Gry wieloosobowe, gry z niepełną informacją, gry jednoturowe

Paweł Rychlikowski

Instytut Informatyki UWr

17 kwietnia 2024

Monte Carlo Tree Search

Algorytm odpowiedzialny za przełom w:

a. W grze w Go
b. W General Game Playing

Główne idee. Przypomnienie

1. Oceniamy sytuację wykonując symulowane rozgrywki.
2. Budujemy drzewo gry (na początku składające się z jednego węzła – stanu przed ruchem komputera)
3. Dla każdego węzła utrzymujemy statystyki, mówiące o tym, kto częściej wygrywał gry rozpoczynające się w tym węźle (pamiętamy też listę możliwych ruchów, ewentualnie wraz ze stanami)
4. Selekcję wykonujemy na każdym poziomie (UCB), na końcu rozwijamy wybrany węzeł dodając jego dzieci i przeprowadzając rozgrywkę

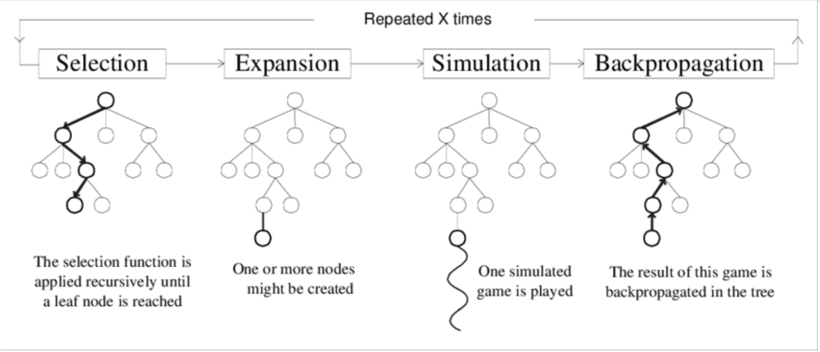
MCTS. Podstawowe operacje

1. **Selection**: wybór węzła do którego wejdziemy (być może wcześniej go tworząc) rozwinięcia
2. **Expansion**: rozwinięcie drzewa (dodanie węzła)
3. **Simulation**: symulowana rozgrywka (zgodnie z jakąś polityką), zaczynające się od wybranego węzła
4. **Backup**: uaktualnienie statystyk dla rozwiniętego węzła i jego przodków

Uwaga

Korzystność ruchu oceniamy z punktu widzenia aktualnego gracza.

MCTS. Rysunek



Opcje

- Jeżeli wejdziemy do węzła, w którym nie wszystkie dzieci są rozwinięte, to możemy wymusić, że brany jest pod uwagę nieuwzględniony ruch.
- Możemy korzystać z jakiejś heurystyki oceniającej ruchy i brać tylko obiecujące ruchy

MCTS. Dodatkowe uwagi

- Możemy grać nie do końca, wówczas jednak potrzebujemy jakiejś funkcji heurystycznej oceniającej sytuację na planszy
- Rozgrywka nie musi być prostym losowaniem, p-stwo ruchu może zależeć od jego (szybkiej!) oceny (to powinno być coś bardzo szybkiego, niekoniecznie funkcja z podpunktu wyżej)

Uwaga

Im więcej obrotów pętli, tym lepsza gra – precyzyjne sterowanie trudnością i czasem działania.

Wybór ruchu

- Naturalny wybór: ruch do najlepiej ocenianej sytuacji
- Inna opcja: ruch do sytuacji, w której byliśmy najwięcej razy

Komentarz do wyboru ruchu

- W pewnym sensie opcje są podobne: UCB też raczej wybiera dobre ruchy (eksploatacja!)
- Wybierając częstą sytuację, uwzględniamy wiarygodność szacunków
- Pojedyncza bardzo korzystna partia zmienia stosunkowo niewiele

Jeszcze o rozgrywce i wyborze węzła w MCTS

- Ciekawa idea: **all-moves-as-first**: w danej sytuacji na planszy szacujemy jakość ruchów widzianych (w symulacjach, w $\alpha\beta$-search też by się dało to zastosować) niezależnie od tego, w którym momencie się zdarzyły
- Motywacja: w tej sytuacji **zawsze** jak ruszę hetmanem na B5 to wygrywam
- Możemy liczyć wartość ruchu jako średni wynik rozgrywki, w której ten ruch był wykonany.
- **Uwaga**: nie $Q(s, a)$, ale $Q(a)$! (ta wartość nie zależy od konkretnego momentu, w którym ruch został wykonany)

Więcej szczegółów w pracy S.Gelly, D.Silver, *Monte-Carlo Tree Search and Rapid Action Value Estimation in Computer Go*

Stosowalność MCTS

- Nie tylko do gier!
- Można stosować do *poważnych* zadań, związanych z przeszukiwaniem (bez oponenta)
  - Na przykład do rozwiązywania więzów (pewnie szczegóły na ćwiczeniach)

Gry częściowo obserwowalne

- Ciekawe do analizy są gry, w których agenci nie mają pełnej wiedzy o świecie.
- Klasyczne przykłady to gry karciane, ale nie tylko.

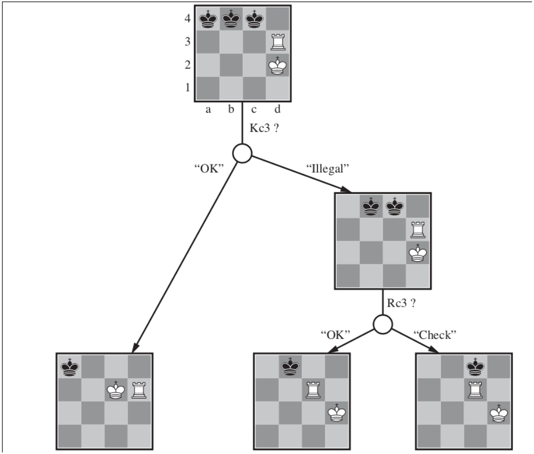
Kriegspiel

- Mamy dwóch graczy, arbitra i 3 szachownice.
- Gracze widzą na szachownicy swoje pionki, mogą tworzyć też hipotezy o bierkach przeciwnika.
- Arbiter zna położenie wszystkich figur i udziela graczom pewnych (skąpych) informacji.
  a) przede wszystkim ocenia, czy ruch jest możliwy (komunikacja osobista, dobry ruch jest od razu wykonywany, w przypadku złego, gracz proponuje kolejny, aż do skutku)
  b) odpowiada na pytanie: „czy ja (gracz) mam jakieś bicie?"
  c) informuje obu graczy, że „na polu X zbito bierkę" (nie podając jaka bierka jest zbita, a jaka biła)
  d) Mówi o szachu (do ubu graczy), dodając, że zagrożenie jest w wierszu, kolumnie, przekątnej lub przez skoczka
- Tak poza tym, to całkiem normalne szachy.

Podobno ludzie radzą sobie z tą grą całkiem nieźle...

Końcówka w Kriegspiel

Przykładowa końcówka, gracz biały dowiedział się, że czarnemu został tylko król i jest on na jednym z 3 pól.



Uwaga

W stanie gry powinniśmy umieścić możliwe ustawienia bierek przeciwnika

Trochę jak z komandosem...

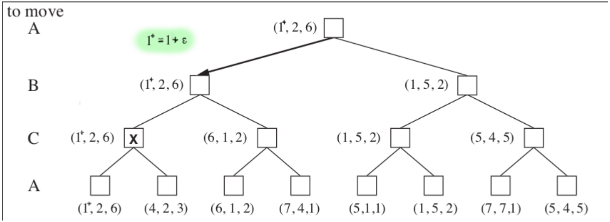
Pytanie

A jak grać w brydża, bądź inną grę karcianą?

Idea (do rozwinięcia na ćwiczeniach)

losowanie układu kart i gra w otwarte karty dla wylosowanego układu, czynności powtarzamy wiele razy

Gry z większą liczbą uczestników



Gry wieloosobowe. Problemy

- Strategia maksymalizująca korzyść pojedynczego gracza w oczywisty sposób nieoptymalna (A mógłby się dogadać z B).
- Kwestie sojuszów, zrywania sojuszów, budowania wiarygodności.
- Czasem używa się: paranoidalnego założenia – gra wieloosobowa staje się jednoosobową, w której oni wszyscy chcą mi zaszkodzić.

Zagadka: co robią robotnik i kołchoźnica?

Odpowiedź

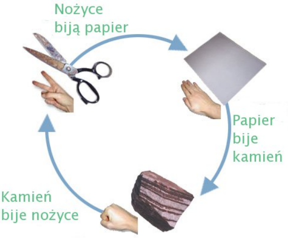

YOU CAN WIN
WIN
ROCK PAPER PAT ER SICGO RSS SRIOORS
IF YO YO WANT!

Gry z jedną turą

- Powiemy sobie trochę o grach z jedną turą
- Ale takich, w których gracze podejmują swoje decyzje jednocześnie

Rozważamy gry z sumą zerową.

Papier, nożyce, kamień



Źródło: Wikipedia

Macierz wypłat

Grę definiuje **macierz wypłat**. Przykładowo poniżej dla P-N-K

Max/Min	Papier	Nożyce	Kamień
Papier	0	-1	+1
Nożyce	+1	0	-1
Kamień	-1	+1	0

Strategie

- Czysta strategia: zawsze akcja a
- Mieszana strategia: rozkład prawdopobieństwa na akcjach

PNK – spostrzeżenia

- **Oczywisty fakt:** każdą strategię stałą można pokonać (też stałą strategią)
- **Fakt 1:** każdą strategię mieszaną można (prawie) pokonać za pomocą strategii stałej:
  Mój przeciwnik gra losowo, ale z przewagą kamienia – zatem ja daję **zawsze** papier
- **Fakt 2:** Optymalna strategia jest mieszana (w tej grze każde z $p = \frac{1}{3}$)
- **Fakt 3:** Znajomość optymalnej strategii mieszanej gracza A, nie daje żadnej przewagi graczowi B (i odwrotnie)

PNK – uwagi końcowe

- W prawdziwym P-N-K dochodzi kilka innych aspektów:
  - Grają ludzie, którzy nie potrafią realizować losowości, Który człowiek (nie dysponując kostką do gry), przegrawszy 3 razy z rzędu jako papier pokaże papier?
  - za to wysyłają swoimi ciałami różne informacje, które można analizować
- Zatem ma sens organizowanie zawodów w PNK
- Sens miałyby również zawody ludzko-komputerowe, realizowane on-line (agent musiałby zgadnąć, czy gra z człowiekiem, czy z maszyną i czy opłaca się próbować zgadnąć model losowania używany przez człowieka)

Gra w zgadywanie (Morra 2)

- Mamy dwóch graczy:
  A) Zgadywacz
  B) Zmyłek

  którzy na sygnał pokazują 1 lub 2 palce.
- Jeżeli Zgadywacz nie zgadnie (pokazał coś innego niż Zmyłek), daje Zmyłkowi 3 dolary.
- Jeżeli Zgadywacz zgadnie, to dostaje od Zmyłka:
  - jak pokazali 1 palec, to 2 dolary
  - jak pokazali 2 palce, to 4 dolary

Pytanie

Jak grać w tę grę? (prośba o podanie wstępnych intuicji)

Macierz wypłat

Definicja

Taką grę zadajemy za pomocą macierzy wypłat, w której $V_{a,b}$ jest wynikiem gry z punktu widzenia pierwszego gracza.

Nasza gra:

Zg/Zm	1 palec	2 palce
1 palec	2	-3
2 palce	-3	4

Proste fakty

- Jak Zmyłek będzie grał cały czas to samo, to Zgadywacz wygra każdą turę (i odwrotnie)
- Muszą zatem stosować strategie mieszane, ale jakie?

Wartość gry

Definicja

Wartość gry dla dwóch strategii graczy jest równa:

$$V(\pi_A, \pi_B) = \sum_{a,b} \pi_A(a)\pi_B(b)V(a, b)$$

Przykładowo: Zgadywacz zawsze zgaduje 1, Zmyłek wybiera akcję losowo z prawdopodobieństwem 0.5.

Wynik: $-\frac{1}{2}$ (tak samo często zyskuje 2 jak traci 3 dolary)

Strategia mieszana vs czysta

Uwaga

Jeżeli gracz A zapowie, że będzie grał strategią mieszaną (i ją poda), wówczas gracz B może grać strategią czystą (i osiągnie optymalny wynik).

Dlaczego?

Odpowiedź

- Możemy dla każdej akcji policzyć wartość oczekiwaną wypłaty
- i wybrać (dowolną) najlepszą akcję
- (Jeżeli takich akcji jest więcej, wówczas można też dowolnie losować między nimi)

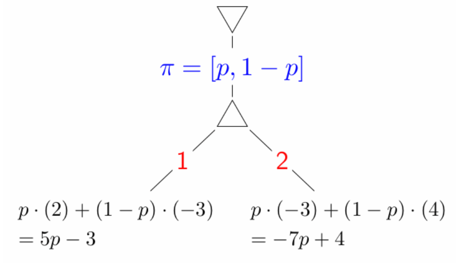
Gra w zgadywanie (Morra 2). Przypomnienie

- Mamy dwóch graczy:
  - A) Zgadywacz
  - B) Zmyłek

  którzy na sygnał pokazują 1 lub 2 palce.
- Jeżeli Zgadywacz nie zgadnie (pokazał coś innego niż Zmyłek), daje Zmyłkowi 3 dolary.
- Jeżeli Zgadywacz zgadnie, to dostaje od Zmyłka:
  - jak pokazali 1 palec, to 2 dolary
  - jak pokazali 2 palce, to 4 dolary

Znalezienie optymalnej strategii

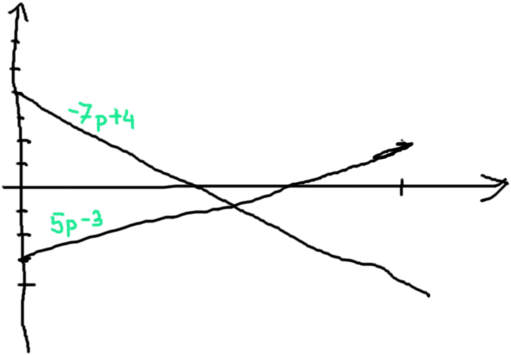
Zaczyna gracz B – Zmyłek.
Wybiera strategię mieszaną z parametrem p



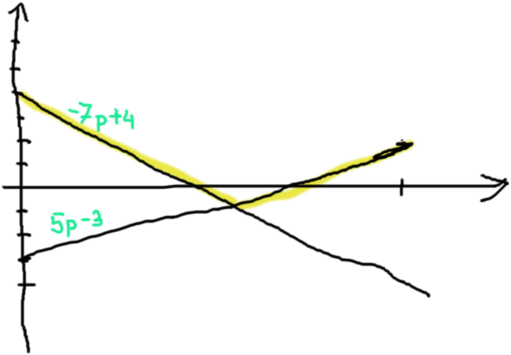
Wartość takiej gry to

$$\min_{p \in [0,1]} (\max(5p - 3, -7p + 4))$$

Optymalna strategia. Wykresy



Optymalna strategia. Wykresy



Znalezienie optymalnej strategii (2)

- W powyższej grze, Zmyłek osiągnie najlepszy wynik, gdy przyjmie $p = \frac{7}{12}$, wynik ten to $-\frac{1}{12}$
- Ok, on zaczynał, miał trudniej – a gdyby zaczynał Zgadywacz? I podał swoją strategię mieszaną?

Wynik gry

Wynik jest dokładnie taki sam, czyli $-\frac{1}{12}$!

Twierdzenie von Neumana

Twierdzenie, von Neuman, 1928

Dla każdej jednoczesnej gry dwuosobowej o sumie zerowej ze skończoną liczbą akcji mamy:

$$\max_{\pi_A} \min_{\pi_B} V(\pi_A, \pi_B) = \min_{\pi_B} \max_{\pi_A} V(\pi_A, \pi_B)$$

dla dowolnych mieszanych polityk π_A, π_B.

- Można ujawnić swoją politykę optymalną!
- **Dowód**: pomijamy, programowanie liniowe, przedmiot J.B.
- Algorytm: programowanie liniowe

- Można o grze wieloturowej myśleć jako o grze jednoturowej
- Gracze na sygnał kładą przed sobą opis strategii (program)

Uwaga

Optymalną strategią jest MiniMax (ExpectMiniMax w grach losowych). Ale wiedząc o strategii gracza różnej od optymalnej możemy oczywiście ugrać więcej.

Co pomijamy

- Gry o sumie niezerowej, w których dochodzi możliwość kooperacji.
- Punkt równawagi Nasha (jest zawsze para strategii, że żaden gracz nie chce jej zmienić, wiedząc, że ten drugi nie zmienia). **Również dla gier o sumie niezerowej!**
- Agent musi zdecydować, czy ma być miły dla innego agenta (i budować reputację przy wielu rozgrywkach, słynny dylemat więźnia).

Procesy decyzyjne Markowa

Procesy decyzyjne Markowa (MDP)

- Coś pomiędzy grami a zwykłym zadaniem przeszukiwania
- (zwłaszcza jeżeli przypomnimy sobie gry z węzłami losowymi)
- a jednocześnie krok w stronę uczenia ze wzmocnieniem

MDP a przeszukiwanie

Standardowe przeszukiwanie

Znamy mechanikę świata i wiemy, że akcja w stanie da nam konretny rezultat (inny stan).

MDP

Znamy mechanikę świata i wiemy, że akcja w stanie da nam pewien rozkład prawdopodobieństwa na następnych stanach.

Nie wiemy, co dokładnie się stanie, ale wiemy co **może** się stać i z jakim prawdopodobieństwem.

Własność Markowa

- Przyszłość zależy od ostatniego stanu.
- Nie zależy od historii...
- Chyba, że jej fragment (o długości N) uznamy za część stanu.

Ważna uwaga

Zakładamy **skończoną** liczbę stanów

Uwaga na wulkany (1)

- Dobrze omawia się MDP na prostych światach na prostokątnej kratce.
- I od takich modeli zaczniemy.

Generalnie myślimy na początku o przestrzeni stanów na tyle małej, że nie będzie kłopotów z pamiętaniem różnych wartości dla **każdego stanu**.

Uwaga na wulkany (2)

Volcano crossing

		-50	20
		-50	
2			

CS221 / Autumn 2017 / Liang & Ermon

Mechanika świata wulkanów

		-50	20
		-50	
2			

- Możliwe 4 akcje (UDLR)
- W normalnym przypadku efekt oczywisty (próba wyjścia poza planszę oznacza pozostanie na polu)
- Z prawdopodobieństwem p możemy się poślizgnąć, wówczas poruszamy się w losowym kierunku.
- Dojście do pola z liczbą kończy grę (i odpowiednią dostajemy wypłatę).

Inny przykład. Gra w kości

Uwaga

Nagroda może być przydzielana w sposób ciągły, nie tylko w stanie końcowym.

- Mamy dwie opcje: pozostanie albo rezygnacja.
- rezygnacja oznacza wypłatę **10$**
- pozostanie to wypłata **4$** po której rzucamy kostką.
- Interpretacja wyniku:
  - 1,2 – koniec gry
  - 3,4,5,6 – gramy dalej

Pytanie

Ile mamy stanów? Odpowiedź: 2