Giải thuật di truyền

Bởi:

Học Viện Công Nghệ Bưu Chính Viễn Thông

Giải thuật di truyền

Nhiễm sắc thể

Các thuật giải di truyền (GAs: Genetic Algorithms) cũng như các thuật toán tiến hoá khác hình thành dựa trên quan niệm cho rằng quá trình tiến hoá tự nhiên là quá trình hợp lý, hoàn hảo. Tự nó đã mang tính tối ưu [12]. Quan điểm trên như một tiên đề, không chứng minh, nhưng phù hợp với thực tế khách quan.

Mục tiêu nghiên cứu của GAs có thể được khái quát như sau:

Trừu tượng hoá và mô phỏng quá trình thích nghi trong hệ thống tự nhiên.

Thiết kế phần mềm, chương trình mô phỏng, nhằm duy trì các cơ chế quan trọng của hệ thống tự nhiên.

Giải thuật di truyền sử dụng một số thuật ngữ của ngành di truyền học [12] như: nhiễm sắc thể, quần thể (Population), Gen.... Nhiễm sắc thể được tạo thành từ các Gen (được biểu diễn của một chuỗi tuyến tính). Mỗi Gen mang một số đặc trưng và có vị trí nhất định trong nhiễm sắc thể. Mỗi nhiễm sắc thể sẽ biểu diễn một lời giải của bài toán.

Các toán tử di truyền

Toán tử sinh sản

Toán tử sinh sản gồm hai quá trình: quá trình sinh sản (phép tái sinh), quá trình chọn lọc (phép chọn).

Phép tái sinh

Phép tái sinh là quá trình các nhiễm sắc thể được sao chép trên cơ sở độ thích nghi. Độ thích nghi là một hàm được gán giá trị thực, tương ứng với mỗi nhiễm sắc thể trong quần thể. Quá trình này, được mô tả như sau:

Xác định độ thích nghi của từng nhiễm sắc thể trong quần thể ở thế hệ thứ t, lập bảng cộng dồn các giá trị thích nghi (theo thứ tự gán cho từng nhiễm sắc thể). Giả sử, quần thể có n cá thể. Gọi độ thích nghi của nhiễm sắc thể i tương ứng là *fi* tổng cộng dồn thứ i là fti được xác định bởi:

$$f_{ti} = \sum_{j=1}^{i} f_j$$

Gọi *Fn* là tổng độ thích nghi của toàn quần thể. Chọn một số ngẫu nhiên *f* trong khoảng từ 0 tới *Fn*. Chọn cá thể thứ k đầu tiên thoả mãn $f \geq ftk$ đưa vào quần thể mới.

Phép chọn

Phép chọn là quá trình loại bỏ các nhiễm sắc thể kém thích nghi trong quần thể. Quá trình này được mô tả như sau:

- Sắp xếp quần thể theo thứ tự mức độ thích nghi giảm dần.

- Loại bỏ các nhiễm sắc thể ở cuối dãy. Giữ lại n cá thể tốt nhất.

Toán tử ghép chéo

Ghép chéo là quá trình tạo nhiễm sắc thể mới trên cơ sở các nhiễm sắc thể cha-mẹ bằng cách ghép một đoạn trên nhiễm sắc thể cha-mẹ với nhau. Toán tử ghép chéo được gán với một xác suất $p\ c$. Quá trình được mô tả như sau:

Chọn ngẫu nhiên một cặp nhiễm sắc thể (cha-mẹ) trong quần thể. Giả sử, nhiễm sắc thể cha-mẹ có cùng độ dài m.

Tạo một số ngẫu nhiên trong khoảng từ 1 tới m-1 (gọi là điểm ghép chéo). Điểm ghép chéo chia nhiễm sắc thể cha-mẹ thành hai chuỗi con có độ dài *m1, m2*. Hai chuỗi con mới được tạo thành là: *m11+ m22 và m21+m12.*

Đưa hai nhiễm sắc thể mới vào quần thể.

Toán tử đột biến

Đột biến là hiện tượng nhiễm sắc thể con mang một số đặc tính không có trong mã di truyền của cha-mẹ.

• Chọn ngẫu nhiên một nhiễm sắc thể trong quần thể;

• Tạo một số ngẫu nhiên k trong khoảng từ 1 tới m, $1 \leq k \leq m$;

• Thay đổi bit thứ k. Đưa nhiễm sắc thể này vào quần thể để tham gia quá trình tiến hoá ở thế hệ tiếp theo.

Các bước cơ bản của giải thuật di truyền

Một giải thuật di truyền đơn giản bao gồm các bước sau:

Bước 1: Khởi tạo một quần thể ban đầu gồm các chuỗi nhiễm sắc thể.

Bước 2: Xác định giá trị mục tiêu cho từng nhiễm sắc thể tương ứng.

Bước 3: Tạo các nhiễm sắc thể mới dựa trên các toán tử di truyền.

Bước 5: Xác định hàm mục tiêu cho các nhiễm sắc thể mới và đưa vào quần thể. Bước 4: Loại bớt các nhiễm sắc thể có độ thích nghi thấp.

Bước 6: Kiểm tra thỏa mãn điều kiện dừng. Nếu điều kiện đúng, lấy ra nhiễm sắc thể tốt nhất, giải thuật dừng lại; ngược lại, quay về bước 3.

Cơ sở toán học của giải thuật di truyền

Cơ sở lý thuyết của giải thuật di truyền dựa trên biểu diễn chuỗi nhị phân và lý thuyết sơ đồ [12]. Một sơ đồ là một chuỗi, có chiều 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 Gen hoặc một ký tự đại diện ”*”. Sơ đồ biểu diễn không gian con trong không gian tìm kiếm. Không gian con này là tập tất cả các chuỗi trong không gian tìm kiếm mà với mọi vị trí trong chuỗi, giá trị của Gen trùng với giá trị của sơ đồ; kí tự đại diện “*” có thể trùng khớp với bất kỳ ký tự biểu diễn nào.

Sơ đồ (* 1 0 1 0) sẽ khớp với 2 chuỗi: (1 1 0 1 0) và (0 1 0 1 0)

Như vậy, sơ đồ (1 1 0 1 0) và (0 1 0 1 0) chỉ khớp với chuỗi chính nó, còn sơ đồ (* * ** *)khớp với tất cả các sơ đồ có độ dài là 5.

Với sơ đồ cụ thể có tương ứng 2r chuỗi, r: là số ký tự đại diện “*” có trong sơ đồ; ngược lại, một chuỗi có chiều dài m sẽ khớp với 2m sơ đồ.

Một chuỗi có chiều dài m, sẽ có tối đa 3m sơ đồ. Trong một quần thể dân số kích thước n, có thể có tương ứng từ 2m đến nx2m sơ đồ khác nhau.

Thuộc tính của sơ đồ

Các sơ đồ khác nhau có 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 o(S)) là tổng số vị trí 0, 1 có trong sơ đồ. Đây là các vị trí cố định (không phải vị trí của các ký tự đại diện) trong sơ đồ. Bậc có thể xác định bằng cách lấy chiều dài của chuỗi trừ đi số ký tự đại diện.

Trong sơ đồ S = (* * 1 0 * 1 *) có bậc o(S) = 7 - 4 = 3;

Chiều dài xác định của sơ đồ S (ký hiệu là δ(S)) là khoảng cách giữa 2 vị trí cố định ở đầu và cuối. Chiều dài của sơ đồ xác định độ nén thông tin chứa trong sơ đồ đó. Trong ví dụtrên δ(S)= 6-3 =3. Như vậy, nếu sơ đồ chỉ có một vị trí cố định thì chiều dài xác định của sơđồ sẽ bằng 0.

Chiều dài của sơ đồ giúp ta tính xác suất tồn tại của sơ đồ do ảnh hưởng của ghép chéo.

Đặc điểm hội tụ của giải thuật di truyền

Khi áp dụng giải thuật GAs cho các vấn đề thực tế thường rất khó khăn. Lý do:

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

• Số bước lặp, khi cài đặt thường không xác định trước;

• Kích thước quần thể thường có giới hạn.

Trong một số trường hợp, GAs không thể tìm được lời giải tối ưu. Lý do, GAs hội tụ sớm về lời giải tối ưu cục bộ. Hội tụ sớm là vấn đề của giải thuật di truyền cũng như các giải thuật tối ưu khác. Nếu hội tụ xảy ra quá nhanh thì các thông tin đáng tin cậy đang phát triển trong quần thể thường bị bỏ qua. Nguyên nhân của sự hội tụ sớm liên quan tới hai vấn đề:

• Quy mô và loại sai số do cơ chế tạo mẫu;

• Bản chất của hàm mục tiêu.

Cơ chế tạo mẫu

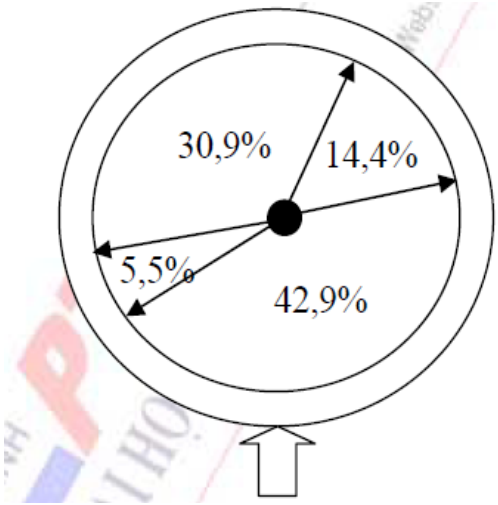
Có hai vấn đề quan trọng trong tiến trình tiến hoá của giải thuật di truyền là: tính đa dạng của quần thể và áp lực chọn lọc [12]. Hai yếu tố này liên quan mật thiết với nhau: khi tăng áp lực chọn lọc thì tính đa dạng của quần thể sẽ giảm và ngược lại. Nói cách

khác, áp lực hội tụ mạnh sẽ dẫn tới sự hội tụ sớm của giải thuật. Nhưng nếu áp lực chọn lọc yếu có thể làm cho tìm kiếm thành vô hiệu. Như vậy, cần thoả hiệp hai vấn đề. Hiện nay, các phương pháp đưa ra đều có khuynh hướng dể đạt tới mục đích này.

Năm 1975 DeJong đã xem xét một số biến thể của chọn lọc đơn giản bằng cách đưa ra: mô hình phát triển ưu tú, mô hình giá trị mong đợi và mô hình nhân tố tập trung.

Năm 1981 Brindle xem xét một số biến thể khác như: tạo mẫu tất định, tạo mẫu hỗn loạn, tạo mẫu hỗn loạn phần dư không thay thế, đấu tranh hỗn loạn, tạo mẫu hỗn loạn phần dư có thay thế.

Năm 1987 Baker nghiên cứu phương pháp tạo mẫu không gian hỗn loạn. Phương pháp này dùng cách "quay" bánh xe định tỷ lệ trước để thực hiện chọn lọc. Bánh xe này được thiết kế theo chuẩn, quay với số khoảng chia đều theo kích thước quần thể.



Tỷ lệ thích nghi của các nhiễm sắc thể trên bánh xe Roulette

Người ta thực hiện việc sinh sản bằng cách quay bánh xe Roulette với số lần bằng số nhiễm sắc thể trên bánh xe Roulette. Đối với bài toán này số lần quay bánh xe Roulette là 4. Nhiễm sắc thể 1 có giá trị thích nghi là 169, tương ứng 14,4 % tổng độ thích nghi. Như vậy, nhiễm sắc thể 1 chiếm 14.4% trên bánh xe Roulette. Mỗi lần quay nhiễm sắc thể 1 sẽ chiếm khe với giá trị 0,144.

Khi yêu cầu sinh ra 1 thế hệ mới, một vòng quay của bánh xe Roulette được đánh trọng số phù hợp sẽ chọn ra một cá thể để sinh sản. Bằng cách này, những nhiễm sắc thể có độ thích nghi cao sẽ có cơ hội được chọn lớn. Như vậy, sẽ có 1 số lượng con cháu lớn trong các thế hệ kế tiếp.

Hàm mục tiêu

Cứ sau mỗi thế hệ được hình thành, chúng ta cần tính lại độ thích nghi cho từng cá thể để chuẩn bị cho một thế hệ mới. Do số lượng các cá thể tăng lên, độ thích nghi giữa các cá thể không có sự chêch lệch đáng kể. Do đó, các cá thể có độ thích nghi cao chưa hẳn chiếm ưu thế trong thế hệ tiếp theo. Vì vậy, cần ấn định tỷ lệ đối với hàm thích nghi nhằm tăng khả năng cho các nhiễm sắc thể đạt độ thích nghi cao. Có 3 cơ chế định tỷ lệ như sau.

Định tỷ lệ tuyến tính

Độ thích nghi được xác định theo công thức:

$$f_i^{'} = a * f_i + b$$

Cần chọn các tham số a, b sao cho độ thích nghi trung bình được ánh xạ vào chính nó. Tăng độ thích nghi tốt nhất bằng cách nhân nó với độ thích nghi trung bình. Cơ chế này có thể tạo ra các giá trị âm cần xử lý riêng. Ngoài ra, các tham số a, b thường gắn với đời sống quần thể và không phụ thuộc vào bài toán.

Phép cắt Sigma

Phương pháp này được thiết kế vừa để cải tiến phương pháp định tỷ lệ tuyến tính vừa để xử lý các giá trị âm, vừa kết hợp thông tin mà bài toán phụ thuộc. Ở đây, độ thích nghi mới được tính theo công thức:

$$f_i^{'} = f_i + (\overline{f} - c * \sigma)$$

trong đó c là một số nguyên nhỏ (thường lấy giá trị từ 1 tới 5);σ là độ lệch chuẩn của quần thể. Với giá trị âm thì f' được thiết lập bằng 0.

Định tỷ lệ cho luật dạng luỹ thừa

Trong phương pháp này, độ thích nghi lúc khởi tạo có năng lực đặc biệt:

$$f_i^{'} = f_i^{k}$$

với 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. Bằng thực nghiệm cho thấy nên chọn k =1.005.

Điều kiện dừng của giải thuật

Chúng ta sẽ khảo sát điều kiện đơn giản nhất để dừng khi số thế hệ vượt quá một ngưỡng cho trước. Trong một số phiên bản về chương trình tiến hoá không phải mọi cá thể đều tiến hoá lại. Vài cá thể trong đó có khả năng vượt từ thế hệ này sang thế hệ khác mà không thay đổi gì cả. Trong những trường hợp như vậy, chúng ta đếm số lần lượng hàm. Nếu số lần lượng hàm vượt quá một hằng xác định trước thì dừng việc tìm kiếm.

Chúng ta nhận thấy, các điều kiện dừng ở trên giả thiết rằng người sử dụng đã biết đặc trưng của hàm, có ảnh hưởng như thế nào tới chiều dài tìm kiếm. Trong một số trường hợp khó có thể xác định số lượng thế hệ (hay lượng giá hàm) phải là bao nhiêu. Giải thuật có thể kết thúc khi cơ hội cho một cải thiện quan trọng chưa bắt đầu.

Có hai loại điều kiện dừng cơ bản. Các điều kiện này dùng các đặc trưng tìm kiếm để quyết định ngừng quá trình tìm kiếm .

Dựa trên cấu trúc nhiễm sắc thể: do sự hội tụ của quần thể bằng cách kiểm soát số alen được hội tụ, ở đây alen được coi như hội tụ nếu một số phần trăm quần thể đã định trước có cùng (hoặc tương đương đối với các biểu diễn không nhị phân) giá trị trong alen này. Nếu số alen hội tụ vượt quá số phần trăm nào đó của tổng số alen, việc tìm kiếm sẽ kết thúc.

Dựa trên ý nghĩa đặc biệt của một nhiễm sắc thể: đo tiến bộ của giải thuật trong một số thế hệ cho trước. Nếu tiến bộ này nhỏ hơn một hằng số ε xác định, kết thúc tìm kiếm.