổi của môt bit là p_m nên xác suất tồn tại của một bit là 1-p_m. Một lần đột biến độc lập với các đột biến khác, vì thế xác suất tồn tại của sơ đồ *S* qua đột biết (nghĩa là chuỗi các đột biến một bit) là:

$$p_s(S) = (1 - p_m)^{o(S)}$$

Do $p_m << 1$, xác suất này có thể tính gần đúng là

$$p_s(S) \approx 1 - o(S) \times p_m$$

Trở lại với sơ đồ S_0 và thí dụ đạng chạy (p_m = 0.001):

$$p_s(S) \approx 1 - 3 \times 0.001 = 0.97$$

Tác động kết hợp của phương pháp chọn lọc, lại tạo và đột biến cho ta dạng mới của phương trình tăng trưởng sơ đồ sinh sản:

$$\xi(S,t+1) \geq \xi(S,t) \times (eval(S,t)/\overline{F(t)}) \times (1 - p_c \times \frac{\delta(S)}{m-1} - o(S) \times p_m) \qquad (4.3)$$

Cũng như 4.1, 4.2 phương trình 4.3 cũng cho ta biết về kỳ vọng số chuỗi phù hợp với sơ đồ *S* trong thế hệ tiếp theo là hàm theo số chuỗi phù hợp với sơ đồ thích nghi tương đối của sơ đồ, chiều dài xác định và bậc của sơ đồ. Ta cũng thấy rằng các sơ đồ trên trung bình có chiều dài xác định ngắn và bậc thấp vẫn có số chuỗi phù hợp với tốc độ tăng theo luỹ thừa

Đối với sơ đồ S_0:

$$(eval(S,t)/\overline{F(t)})(1 - p_c \times \frac{\delta(S)}{m-1} - o(S) \times p_m) = 1.396751 \times 0.954375 = 1.333024$$

Điều này nghĩa là sơ đồ ngắn, bậc thấp, trên trung bình S_0 vẫn nhận một số chuỗi tăng theo luỹ thừa trong các thế hệ kế tiếp. Tại thời điểm t+1 ta có 3×1.333024 = 4 chuỗi phù hợp với S_0 không kém nhiều so với 4.19 hay 4.12 – giá trị mà ta chỉ tính đến việc lựa chọn hay giá trị mà ta chỉ tính đến việc chọn lọc và lai tạo. Tại thời điểm t+2 ta có $3 \times 1.333024^2 = 5.33$ chuỗi như vậy (cũng không kém 5.85 hoặc 5.67 nhiều)

Tóm lại phương trình tăng trưởng 4.1 cho biết chọn lọc làm tăng tốc độ tạo mẫu của các sơ đồ trên trung bình, và cho biết thay đổi này là theo luỹ thừa. Việc tăng trưởng tự nó không đưa ra sơ đồ mới nào. Điều này chính là lý do mà phép lai được đưa vào để giúp trao đổi thông tin cấu trúc nhưng ngẫu nhiên. Ngoài ra phép lai đưa tính biến thiên cao hơn vào quần thể. Tác động kết hợp (gây rối) của những phép này đối với một sơ đồ không hề quan trọng nếu sơ đồ ngắn và có bậc thấp. Còn phương trình tăng trưởng 4.3 là cơ sở quan trọng của phát biểu sau:

Định lý: Các sơ đồ ngắn, bậc thấp, trên trung bình nhận số chuỗi tăng theo luỹ thừa trong các thế hệ tiếp theo của giải thuật di truyền

Kết quả tức thời của định lý này là GA khảo sát không gian tìm kiếm bằng những sơ đồ ngắn, bậc thấp, do đó những sơ đồ này được dùng để tráo đổi thông tin trong khi lai

5. Các phương pháp biểu diễn nhiễm sắc thể và các toán tử di truyền chuyên biệt

Khi ứng dụng giải thuật di truyền vào thực tế, đôi khi gặp những bài toán đòi hỏi một cách biểu diễn lời giải thích hợp, nếu không giải thuật di truyền khó cho lời giải tốt được, thường là hội tụ sớm về một lời giải tối ưu không toàn cục

Biểu diễn nhị phân truyền thống có một số bất lợi khi áp dụng GA giải các bài toán số cần độ chính xác cao, trong một không gian có số chiều lớn. Thí dụ tối ưu hàm 100 biến, mỗi biến nhận giá trị trong khoảng [-500,500], chính xác đến 6 số lẻ thì chiều dài của véc tơ lời giải nhị phân phải là 3000 và phát sinh một không gian tìm kiếm khoảng 10^{1000} phần tử. Tìm kiếm trong một không gian như thế giải thuật di truyền thực hiện rất kém hiệu quả

Với lý do trên trong phần này chúng ta sẽ thử nghiệm với các gien mã hoá là các số thực cùng với các toán tử di truyền chuyên biệt ứng với cách mã hoá số thực này

5.1. Biểu diễn thực

Trong biểu diễn thực, mỗi véc tơ nhiễm sắc thể được mã hoá thành vectơ thực có cùng chiều dài với véc tơ lời giản. Mỗi phần tử được chọn lúc khởi tạo sao cho thuộc miền xác định của nó, và các toán tử được thiết kế để bảo toàn các ràng buộc này (không có vấn đề như vậy trong biểu diễn nhị phân, nhưng thiết kế của các toán tử này khá đơn giản, ta không thấy điều đó là bất lợi, mặt khác nó lại cung cấp các lợi ích khác được trình bày dưới đây)

Ví dụ: Xét bài toán cực đại hàm 4 biến $f(x_1, x_2,\ldots, x_4)$ với miền ràng buộc:

$x_1 \in [-0.481, 0.519]$, $x_2 \in [-1.851, -0.815]$

$x_3 \in [-4.631, -3.631]$, $x_4 \in [-0.053, 0.053]$

Giả sử kích thước quần thể pop_size = 10, tập hợp véc tơ biểu diễn sẽ là:

s_1 = (-0.470, -1.811, -4.301, -0.051)

s_2 = (-0.130, -1.420, -4.090, -0.031)

s_3 = (-0.221, -0.901, -4.361, -0.010)

s_4 = (-0.370, -0.950, -4.071, -0.051)

s_5 = (-0.320, -0.930, -3.950, -0.031)

s_6 = (-0.351, -0.970, -4.410, -0.011)

s_7 = (-0.471, -0.991, -3.710, -0.030)

s_8 = (-0.030, -0.920, -3.971, -0.011)

s_9 = (-0.071, -0.911, -4.520, -0.011)

s_{10} = (-0.361, -0.901, -4.160, -0.001)

Sự chính xác của cách tiếp cận như thế chỉ tuỳ thuộc máy tính nhưng nói chung là tốt hơn nhiều so với biểu diễn nhị phân. Đương nhiên ta luôn có thể tăng độ chính xác của biểu diễn nhị phân khi thêm các bit, nhưng điều đó làm giải thuật chậm đi đáng kể như đã thảo luận ở phần trước

Thêm nữa biểu diễn thực có khả năng biểu diễn một miền rất rộng (hoặc các trường hợp miền xác định không biết trước cụ thể). Mặt khác trong biểu diễn nhị phân, độ chính xác sẽ giảm khi tăng kích thước miền, do chiều dài nhị phân cố định cho trước. Hơn nữa với biểu diễn thực việc thiết kế các công cụ đặc biệt để xử lý các ràng buộc không tầm thường sẽ dễ hơn

5.2. Các toán tử chuyên biệt hoá

Các toán tử ta sẽ sử dụng rất khác các toán tử cổ điển, vì chúng làm việc trong một không gian khác (có giá trị thực). Hơn nữa một vài toán tử không đồng bộ, nghĩa là hành động của chúng phụ thuộc vào tuổi của quần thể

5.2.1 Nhóm toán tử đột biến

+ Đột biến đồng bộ: Đột biến đồng bộ được định nghĩa tương tự với định nghĩa của phiên bản cổ điển: nếu $s_v = <v_1, \ldots, v_n>$ là nhiễm sắc thể, thì mỗi phần tử v_k có cơ hội trải qua tiến trình đột biến ngang nhau. Kết quả của một lần ứng dụng toán tử này là véc tơ $s^t_v = <v_1, \ldots, v'_k, \ldots, v_n>$ và v'_k là giá trị ngẫu nhiên trong miền tham số tương ứng

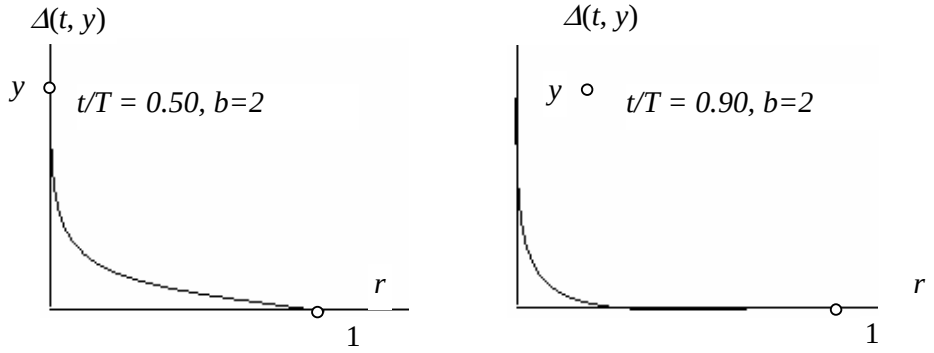
Thí dụ: Giả sử phần tử thứ 3 của véc tơ s_3 = (-0.221, -0.901, -4.361, -0.010) được chọn cho đột biến, biết $x_3 \in$ [-4.631, -3.631] do đó x'_3 được chọn ngẫu nhiên trong miền [-4.631, -3.631] , chẳng hạn x'_3 =-4.12

+ Đột biến không đồng bộ: Đột biến không đồng bộ là một trong những toán tử có nhiệm vụ về tìm độ chính xác của hệ thống. Nó được định nghĩa như sau: nếu $s^t_v = <v_1, \ldots, v_m>$ là nhiễm sắc thể và phần tử v_k được chọn đột biến này (miền của v_k là $[l_k, u_k]$), kết quả là một vectơ $s^{t+1}_v = <v_1, \ldots, v'_k, \ldots, v_m>$ với $k \in [1, \ldots, n]$ và

$$v'_k = \begin{cases} v_k + \Delta(t, u_k - v_k) & \text{neu chu so ngau nhien la } 0 \\ v_k - \Delta(t, v_k - l_k) & \text{neu chu so ngau nhien la } 1 \end{cases}$$

trong đó, hàm $\Delta(t, y)$ trả về giá trị trong khoảng $[0, y]$ sao cho xác suất của $\Delta(t, y)$ gần bằng 0 sẽ tăng khi t tăng. Xác suất này buộc toán tử tìm kiếm không gian thoật đầu là đồng bộ (khi t nhỏ) và rất cục bộ ở những giai đoạn sau. Ta sử dụng hàm sau:

$\Delta(t, y) = y \times (1 - r^{(1-\frac{t}{T})^b})$, với r là số ngẫu nhiên trong khoảng [0..1], T là số thế hệ tối đa và b là tham số hệ thống xác định mức độ không đồng bộ. Hình 2 biểu diễn giá trị của Δ đối với hai lận được chọn, hình này hiển thị rõ ràng cách ứng xử của toán tử



Hình 2. ***Δ(t, y) đối với hai lần chọn***

Hơn nữa ngoài cách áp dụng đột biến chuẩn ta có một số cơ chế mới: đột biến không đồng bộ cũng được áp dụng cho một vectơ lời giải thay vì chỉ một phần tử duy nhất của nó, khiến cho vectơ hơi trượt trong không gian lời giải

Thí dụ: Giả sử phần tử thứ 2 của vectơ s_4 = (-0.370, -0.950, -4.071, -0.051) được chọn cho đột biến, biết $x_2 \in$ [-1.851, -0.815] , lúc đó:

Vì *k* chẵn nên: $x'_2 = x_2 - \Delta(t, (0.950-(-1.815))) = x_2 - \Delta(t, 0.865)$

Giả sử $r = 0.4$, $t / T = 0.5$, $b = 2$ ta có:

$$\Delta(t,0.865) = 0.865 \times (1 - 0.4^{0.5^2}) = 0.865 \times 0.6 = 0.519$$

Do đó x'_2 =-0.950 - 0.519 = -1.469 $\in$ [-1.851, -0.815]

5.2.1 Nhóm toán tử lai tạo

+ **Lai đơn giản**: Phép lai đơn giản được xác định như sau:

Nếu $s^t_v = \langle v_1, \ldots, v_m \rangle$ và $s^t_w = \langle w_1, \ldots, w_m \rangle$ được lai ghép ở vị trí thứ *k*, thì kết quả là: $s^t_v = \langle v_1, \ldots, v_k, w_{k+1}, \ldots, w_m \rangle$ và $s^t_w = \langle w_1, \ldots, w_k, v_{k+1}, \ldots, v_m \rangle$

Thí dụ: Chọn hai nhiễm sắc thể:

s_5 = (-0.320, -0.930, -3.950, -0.031)

s_6 = (-0.351, -0.970, -4.410, -0.011)

Cho lai ghép ở vị trí thứ 3, ta có kết quả:

s'_5 = (-0.320, -0.930, -3.950, -0.011)

s'_6 = (-0.351, -0.970, -4.410, -0.031)

+ **Lai số học đơn**: Phép lai số học đơn được xác định như sau:

Nếu $s^t_v = <v_1, \ldots, v_m>$ và $s^t_w = <w_1, \ldots, w_m>$ được lai ghép thì kết quả là $s^{t+1}_v = <v_1, \ldots, v'_k, \ldots, v_m>$ và $s^{t+1}_w = <w_1, \ldots, w'_k, \ldots, w_m>$, ở đó $1 \le k \le m$, $v'_k = a \times v_k + (1 - a) \times v_k$ và $w'_k = a \times w_k + (1 - a) \times w_k$, a là giá trị động được xác định theo véc tơ s_v và s_w. Chính xác hơn a được chọn trong phạm vi:

$$a \in \begin{cases} [\max(\alpha, \beta), \min(\gamma, \delta)] \\ [0,0] \qquad if\ v_k = w_k \\ \max(\gamma, \delta), \min(\alpha, \beta) \end{cases}$$

Trong đó:

$\alpha = (l_k - w_k)/(v_k - w_k)$, $\beta = (u_k - v_k)/(w_k - v_k)$

$\gamma = (l_k - v_k)/(w_k - v_k)$, $\delta = (u_k - w_k)/(v_k - w_k)$

Thí dụ: Chọn hai nhiễm sắc thể:

s_5 = (-0.320, -0.930, -3.950, -0.031)

s_6 = (-0.351, -0.970, -4.410, -0.011)

Cho lai ghép tại vị trí thứ 3 biết $x_3 \in$ [-4.631, -3.631], ta có kết quả

α = ((-4.631)-(-4.410))/((-3.950)-(-4.410)) = -0.48

β = ((-3.631)-(-3.950))/((-4.410)-(-3.950)) = -0.69

γ = ((-4.631)-(-3.950))/((-4.410)-(-3.950)) = 1.48

δ = ((-3.631)-(-4.410))/((-3.950)-(-4.410)) = 1.69

$a \in$ [-0.48, 1.48]

Giả sử a = 1, ta có:

s'_5 = (-0.320, -0.930, -4.410, -0.031)

s'_6 = (-0.351, -0.970, -3.950, -0.011)

+ **Lai ghép số học toàn cục**: Lai ghép số học toàn cục là tổ hợp tuyến tính của hai véc tơ được xác định như sau:

Nếu $s^t_v = <v_1, \ldots, v_m>$ và $s^t_w = <w_1, \ldots, w_m>$ được lai ghép thì kết quả là $s^{t+1}_v = a \times s^t_w + (1 - a) \times s^t_v$ và $s^{t+1}_w = a \times s^t_v + (1 - a) \times s^t_w$, với a là một tham số tĩnh $\in$ [0,1]

Thí dụ: Chọn 2 nhiễm sắc thể:

s_5 = (-0.320, -0.930, -3.950, -0.031)

s_6 = (-0.351, -0.970, -4.410, -0.011)

Cho lai ghép, giả sử $a = 0.6$

Ta có các phần tử của s'_5 là:

pht1 = (0.351)×0.6+0.4×(-0.320) = 0.083

pht2 = (-0.970)×0.6+0.4×(-0.930) = -0.954

pht3 = (-4.410)×0.6+0.4×(-0.950) = -4.226

pht4 = (-0.011)×0.6+0.4×(0.031) = 0.006

s'_5 = (0.083, -0.954, -4.226, 0.006)

Các phần tử của s'_6 là:

pht1 = (-0.320)×0.6+0.4×(0.351) = -0.052

pht2 = (-0.930)×0.6+0.4×(-0.970) = -0.946

pht3 = (-0.950)×0.6+0.4×(-4.410) = -4.134

pht4 = (0.031)×0.6+0.4×(-0.011) = 0.014

s'_5 = (-0.052, -0.946, -4.134, 0.014)

6. Một số vấn đề của giải thuật di truyền

Với một bài toán được hình thức hoá tốt, ta có thể chứng minh được giải thuật di truyền hội tụ được về lời giải tối ưu. Tuy nhiên các ứng dụng thực tiễn thường khác xa với lý thuyết do

+ Cách biểu diễn nhiễm sắc thể có thể tạo ra không gian tìm kiếm khác với không gian thực sự của bài toán

+ Số bước lặp không đủ vì ta phải xác định trước

+ Kích thước quần thể bị giới hạn

Như vậy trong một số trường hợp, GA không không thể tìm được lời giải tối ưu, nguyên nhân chính là do GA hội tụ sớm về các lời giải tối ưu cục bộ. Hội tụ sớm là vấn đề lớn của giải thuật di truyền cũng như các giải thuật tối ưu khác. Nếu hội tụ quá nhanh, thì các thông tin đáng giá đã phát triển trong quần thể thường bị mất mát

Đã có một số chiến lược được giới thiệu nhằm chống lại sự hội tụ sớm của GA, như (1) chiến lược ghép đôi, được gọi là tránh hỗn giao, (2) sử dụng phép lai đồng dạng hay (3) khám phá những chuỗi trùng lắp trong quần thể

Tuy nhiên, hầu hết những nghiên cứu về lĩnh vực này đề liên quan đến

+ Quy mô và loại sai số do cơ chế tạo mẫu và

+ Bản chất của hàm mục tiêu

6.1. Cơ chế tạo mẫu

Có hai vấn đề quan trọng trong tiến trình tiến hoá của tìm kiếm di truyền: Tính đa dạng quần thể và áp lực chọn lọc. Những yếu tố này liên quan mạnh mẽ với nhau: khi tăng áp lực chọn lọc thì tính đa dạng của quần thể sẽ giảm đi và ngược lại. Nói cách khác, áp lực chọn lọc mạnh ủng hộ tính hội tụ sớm của tìm kiếm GA, nhưng nếu áp lực chọn lọc yếu có thể làm cho tìm kiếm thành vô hiệu. Như vậy việc cân đối giữa hai yếu tố này rất quan trọng, các cơ chế tạo mẫu đều có khuynh hướng đạt đến mục đích này

Back và Hoffmeister đã khảo sát về các đặc trưng của thủ tục chọn lọc. Họ chia các thủ tục chọn lọc thành hai loại: động và tĩnh – chọn lọc tĩnh yêu cầu xác suất chọn lọc là một hằng cho mọi thế hệ (thí dụ, chọn theo hạng), trong khi chọn lọc động thì không cần điều này (chọn lọc theo tỷ lệ). Một cách phân loại khác là chia các thủ tục thành các loại *tiêu diệt* và *bảo tồn* – chọn lọc bảo tồn yêu cầu có xác suất chọn lọc không bằng không đối với mỗi cá thể, trong khi chọn lọc tiêu diệt thì không. Các thủ tục chọn lọc tiêu diệt lại được chia thành hai loại phải và trái: trong chọn lọc tiêu diệt trái các cá thể tốt nhất bị ngăn lại không cho sinh sản để tránh hội tụ sớm do các siêu cá thể (chọn lọc phải thì không). Ngoài ra, một số thủ tục chọn lọc thuần chủng, nghĩa là cha-mẹ chỉ được phép sinh con trong một thế hệ thôi (nghĩa là thời gian sống của mỗi cá thể bị giới hạn trong một thế hệ thôi bất kể độ thích nghi của nó) [2]. Một số chọn lọc có tính thế hệ theo nghĩa tập các cha-mẹ được giữ cố định cho đến khi tất cả các con của thế thệ tiếp theo được sinh ra hoàn toàn, trong các cách chọn lọc trên đường tiến thì con sẽ lập tức thay thế cha-mẹ nó. Một số chọn lọc ưu tú theo nghĩa là một số (hay tất cả) cha-mẹ được phép đồng thời quan chọn lọc cùng với con của chúng

Giải thuật di truyền cải tiến (modGA) được dưới thiệu trong chuyên đề này được trình bày trong [2] . Khác với thuật giải di truyền cổ điển, trong modGA ta không phải thực hiện bước chọn lọc "tái sinh $P'(t)$ từ $P(t\text{-}1)$" mà chọn các nhiễm sắc thể r một cách độc lập (không cần thiết phải khác biệt) để sinh con và các nhiễm sắc thể (phân biệt) bị loại. Sau khi các bước "chọn-cha-mẹ" và "chọn-loại" của modGA đã được thực hiện, có ba nhóm chuỗi (không nhất thiết rời rạc) trong quần thể:

+ r chuỗi (không nhất thiết phân biệt) để sinh sản (cha-mẹ)

+ chính xác r chuỗi bị loại, và

+ các chuỗi còn lại là chuỗi trung bình

Số chuỗi trung tính trong một quần thể (it nhất là $pop\_size - 2r$ và nhiều nhất là $pop\_size - r$) tuỳ thuộc vào số cha-mẹ đã chọn lọc phân biệt và số các chuỗi gối lên nhau trong loại "cha-mẹ" và loại "loại". Rồi quần thể mới $P(t+1)$ được tạo ra, cần có $pop\_size - r$ chuỗi (tất cả các chuỗi, trừ những chuỗi đã chọn để loại) và r con của r cha-mẹ

Thủ tục modGA()

```
{       t ← 0;
        khởi_tạo P(t);
        lượng_giá P(t);
        do {    t ← t+1;
                chọn_ cha_mẹ từ P(t-1);
                chọn_các_cá_thể_ loại_từ P(t-1);
                tạo P(t) // tái sinh cha-mẹ;
                lượng_giá P(t);
            } while (điều_ kiện_dừng);
}
```

Như đã trình bày, giải thuật có một bước khó giải quyết, làm sao chọn r nhiễm sắc thể để loại. Rõ ràng ta muốn thực hiện chọn lọc theo cách mà các nhiễm sắc thể mạnh hơn ít khả năng bị loại hơn, muốn thế ta thay đổi phương thức tạo quần thể mới $P(t+1)$ như sau:

bước 1: chọn r cha-mẹ từ $P(t)$. Mỗi nhiễm sắc thể đã chọn (hay mỗi bản sao của một số nhiễm sắc thể được chọn) được đánh dấu là có thể áp dụng chính xác cho một toán tử di truyền cố định

bước 2: chọn $pop\_size - r$ nhiễm sắc thể khác nhau từ $P(t)$ và sao chép vào $P(t+1)$

bước 3: cho r nhiễm sắc thể cha-mẹ, để có thể sinh được chính xác r con

bước 4: chèn r con mới này vào quần thể $P(t+1)$

Những chọn lọc trên (bước 1,2) được thực hiện tuỳ thuộc độ thích nghi của các nhiễm sắc thể (phương pháp tạo mẫu không gian hỗn loạn). Với phép chọn lọc này ta thấy:

+ Trước tiên cả cha-mẹ và con đều có cơ may để hiện hữu trong thế hệ mới: một cá thể trên trung bình sẽ có nhiều cơ hội để được chọn là cha-mẹ (bước 1) và đồng thời được chọn trong quần thể mới của *pop_size – r* phần tử (bước 2), và như vậy một hoặc nhiều con của nó sẽ nhận một số trong *r* vị trí còn lại

+ Thứ hai, ta áp dụng các toán tử di truyền trên toàn bộ các cá thể tương phản với các bit cá thể (đột biến cổ điển). Điều này có thể tạo ra cách xử lý đồng nhất tất cả các toán tử được dùng trong chương trình tiến hoá. Như vậy nếu ba toán tử được dùng (đột biến, lai tạo, đảo), một số cha-mẹ sẽ qua đột biến, một số khác qua lai tạo và số còn lại bị đảo

Cách tiếp cận modGA có một số đặc tính lý thuyết như thuật giải di truyền cổ điển. Ta có thể viết lại phương trình tăng trưởng 4.3 thành:

$$\xi(S,t+1) \geq \xi(S,t) \times p_s(S) \times p_g(S) \qquad 6.1$$

trong đó $p_s(S)$ biểu diễn xác suất sinh tồn của lược đồ *S* và $p_g(S)$ biểu diễn tăng trưởng của lược đồ *S*. Tăng trưởng của lược đồ *S* xảy ra trong giai đoạn chọn lọc (giai đoạn tăng trưởng) khi nhiều bản sao của các lược đồ trên trung bình được sao chép vào quần thể mới, $p_g(S) = eval(S,t)/\overline{F}$. Các nhiễm sắc thể được chọn phải sống sót qua lai tạo và đột biến $p_s(S) = 1 - p_c \frac{\delta(S)}{m-1} - p_m o(S) < 1$

Phương trình 6.1 hàm ý là lược đồ ngắn, bậc thấp, $p_s(S) \times p_g(S) > 1$ do đó mà những lược đồ như thế nhận số lần thử tăng theo luỹ thừa trong các thế hệ tiếp theo. Điều tương tự cũng đúng cho phiên bản modGA. Số nhiễm sắc thể cần chọn của lược đồ S trong modGA cũng là kết quả của số nhiễm sắc thể trong quần thể cũ $\xi(S,t)$, xác suất sinh tồn $p_s(S) < 1$ và xác suất tăng trưởng $p_g(S)$ - điểm khác biệt duy nhất là sự thông dịc trong các thời kỳ và co rút và bậc tương đối của chúng. Trong phiên bản modGA, thời kỳ co rút là: *n- r* nhiễm sắc thể được chọn cho quần thể mới. Xác suất sinh tồn được định nghĩa là một phần các nhiễm sắc thể của lược đồ *S* không bị chọn loại. Thời kỳ tăng trưởng xảy ra tiếp đó được biểu hiện trong việc xuất hiện của *r* con mới. Xác suất tăng trưởng $p_g(S)$ của lược đồ này do các con mới sinh ra từ *r* cha-mẹ. Lần nữa, đối với các lược đồ ngắn, bậc thấp $p_s(S) \times p_g(S) > 1$ vẫn đúng và những lược đồ đó nhận các lần thử tăng theo luỹ thừa trong các thế hệ kế tiếp

Một trong những ý tưởng về giải thuật modGA là việc sử dụng các tài nguyên lưu trữ tốt hơn: kích thước quần thể. Giải thuật mới tránh được việc để lại những bản sao y hệt của cùng các nhiễm sắc thể trong quần thể mới (điều này cũng có thể xảy ra ngẫu nhiên, nhưng thường rất hiếm). Mặt khác, thuật giải cổ điển rất dễ tạo nhiều phiên bản như thế. Hơn nữa, nhiều sự cố như thế về các siêu cá thể tạo xác suất cho một phản ứng dây truyền: có cơ hội cho một số lượng lớn hơn các bản sao y như thế trong quần thể kế tiếp… Cách này khiển kích thước quần thể bị giới hạn có thể thực sự chỉ biểu diễn một số giảm của cá nhiễm sắc thể đồng nhất. Sử dụng không gian nhỏ sẽ làm giảm hiệu quả của thuận giải. Chú ý rằng cơ sở lý thuyết của thuật giải di truyền giả định kích thước quần thể vô hạn. Trong modGA ta có thể có một số thành viên của gia đình nhiễm sắc thể, nhưng tất cả các thành viên như thế đều khác nhau (ý ta nói đến những con cùng cha mẹ)

Một thí dụ về một nhiễm sắc thể với một giá trị mong đợi trong $P(t+1)$ bằng p = 3, cũng giả định là thuật giải di truyền cổ điển có xác suất lại tạo và đột biến p_c = 0.3 và p_m = 0.003. Sau bước chọn lọc sẽ có chính xác p = 3 bản sao nhiễm sắc thể này trong quần thể $P(t+1)$ trước khi sinh. Sau khi snh, cho chiều dài nhiễm sắc thể bằng 20, số bản sao y hệt của nhiễm sắc thể này còn lại trong $P(t+1)$ sẽ là $p\times(1-p_c-p_m\times m)$ = 1.92. Do đó có thể yên tâm khi nói rằng quần thể kế tiếp sẽ có hai bản sao y của nhiễm sắc thể đó, giảm đi số nhiễm sắc thể khác

6.2. Các đặc trưng của hàm

Thuật giải modGA cung cấp một cơ chế mới tạo một quần thể mới từ quần thể cũ. Tuy nhiên dường như một số độ đo bổ sung có lẽ liên quan đến đặc trưng của hàm đang cần tối ưu. Qua nhiều năm, ta đã thấy được ba hướng cơ bản. Một trong các hướng vay mượn kỹ thuật mô phỏng tôi luyện thép là thay đổi độ hỗn loạn của hệ thống (tốc độ hội tụ quần thể được kiểm soát bằng các toán tử nhiệt động học, các toán tử này được sử dụng tham số nhiệt độ toàn cục)

Một hướng khác là chỉ tái sinh tương ứng với thứ hạng thay vì theo giá trị thực, do đó việc xếp hạng một cách tự động dẫn đến việc định tỷ lệ đồng nhất qua quần thể

Hướng cuối tập trung vào việc thử cố định chính hàm cần tối ưu bằng cách đưa vào một cơ chế định tỷ lệ. Theo Goldberg ta chia các cơ chế ra thành 3 loại

1. Định tỷ lệ tuyến tính: Theo phương pháp này, độ thích nghi các nhiễm sắc thể hiện có được xác định theo công thức: $f_i = a\times f_i + b$, các tham số a, b được chọn sao cho độ thích nghi trung bình được ánh xạ vào chính nó và tăng độ thích nghi tốt nhất bằng cách nhân với độ thích nghi trung bình. Cơ chế này, dù

rất mạnh, có thể tạo ra các giá trị âm cần phải xử lý riêng. Ngoài ra các tham số *a*, *b* thường gắn với đời sống quần thể và không phụ thuộc vào bài toán

2. Phép cắt Sigma: Phương pháp được thiết kế là một cải tiến về định tỷ lệ tuyến tính vừa để xử lý các giá trị âm, vừa để kết hợp thông tin phục thuộc bài toán vào chính ánh xạ. ở đây độ thích nghi mới được tính theo công thức $f_i = a + (\overline{f_i} - c \times \sigma)$, trong đó c là số nguyên nhỏ (thường là một số trong khoảng từ 1 đến 5) còn σ là độ lệnh chuẩn của quần thể, với giá trị âm thì *f* được thiết lập bằng 0

3. Định tỷ lệ luật luỹ thừa: Trong phương pháp này, thích nghi lúc khởi tạo được coi như năng lực đặc biệt $f_i^{'} = f_i^k$ với một số *k* gần bằng 1. Tham số *k* định tỷ lệ hàm *f*, tuy nhiên một số nhà nghiên cứu cho rằng nên chọn *k* độc lập với bài toán, cũng trong các nghiên cứu đó các tác giả đã cho *k* = 1.005

Kết luận:

Chuyên đề đã thực hiện việc nghiên cứu các khái niệm cơ bản của giải thuật di truyền như phép biểu diễn nhiễm sắc thể dưới dạng nhị phân, các toán tử di truyền và cơ chế hoạt động của giải thuật.

Phương pháp biểu diễn nhiễm sắc thể dưới dạng số thực cũng đã được xem xét cùng với các sự thay đổi tương ứng của các toán tử chuyên biệt

Chuyên đề cũng đã đề cập đến lý thuyết sơ đồ, một lý thuyết quan trọng trong giải thuật di truyền

Mục tiêu của các nghiên cứu trên là nắm chắc nền tảng cơ bản của giải thuật, để có thể xây dựng các giải thuật di truyền , ứng dụng vào phát triển các phương pháp lập luận mờ sử dụng đại số gia tử, nội dung quan trọng của luận án

Tài liệu tham khảo

1. Trí tuệ nhân tạo – Lập trình tiến hoá, Nguyễn Đình Thúc, Nhà xuất bản giáo dục, 2001

2. Giải thuật di truyền – cách giải tự nhiên các bài toán trên máy tính, Hoàng Kiếm, Nhà xuất bản khoa học kỹ thuật, 2000