Dobývání znalostí

Doc. RNDr. Iveta Mrázová, CSc.

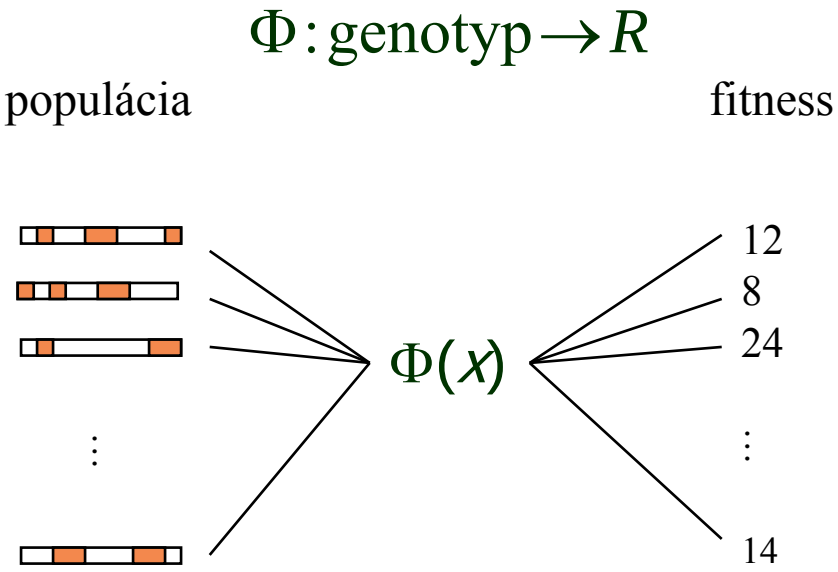
Katedra teoretické informatiky
Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

Genetické algoritmy (GA)

- Holland 60. roky 20. stor.
- **Populácia** umelých chromozómov sa cyklicky podrobuje
  - selektívnej reprodukcii preferujúcej výkonnejších jedincov a
  - náhodným zmenám
- **Umelý chromozóm** (genotyp) = reťazec symbolov kódujúci vlastnosti jedinca (fenotyp)
  - napr. binárna hodnota premennej alebo postupnosť hodnôt premenných
  - mnoho typov kódovaní
    - binárne, Greyov kód, reálne hodnoty
  - abecedy – napr. binárna, ternárna, …

GA – fitness funkcia

- **Fitness funkcia (účelová, cieľová funkcia)** – kritérium výkonnosti, zobrazenie: genotyp → reálne (celé) číslo
  - čím vyššia hodnota, tým je jedinec lepší



Genetické algoritmy

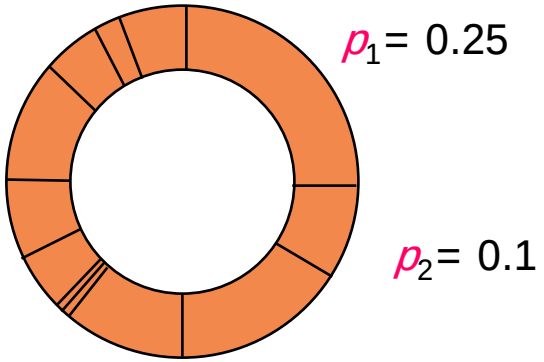
Jednoduchý genetický algoritmus (Goldberg 1989)

- vytvor populáciu N náhodne vygenerovaných chromozómov $x_1, x_2, \ldots, x_N$
- opakuj
  - dekóduj všetky chromozómy a spočítaj ich fitness $f_i = \Phi(x_i)$
  - vytvor novú populáciu selektívnou reprodukciou
  - rekombinuj chromozómy – kríženie
  - mutuj chromozómy
- skonči, keď sa objaví hľadaný jedinec, alebo fitness najlepšieho nerastie

Jednoduchá selekcia

- Zo starej populácie vytvárame novú kopírovaním chromozómov tak, že čím lepší jedinec, tým viacej jeho kópií sa môže objaviť v novej populácii
- reprodukcia **ruletou**:
  - každý jedinec dostane na kole rulety priehradku veľkosti p_i úmernej jeho fitness funkcii – to bude pravdepodobnosť, že bude reprodukovaný (predpokladáme nezápornú fitness funkciu)

$$p_i = \frac{f_i}{\sum_{j=1}^{n} f_j}$$



  - ruletou sa otočí N krát

Varianty selekcie (1)

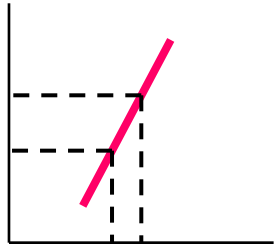
- **Problémy jednoduchej selekcie**

  a) keď všetci jedinci majú podobnú fitness

  náhodné prehľadávanie s genetickým posunom

  b) keď jeden až dvaja jedinci majú fitness o mnoho vyššiu než zvyšok populácie

  takmer všetci jedinci v novej generácii budú kópiou toho istého jedinca, predčasná konvergencia



- **Riešenie:**

  1. **škálovanie [scaling]** (typicky lineárna transformácia)

  pre a) zväčšiť rozdiely, pre b) zmenšiť

  2. **selekcia podľa poradia [rank based]** – jedinci sa zoradia podľa fitness a pravdepodobnosť reprodukcie je úmerná poradiu jedinca, nie jeho fitness

Varianty selekcie (2)

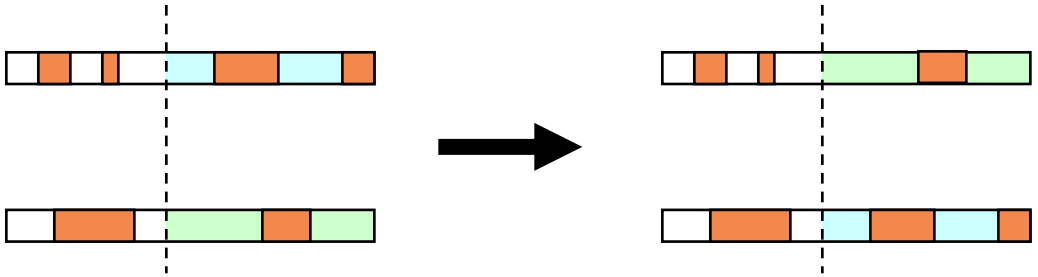
3. **S orezávaním [truncation]** – jedincov usporiadame podľa veľkosti fitness, M najlepších jedincov okopírujeme O - krát tak, že $N = M \times O$
4. **Turnajom [tournament]** – náhodne sa vyberú 2 jedinci a vygeneruje sa náhodné číslo $r \in \langle 0, 1 \rangle$, ak je $r < T$, kde $T \in \langle 0, 1 \rangle$ je preddefinovaný parameter, tak bude okopírovaný jedinec s vyššou fitness inak ten druhý

- Pri ďalších genetických operáciách sa môže doteraz najlepší jedinec stratiť. Pomoc: **elitizmus [elitism]** – S najlepších jedincov je bez zmeny okopírovaných do novej generácie

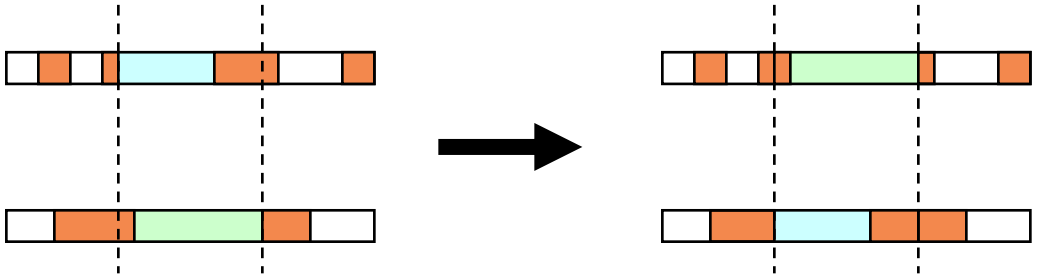
Kríženie

- Jedinci sú náhodne spárovaní a každý pár je s danou pravdepodobnosťou skrížený

Jednobodové kríženie

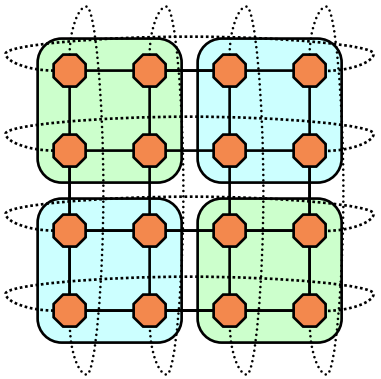


Viacbodové kríženie



Mutácia, distribuované GA

- Každý prvok chromozómu je s danou pravdepodobnosťou zmenený
  - Bit sa neguje, v prípade iných oborov hodnôt sa nahradí náhodnou hodnotou z daného oboru hodnôt, alebo sa pričíta náhodná hodnota podľa nejakého rozdelenia so stredom 0
- Jedinci v populácii sú rozmiestnení napr. v dvojrozmernom priestore (napr. toroid); selekcia a kríženie sa dejú iba lokálne, subpopulácie sa prekrývajú – tým je možné šírenie „dobrých“ vlastností cez celú populáciu

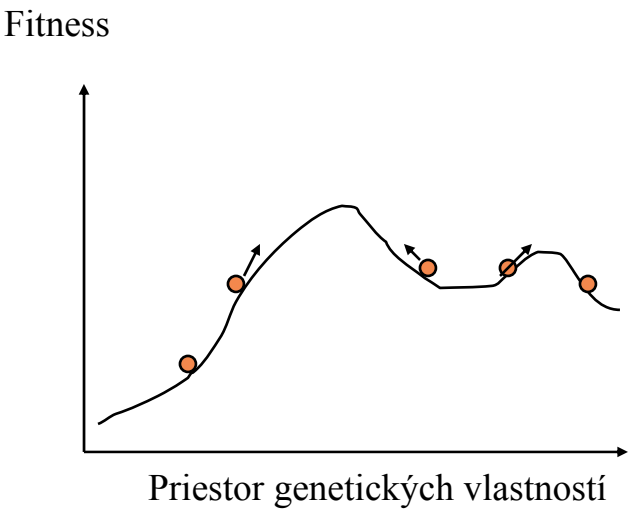


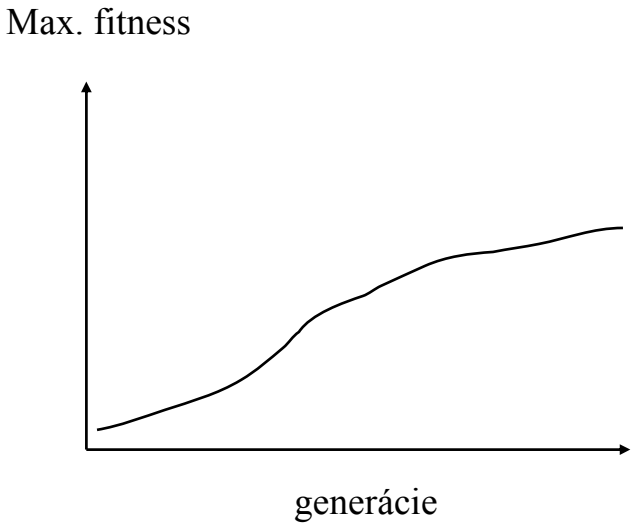
Nedeterminizmus

- Celý GA je nedeterministický, používa náhodné veličiny
- Typicky sa sleduje priemerná a maximálna fitness, GA sa zastaví, keď je dosiahnutá dopredu zadaná hodnota fitness, alebo zadaný počet generácií, alebo sa fitness v priebehu niekoľkých generácií nemení.
- Následne sa GA spúšťa opakovane aj s pozmenenými parametrami
- Výsledkom je najlepší jedinec vybraný z finálnych generácií zo všetkých behov GA.

Základné bloky a teória schém

- Fitness si môžeme predstaviť ako mnohorozmernú nadrovinu udávajúcu hodnotu fitness vo všetkých možných hodnotách genotypu – fitness krajina s kopcami a údoliami; pre chromozóm dĺžky 1 je to funkcia jednej premennej





Schémy

- Schéma je šablóna popisujúca nejakú skupinu reťazcov
  - $1*1 = \{101, 111\}$ * zastupuje ľubovoľnú prípustnú hodnotu
- populácia s N jedincami s reťazcami dĺžky k obsahuje niečo medzi 2^k a $N\,2^k$ schémami
- schéma môže popisovať komponentu chromozómu, ktorá zaručuje vysokú fitness
- potenciálne umožňuje preskúmať viacej reťazcov než ich je v populácii
- Holland 1975 – GA spracuje v jednom kroku až N^3 schém, aj keď má iba N reťazcov (implicitný paralelizmus GA)
- niektoré schémy majú vyššiu priemernú fitness;
- dlhé schémy sa ľahko rozbijú *1*****1* ***11***

Schémy (2)

- Najdôležitejšie sú schémy s krátkou dĺžkou. Schémy s krátkou dĺžkou a vysokou fitness sa objavujú v exponenciálne mnohých potomkoch v priebehu GA.
- Schémy sú považované za základné bloky evolúcie a kríženie je hlavný operátor, pretože umožňuje preskúmať kombinácie schém
- Funguje to však iba pri vhodnom kódovaní, kde základné bloky majú krátku dĺžku.
- POZOR: niekedy kríženie „kazí“ výsledky a pravdepodobnosť kríženia sa preto nastavuje na 0, alebo na veľmi malú hodnotu.

Veta o schémach

- $o(H)$ ozn. **rád schémy** H = počet pevných pozícií v schéme (s hodnotou 0 alebo 1 pre binárnu abecedu). Napr.
  - $o(011*1**) = 4$
  - $o(1******) = 1$
- $\delta(H)$ ozn. **dĺžku schémy** H = vzdialenosť medzi prvou a poslednou pevnou pozíciou (s hodnotou 0 alebo 1 pre binárnu abecedu)
  - $\delta(011*1**) = 4$
  - $\delta(1******) = 0$
- analyzujeme vývoj GA – vplyv reprodukcie, kríženia a mutácie na počet reťazcov zodpovedajúcich danej schéme

Vplyv reprodukcie na očakávaný počet reťazcov danej schémy v populácii (1)

- Predpokladajme, že v kroku ***t*** obsahuje populácia ***m*** reťazcov odpovedajúcich schéme ***H*** (***m = m (H , t)***)
- pri reprodukcii je reťazec A_i vybraný do ďalšej populácie (podľa jeho ohodnotenia $f_i = \Phi(A_i)$) s pravdepodobnosťou

$$p_i = \frac{f_i}{\sum_{j=1}^{n} f_j}$$

- v ďalšej populácii veľkosti a bude očakávaný počet reťazcovodpovedajúcich schéme ***H*** (***m = m (H , t+1)***) určený pomocou

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

- kde ***f (H)*** odpovedá priemernému hodnoteniu reťazcov odpovedajúcich schéme ***H***

Vplyv reprodukcie na očakávaný počet reťazcov danej schémy v populácii (2)

- ***f (H)*** odpovedá priemernému hodnoteniu reťazcov odpovedajúcich schéme ***H*** v kroku ***i***
  - potom pre
$$\bar{f} = \frac{\sum_j f_j}{n}$$
  - dostávame
$$m(H, t+1) = m(H, t) \cdot \frac{f(H)}{\bar{f}}$$

$\Rightarrow$ **pri prostej reprodukcii počet reťazcov odpovedajúcich určitej schéme v popuácii rastie, resp. klesá, podľa ich priemerného ohodnotenia**

Vplyv kríženia na očakávaný počet reťazcov danej schémy v populácii (1)

- Pre každú schému je možné určiť pravdepodobnosť jeho pretrvania po krížení p_S podľa $$p_S = 1 - \frac{\delta(H)}{l-1}$$
  - (teda schéma "prežije", ak bude bod kríženia "mimo jeho dĺžky")
- Ak kríženie nastáva náhodne – s pravdepodobnosťou p_c – bude pravdepodobnosť pretrvania schémy po krížení $$p_S = 1 - p_c \frac{\delta(H)}{l-1}$$

⇒ vplyv kombinácie reprodukcie a kríženia na počet reťazcov schémy H v populácii bude

$$m(H, t+1) \geq m(H, t) \cdot \frac{f(H)}{\bar{f}} \cdot \left[1 - p_c \frac{\delta(H)}{l-1}\right]$$

Vplyv kríženia na očakávaný počet reťazcov danej schémy v populácii (2)

$$m(H,t+1) \geq \quad m(H,t) \cdot \frac{f(H)}{\bar{f}} \cdot \left[1 - p_c \frac{\delta(H)}{l-1}\right]$$

- **Pri reprodukcii a krížení počet reťazcov schémy *H* rastie, resp. klesá, v závislosti na**
  - **ohodnotení schémy**
  - **dĺžke schémy**

Vplyv mutácie na očakávaný počet reťazcov danej schémy v populácii (1)

- Náhodná zmena na jednej pozícii s pravdepodobnosťou p_m
- aby "prežila" schéma H, tak sa musí zachovať každá z jeho pevných pozícií
  - každá pozícia prežíva s pravdepodobnosťou $1 - p_m$
  - všetky mutácie sú navzájom nezávislé
  - schéma H "prežije", ak prežije každá z jeho $o(H)$ pevných pozícií
  - pravdepodobnosť, že schéma H "prežije" mutáciu je

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

  - aproximácia pre malé hodnoty p_m $(<< 1)$

$$(1 - p_m)^{o(H)} \approx 1 - o(H)p_m$$

Vplyv reprodukcie, kríženia a mutácie na očakávaný počet reťazcov danej schémy v populácii

- očakávaný počet reťazcov po reprodukcii, krížení a mutácii

$$m(H,t+1) \geq \quad m(H,t) \cdot \frac{f(H)}{\bar{f}} \cdot \left[1 - p_c \frac{\delta(H)}{l-1} - o(H) \cdot p_m \right]$$

$\Rightarrow$ **najväčšiu "šancu na prežitie" majú krátke schémy s malým počtom pevných pozícií a nadpriemerným ohodnotením**