Genetické algoritmy a jejich aplikace
PB016 Úvod do umělé inteligence

Martin Frodl

1. 12. 2009

Úvod

Nemožnost efektivně vyřešit řadu (zejména $\mathcal{NP}$-těžkých) problémů klasickým systematickým prohledáváním vedla ke vzniku nejrůznějších stochastických (pravděpodobnostních) přístupů, využívajících znalosti konkrétního problému. Na rozdíl od deterministických algoritmů nelze dokázat jejich korektnost (a skutečně, s nenulovou pravděpodobností mohou vrátit jako výsledek suboptimální řešení); praxe však ukazuje, že ve velké většině případů *uspokojivé* řešení v požadovaném čase nalezeno je.

Genetické algoritmy, inspirované jednoduchou myšlenkou darwinovské selekce, se dostaly do širšího povědomí zásluhou Johna H. Hollanda a jeho studie procesů probíhajících v přirozených a umělých systémech.[3][4] Tak jako v přírodě soupeří jedinci v rámci populace o zdroje, pohlavní partnery atd., je i běh genetického algoritmu soutěží mezi jednotlivými řešeními. Na základě kvality je každé řešení více či méně upřednostňováno při reprodukci, tj. tvorbě nové generace řešení, tak, aby se vlastnosti vedoucí k lepšímu výsledku v populaci šířily.

Inspirace přírodními procesy se projevila i v terminologii, a tak se v oblasti genetických algoritmů objevují jinak biologické pojmy jako *chromosom*, *alela*, *mutace*, *crossing-over*, *fitness* a jiné, jejich význam bude objasněn v kontextu.

Obecné schéma genetického algoritmu

Navzdory rozmanitosti široce použitelné třídy genetických algoritmů lze více či méně u všech vysledovat následující společnou kostru:[4]

1. náhodně vygeneruj populaci řešení
2. pro každé řešení vypočítej *fitness*, tj. hodnotu nějaké ohodnocovací funkce
3. na základě fitness vyber dvojice ke zkřížení a vygeneruj jejich potomstvo
4. vytvoř novou populaci z generací rodičů a potomků
5. opět vypočítej fitness všech jedinců
6. pokud bylo nalezeno uspokojivé řešení (tj. s dostatečně vysokou fitness), vrať výslednou populaci, jinak opakuj od bodu 3

Namísto další teorie ukážeme průběh genetického algoritmu na následující motivační úloze, k jejímuž řešení se právě tohoto přístupu s úspěchem využívá.

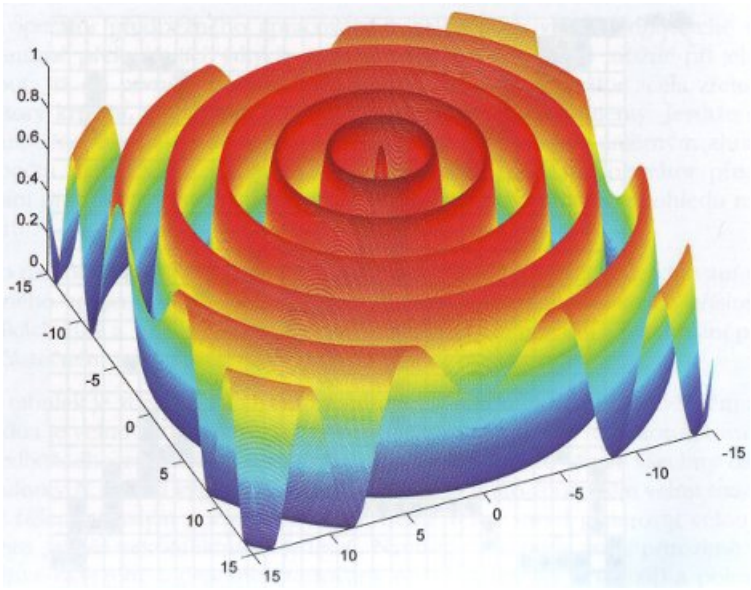
Příklad: hledální globálních extrémů funkce

Úkol: najděte globální maximum funkce

$$f(x,y) = \frac{1 - \sin^2 \sqrt{x^2 + y^2}}{1 + 0.001 \cdot (x^2 + y^2)}. \tag{1}$$

(Analyticky snadno ověříme, že jediné globální maximum se nachází v bodě $(0,0)$, kde $f(0,0) = 1$.)

Jak je z grafu patrné, je funkce f navržená tak, aby vzdorovala pokusům nejrůznějších horolezeckých algoritmů, "šplhajících" z náhodně zvoleného bodu po směru gradientu: takový algoritmus pravděpodobně nalezne pouze jedno z mnoha lokálních maxim. Genetický algoritmus, jak je popsán dále, naopak řešení podobných optimalizačních úloh s více než 90% pravděpodobností nalezne.[2]



Obrázek 1: Graf funkce $f(x,y)$

Omezme se při hledání maxima na množinu $[-100, 100] \times [-100, 100]$. Dále požadujme přesnost výsledku např. na 4 desetinná místa pro každou ze souřadnic x, y. Prohledávaná množina se tak rozpadne na $(200 \cdot 10000)^2 = 4 \cdot 10^{12}$ dílků, kde každý představuje možné řešení. Cílem algoritmu je nalézt takový dílek, ve kterém se nachází nejvyšší hodnota f.

Hledáme datovou strukturu umožňující jednoznačnou reprezentaci každého z dílků. Triviální (přesto hojně využívanou) reprezentací může být dostatečně dlouhá posloupnost bitů, v tomto případě 42prvková, protože

$$2^{41} \approx 2,20 \cdot 10^{12} < 4 \cdot 10^{12} < 4,40 \cdot 10^{12} \approx 2^{42}.$$

Prvních 21 bitů bude reprezentovat x-ovou, posledních 21 bitů y-ovou souřadnici daného řešení. Takovou bitovou posloupnost nazveme *chromosomem*, konkrétní pozice *geny* a hodnoty genů *alelami*. Podobně jako v přírodních populacích nese potomek část genů od jednoho a část od druhého rodiče, dochází i zde při generování potomstva k (pseudo)náhodnému rozdělení rodičovských chromosomů a spojením příslušných částí k sobě. Tuto operaci nazýváme *crossing-overem* (křížením) a jeho spojitost s přírodními procesy je patrná.

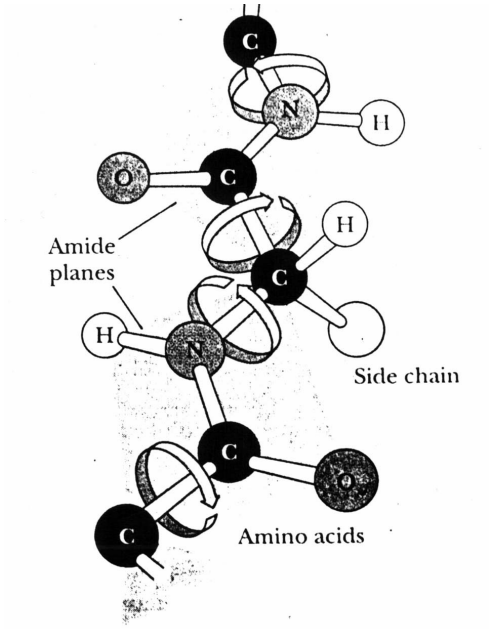
Rodiče							Potomci					
(**1**	**0**	**1**	**1**	**0**	**1**)	→	(**1**	**0**	**1**	**1**	0	0)
(1	1	0	0	0	0)		(1	1	0	0	**0**	**1**)

Kromě křížení může navíc s určitou pravděpodobností docházet k náhodným bodovým změnám, tzv. *mutacím*, které do populace vnášejí další variabilitu.

Jakmile máme sestavenou vhodnou reprezentaci řešení, samotný běh algoritmu již probíhá podle schématu uvedeného výše. Podrobnosti ohledně jednotlivých kroků jsou rozebrány např. v [4].

Aplikace: předpovídání struktury proteinů

Přepovědi struktury komplexních makromolekul, jakými jsou zejména proteiny, pouze na základě znalosti primární struktury (např. sekvence aminokyselin) stále zůstávají velkou výzvou bioinformatikům. Problém nalezení prostorového uspořádání s nejnižší možnou energií (právě takové se v živém



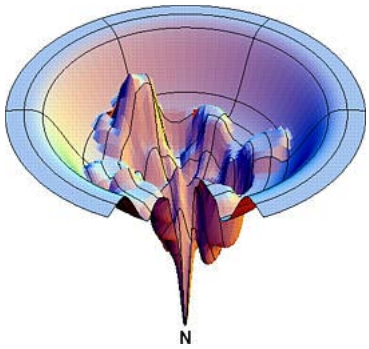
Obrázek 2: Torzní úhly φ (na vazbě C—N) a ψ (na vazbě C—C)

organismu pravděpodobně vyskytuje) zůstává i po redukci na problém nalezení optimální kombinace torzních úhlů φ, ψ (viz Obrázek 2) $\mathcal{NP}$-úplný.[1]

Genetické algoritmy se ukazují jako jedno z nadějných řešení tohoto problému. Modifikujeme-li strukturu chromosomu z úvodního příkladu tak, že místo binárních číslic se na každé pozici může vyskytovat číslo v rozsahu 0°–360°, redukuje se problém na velmi podobný úvodnímu příkladu. Ohodnocovací funkcí je celková energie dané konformace[6] (vypočtená na základě různých fyzikálních interakcí – náboje, hydrofobicity, fyzické kolize...), namísto dvou argumentů jich přijímá řádově $2n$ (kde n je počet aminokyselinových zbytků v proteinu).

Aplikace: předpovědi chování cen akcií (podle [5])

Další možná aplikace genetických algoritmů leží v oblasti financí. Úkolem je najít na základě souboru dosavadních údajů realistickou predikci dalšího vývoje na akciovém trhu. Otázkou, které údaje považovat za relevantní vzhledem k budoucímu stavu, se zabývá celé odvětví ekonomiky, zpravidla se vyu-



Obrázek 3: Ilustrace závislosti energie molekuly na hodnotách torzních úhlů φ, ψ (zjednodušeno pouze na dva úhly)

žívají informace tzv. *technické* (tj. dosavadní vývoj ceny daných akcií), *fundamentální* (tj., zhruba řečeno, současné ekonomické, politické, historické, demografické aj. faktory, které mají na cenu akcií potenciální vliv) nebo nejčastěji kombinace obou přístupů.[5][7]

Předpokládejme, že chceme předpovědět (relativní) výnos akcií konkrétní firmy 12 týdnů do budoucnosti. Sestavme kolekci několika (řekněme 15) údajů reprezentujících např. hodnoty nějakého ekonomického ukazatele v různých časových okamžicích a ke každému z nich dodejme relativní výnos akcií po uplynutí 12 týdnů od příslušného data. Kolik vzorkování provést, je otázkou praxe a závisí pochopitelně na množství dostupných údajů; [5] uvádí 200–600 záznamů pro každý ukazatel.

Úloha genetického algoritmu nyní spočívá v nalezení vhodného pravidla (pravidel), rozhodujících o příznivosti nebo nepříznivosti okolností pro daný ukazatel. Typicky nás může zajímat, za jakých podmínek je výhodné které akcie prodávat nebo naopak nakupovat. Z náhodné populace pravidel tvaru např.

Pravidlo 1: **Pokud** P/E[1]> 30, **pak** prodej.
Pravidlo 2: **Pokud** $P/E < 40$ **a** tempo růstu $> 40\%$, **pak** nakup.
atd.

postupně genetickým algoritmem selektujeme ta pravidla, jejichž apli-

[1]P/E = poměr cena/výnos, price to earnings ratio

kací maximalizujeme čistý výnos. Jak uvádí [5], kombinací několika pravidel můžeme často dosáhnout lepších výsledků než použitím kteréhokoli z nich zvlášť, jak může být patrné na příkladu pravidel 1 a 2 (výše): navzdory nevýhodnému poměru P/E můžeme být výhodné učinit výjimku v důsledku rychle rostoucí ceny.

Další oblasti použití

- design materiálů a aerodynamických tvarů závodních automobilů i běžných dopravních prostředků
- strukturní a funkcionální design staveb, strojů, výrobních jednotek atd.
- směrování paketů v telekomunikačních sítích (routing)
- šifrování a dešifrování
- návrh syntetických molekul (léčiva, materiály, atd.)

Literatura

[1] Crescenzi Pierluigi a kol.: "On the Complexity of Protein Folding." In: *Proceedings of the 30th Symposium on Theory of Computing*, ACM Press, New York, 1998, pp. 61–62.

[2] Geyer-Schulz Andreas: *Fuzzy Rule-Based Expert System and Genetic Machine Learning* (2nd ed.), Heidelberg, Physica-Verlag, 1997.

[3] Holland John Henry: *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, Michigan, 1975.

[4] Hynek Josef. *Genetické algoritmy a genetické programování*. Grada Publishing, Praha, 2008.

[5] Mahfoud Sam, Mani Ganesh: "Financial Forecasting Using Genetic Algorithms." In: *Applied Artificial Intelligence*, Taylor & Francis, 1996, pp. 543–565. Dostupné online: <`http://sce.uhcl.edu/boetticher/ML_DataMining/mahfoud96financial.pdf`>

[6] Schulze-Kremmer Steffen: "Protein Folding Simulation by Force Field Optimisation." *Genetic algorithms and protein folding*. 25. 1. 1996. Web. 25. 11. 2009. Dostupné online: <`http://www.techfak.uni-bielefeld.de/bcd/Curric/ProtEn/121.html`>

[7] "Fundamentální analýza." *Wikipedie: Otevřená encyklopedie*. 26. 11. 2009, 15:08 UTC. 26. 11. 2009, 19:25. Dostupné online: <`http://cs.wikipedia.org/w/index.php?title=Fundamentální_analýza&oldid=4648524`>