Học Máy
(IT 4862)

Nguyễn Nhật Quang

quangnn-fit@mail.hut.edu.vn

Trường Đại học Bách Khoa Hà Nội

Viện Công nghệ thông tin và truyền thông

Năm học 2011-2012

Nội dung môn học:

- Giới thiệu chung
- Đánh giá hiệu năng hệ thống học máy
- Các phương pháp học dựa trên xác suất
- **Các phương pháp học có giám sát**
  - **Giải thuật di truyền (Genetic algorithm)**
- Các phương pháp học không giám sát
- Lọc cộng tác
- Học tăng cường

Giải thuật di truyền – Giới thiệu

- Dựa trên (bắt chước) quá trình tiến hóa tự nhiên trong sinh học
- Áp dụng phương pháp tìm kiếm ngẫu nhiên (stochastic search) để tìm được lời giải (vd: một hàm mục tiêu, một mô hình phân lớp, …) tối ưu
- Giải thuật di truyền (Generic Algorithm – GA) có khả năng tìm được các lời giải tốt thậm chí ngay cả với các không gian tìm kiếm (lời giải) không liên tục rất phức tạp
- Mỗi khả năng của lời giải được biểu diễn bằng một chuỗi nhị phân (vd: 100101101) – được gọi là **nhiễm sắc thể (chromosome)**
  - Việc biểu diễn này phụ thuộc vào từng bài toán cụ thể
- GA cũng được xem như một bài toán học máy (*a learning problem*) dựa trên quá trình tối ưu hóa (*optimization*)

Giải thuật di truyền – Các bước chính

- Xây dựng (khởi tạo) **quần thể (population) ban đầu**
  - Tạo nên một số các giả thiết (khả năng của lời giải) ban đầu
  - Mỗi giả thiết khác các giả thiết khác (vd: khác nhau đối với các giá trị của một số tham số nào đó của bài toán)
- Đánh giá quần thể
  - Đánh giá (cho điểm) mỗi giả thiết (vd: bằng cách kiểm tra độ chính xác của hệ thống trên một tập dữ liệu kiểm thử)
  - Trong lĩnh vực sinh học, điểm đánh giá này của mỗi giả thiết được gọi là **độ phù hợp (fitness)** của giả thiết đó
  - Xếp hạng các giả thiết theo mức độ phù hợp của chúng, và chỉ giữ lại các giả thiết tốt nhất (gọi là **các giả thiết phù hợp nhất – survival of the fittest**)
- Sản sinh ra **thế hệ tiếp theo (next generation)**
  - Thay đổi ngẫu nhiên các giả thiết để sản sinh ra thế hệ tiếp theo (gọi là **các con cháu – offspring**)
- Lặp lại quá trình trên cho đến khi ở một thế hệ nào đó có giả thiết tốt nhất có độ phù hợp cao hơn giá tri phù hợp mong muốn (định trước)

GA(Fitness, θ, n, r_{co}, r_{mu})

Fitness: A function that produces the score (fitness) given a hypothesis

θ: The desired fitness value (i.e., a threshold specifying the termination condition)

n: The number of hypotheses in the population

r_{co}: The percentage of the population influenced by the *crossover* operator at each step

r_{mu}: The percentage of the population influenced by the *mutation* operator at each step

Initialize the population: H ← Randomly generate n hypotheses

Evaluate the initial population. For each h∈H: compute Fitness(h)

while ($\max_{\{h \in H\}}$ Fitness(h) < θ) do

$H^{next} \leftarrow \varnothing$

Reproduction (Replication). Probabilistically select $(1-r_{co}).n$ hypotheses of H to add to H^{next}.

The probability of selecting hypothesis h_i from H is:

$$P(h_i) = \frac{Fitness(h_i)}{\sum_{j=1}^{n} Fitness(h_j)}$$

GA(Fitness, θ, n, r_{co}, r_{mu})

…

Crossover.

Probabilistically select (r_{co}.n/2) pairs of hypotheses from H, according to the probability computation P(h_i) given above.

For each pair (h_i, h_j), produce two offspring (i.e., children) by applying the crossover operator. Then, add all the offspring to H^{next}.

Mutation.

Select (r_{mu}.n) hypotheses of H^{next}, with uniform probability.

For each selected hypothesis, invert one randomly chosen bit (i.e., 0 to 1, or 1 to 0) in the hypothesis's representation.

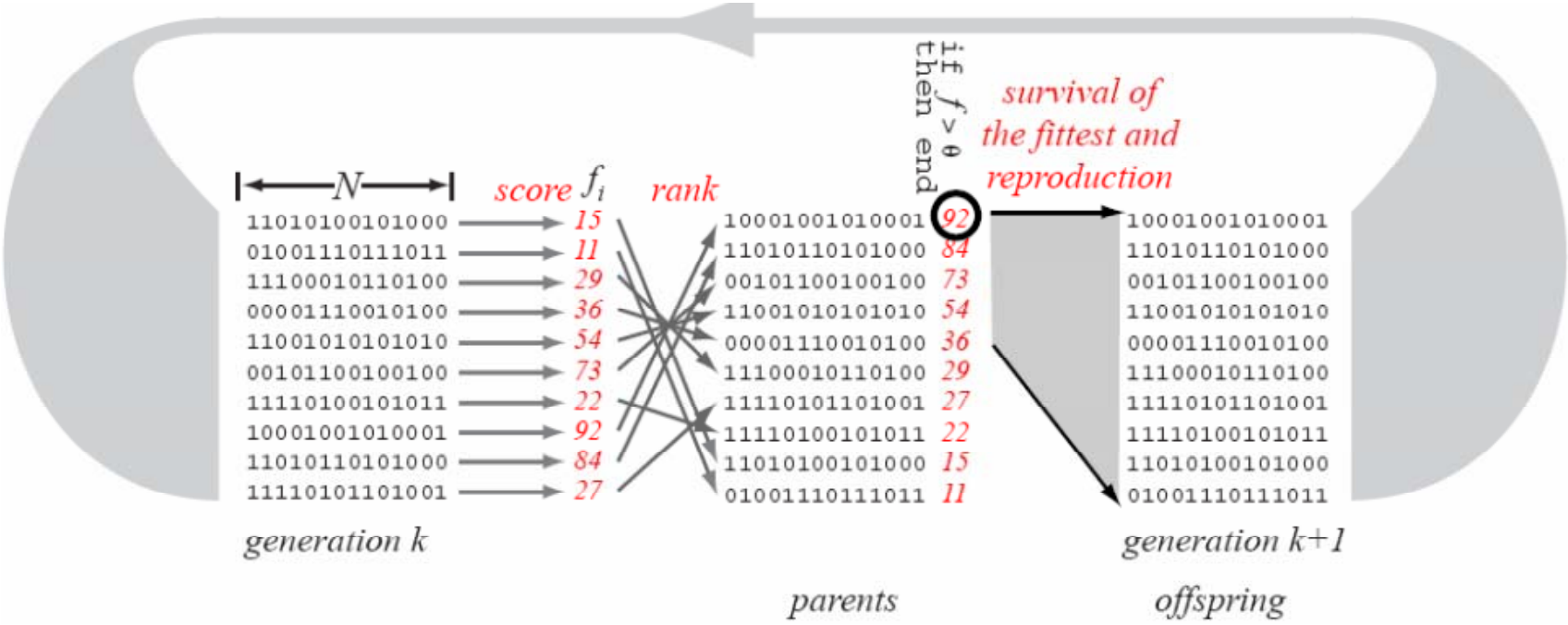
Producing the **next generation**: $H \leftarrow H^{next}$

Evaluate the new population. For each $h \in H$: compute Fitness(h)

end while

return $\text{argmax}_{\{h \in H\}}$Fitness(h)

Giải thuật di truyền – Minh họa



[Duda et al., 2000]

Các toán tử di truyền

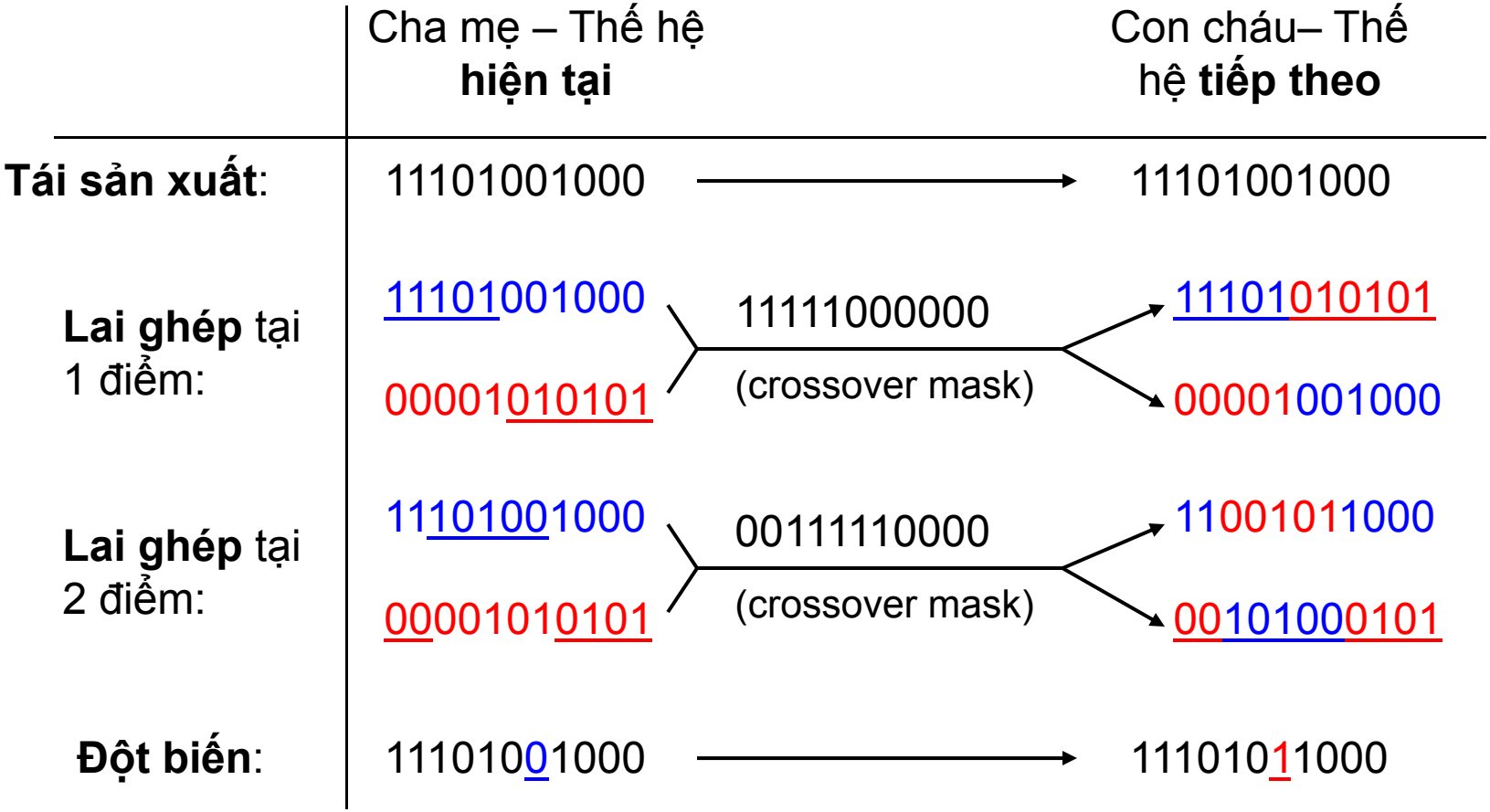
- 3 toán tử di truyền được sử dụng để sinh ra các cá thể con cháu (offspring) trong thế hệ tiếp theo
  - Nhưng chỉ có 2 toán tử lai ghép (*crossover*) và đột biến (*mutation*) tạo nên sự thay đổi
- **Tái sản xuất (Reproduction)**

  → Một giả thiết được giữ lại (không thay đổi)
- **Lai ghép (Crossover)** để sinh ra 2 cá thể mới

  → Ghép ("phối hợp") của hai cá thể cha mẹ
  - Điểm lai ghép được chọn ngẫu nhiên (trên chiều dài của nhiễm sắc thể)
  - Phần đầu tiên của nhiễm sắc thể h_i được ghép với phần sau của nhiễm sắc thể h_j, và ngược lại, để sinh ra 2 nhiễm sắc thể mới
- **Đột biến (Mutation)** để sinh ra 1 cá thể mới

  →Chọn ngẫu nhiên một bit của nhiễm sắc thể, và đổi giá trị (0→1 / 1→0)
  - Chỉ tạo nên một thay đổi nhỏ và ngẫu nhiên đối với một cá thể cha mẹ!

Các toán tử di truyền – Ví dụ

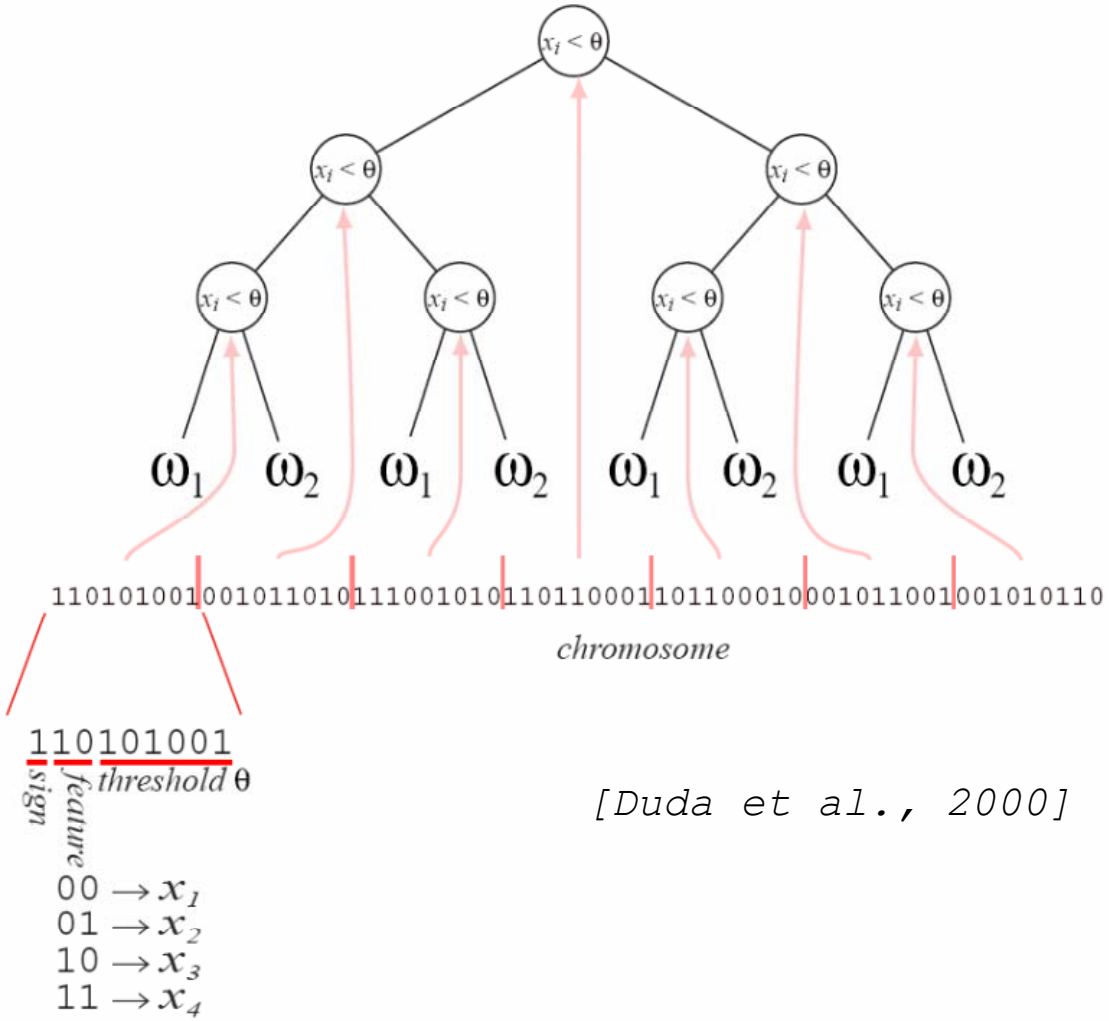


[Mitchell, 1997]

Biểu diễn giả thiết – Ví dụ

Ánh xạ (chuyển đổi) giữa:

- Biểu diễn các nhiễm sắc thể (chuỗi nhị phân), và
- Biểu diễn cây quyết định cho bài toán phân lớp có 2 lớp



[Duda et al., 2000]

Tài liệu tham khảo

- T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.
- R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.