ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

ĐẶNG THỊ MINH PHƯƠNG

BIỂU DIỄN NHIỄM SẮC THỂ TRONG GIẢI THUẬT DI TRUYỀN VÀ CÁC TOÁN TỬ DI TRUYỀN CHUYÊN BIỆT

Chuyên ngành: Khoa học máy tính
Mã số: 60.48.01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2012

LỜI NÓI ĐẦU

Đặt vấn đề

Cho đến nay đã có nhiều thuật toán tìm lời giải tối ưu cho nhiều lĩnh vực bài toán, ví dụ như trong bài toán tìm kiếm trên danh sách, cây, đồ thị các nhà khoa học đã đưa ra thuật toán tìm kiếm quay lui, vét cạn. Các thuật toán này tuy tìm được nghiệm tối ưu nhưng chỉ áp dụng được cho các bài toán có không gian tìm kiếm nhỏ.

Để khắc phục các hạn chế như trên các nhà khoa học cũng đã đưa ra các thuật toán tìm kiếm heurictics, đây là thuật toán có sử dụng các tri thức về lĩnh vực bài toán để nhằm giảm thời gian tìm kiếm. Tuy nhiên các thuật toán này lại vấp phải một vấn đề là các tri thức thường là kinh nghiệm của con người, do đó nó có thể chưa chính xác, đầy đủ và điều này có thể dẫn tới sự chệch hướng trong quá trình tìm kiếm.

Giải thuật di truyền là một trong những kỹ thuật tìm kiếm tối ưu giúp ta giải quyết được những vấn đề đã đặt ra ở trên, nó cho phép ta tìm kiếm lời giải tối ưu trên các không gian lớn, nguyên tắc cơ bản của giải thuật di truyền là mô phỏng quá trình chọn lọc của tự nhiên. Cho đến nay lĩnh vực nghiên cứu về giải thuật di truyền đã thu được nhiều thành tựu, giải thuật di truyền được ứng dụng trong nhiều lĩnh vực phức tạp, các vấn đề khó có thể giải quyết được bằng phương pháp thông thường.

Với những khả năng tiềm tàng của giải thuật di truyền đã là động lực và lý do chính để tác giả chọn đề tài ***"Biểu diễn nhiễm sắc thể trong giải thuật di truyền và các toán tử di truyền chuyên biệt".***

Mục tiêu của đề tài

- Nghiên cứu các khái niệm cơ bản của giải thuật di truyền.

- Nghiên cứu một số phương pháp biểu diễn nhiễm sắc thể trong giải thuật di truyền và các toán tử di truyền tương ứng.

- Nghiên cứu lựa chọn một số bài toán tối ưu và ứng dụng giải thuật di truyền để giải quyết các bài toán này.

Phạm vi của đề tài

- Nghiên cứu các khái niệm cơ bản của giải thuật di truyền.

- Nghiên cứu giải thuật di truyền sử dụng phương pháp biểu diễn nhiễm sắc thể bằng mã hóa nhị phân và các toán tử di truyền tương ứng.

- Nghiên cứu giải thuật di truyền sử dụng phương pháp biểu diễn nhiễm sắc thể bằng mã hóa số thực và các toán tử di truyền tương ứng.

- Nghiên cứu phương pháp biểu diễn nhiễm sắc thể bằng một hoán vị của một tập hợp.

- Ứng dụng giải thuật di truyền sử dụng mã hóa nhị phân và giải thuật di truyền sử dụng mã hóa số thực để xác định độ rộng của các tập mờ trong bài toán xấp xỉ mô hình mờ của Cao-Kandel.

Chương 1

CÁC KHÁI NIỆM CƠ BẢN VỀ GIẢI THUẬT DI TRUYỀN

1.1. Mở đầu

Giải thuật di truyền (Gennetic Algorithm) là giải thuật tìm kiếm, chọn lựa các giải pháp tối ưu để giải quyết các bài toán khác nhau dựa trên cơ chế chọn lọc tự nhiên của ngành di truyền học.

Trong cơ thể sinh vật, các gen liên kết với nhau theo cấu trúc dạng chuỗi gọi là nhiễm sắc thể, nó đặc trưng cho mỗi loài và quyết định sự sống còn của cơ thể đó.

Một loài muốn tồn tại phải thích nghi với môi trường, cơ thể sống nào thích nghi với môi trường hơn thì sẽ tồn tại và sinh sản với số lượng ngày càng nhiều hơn, trái lại những loài không thích nghi với môi trường sẽ dần dần bị diệt chủng.

Môi trường tự nhiên luôn biến đổi, nên cấu trúc nhiễm sắc thể cũng thay đổi để thích nghi với môi trường và ở thế hệ sau luôn có độ thích nghi cao hơn ở thế hệ trước. Cấu trúc này có được nhờ vào sự trao đổi thông tin ngẫu nhiên với môi trường bên ngoài hay giữa chúng với nhau.

Dựa vào đó các nhà khoa học máy tính xây dựng nên một giải thuật tìm kiếm tinh tế dựa trên cơ sở chọn lọc tự nhiên và quy luật tiến hóa gọi là giải thuật di truyền.

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. [1, 2]

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

1.2.1. 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 di truyền, 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. [1, 2]

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).

1.2.2. Giải thuật di truyền đơn giản [1, 2, 3]

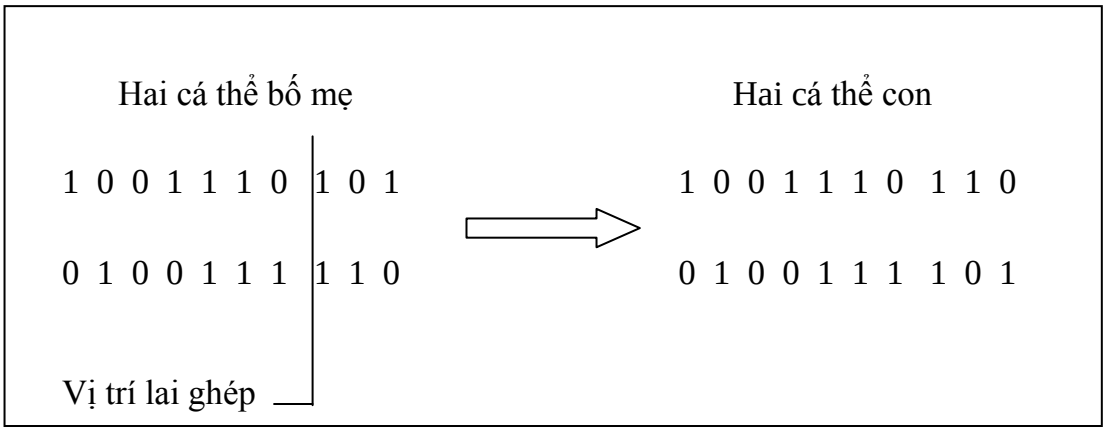
Trong giải thuật di truyền của mình J. H. Holland sử dụng mã hoá nhị phân để biểu diễn các cá thể, lý do là phần lớn các bài toán tối ưu hoá đều có thể được mã hoá thành chuỗi nhị phân khá đơn giản.

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 1.1. Sơ đồ lai ghép 1 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 trên.

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() trình bày dưới đây.

Thủ tục SGA () /* *Giải bài toán tối ưu* */

{ k = 0;

// Khởi tạo quần thể P_0 một cách ngẫu nhiên.

khởi_tạo (P_k);

// Tính giá trị hàm mục tiêu cho từng cá thể.

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; G - số thế hệ cần tiến hoá.

Đó chính 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.

Ví dụ: xét bài toán tìm max của hàm $f(x) = x^2$ với x là số nguyên trên đoạn [0, 31].

Để sử dụng giải thuật di truyền ta mã hóa mỗi số nguyên x trong đoạn

[0, 31] bởi một số nhị phân có độ dài 5, chẳng hạn chuỗi 11000 là mã của số nguyên 24.

Hàm thích nghi được xác định chính là hàm $f(x)=x^2$

Quần thể ban đầu gồm 4 cá thể (kích thước quần thể n=4).

Thực hiện quá trình chọn lọc ta có bảng sau, trong bảng này ta thấy cá thể 2 có độ thích nghi cao nhất nên nó được chọn 2 lần, cá thể 3 có độ thích nghi thấp nhất không được chọn lần nào, mỗi cá thể 1 và 4 được chọn 1 lần

Số hiệu cá thể	**Quần thể ban đầu**	**x**	**Độ thích nghi $f(x)=x^2$**	**Số lần được chọn**
1	0 1 1 0 1	13	169	1
2	1 1 0 0 0	24	576	2
3	0 1 0 0 0	8	64	0
4	1 0 0 1 1	19	361	1