Algorytmy robotyki mobilnej
Eksploracja

Janusz Jakubiak

Katedra Cybernetyki i Robotyki

2020/2021

Informacja o prawach autorskich

Niniejsza prezentacja zawiera materiały z książki Probabilistic Robotics (S. Thurn et al.)

Planowanie i sterowanie probabilistyczne

Motywacja

Zakłócenia informacji wykorzystywanej w podejmowaniu decyzji powodują niepewność efektów wykonywanej akcji.
Jeśli moglibyśmy estymować niepewność informacji – czy moglibyśmy poprawić efekty akcji?

Planowanie i sterowanie probabilistyczne

Rozważane typy niepewności

efekty akcji :
- deterministyczne
- stochastyczne

percepcja :
- pełna obserwowalność
- częściowa obserwowalność

Cel

Zapewnić odporność na zakłócenia nie tylko bieżącej akcji, ale także uwzględniać przewidywaną przyszłą niepewność.

Eksploracja

Eksploracja (zadanie zbierania informacji)

Bezpośrednim celem akcji robota jest uzyskanie informacji (zmniejszenie niepewności).

Przykłady

- budowanie siatki zajętości – maksymalizacja informacji o każdym polu
- przeszukiwanie – wyznaczenie położenia obiektu (np. osoby w budynku)
- aktywna lokalizacja – poprawa jakości informacji o własnym położeniu

Metody

- Częściowo obserwowalne procesy decyzyjne Markowa (Partially Observable Markov Decision Process, POMDP) – ogólne, lecz złożone obliczeniowo
- metody specjalizowane

Terminologia

cel – osiągnięcie określonego rezultatu (np. stanu), zwykle z optymalizacją funkcji kosztu

koszt – zmienna opisująca jakość ruchu (dokładność, czas, długość ścieżki itp.)

nagroda elementarna – funkcja opisująca zależność kosztu w pojedynczym kroku od stanu i sterowania

$$r(x, u)$$

współczynnik dyskontujący ($\gamma \in [0, 1]$) współczynnik zależny od czasu

horyzont planowania T – najczęściej: 1, skończony, nieskończony

Terminologia, c.d.

całkowita nagroda –

$$\sum_{\tau=1}^{T} \gamma^T r_{t+\tau}$$

strategia – plan akcji

$$\pi : z_{1:t-1}, u_{1:t-1} \rightarrow u_t (\text{or} u_{t:t+r})$$

nagroda oczekiwana

$$R_T = E\left[\sum_{\tau=1}^{T} \gamma^T r_{t+\tau}\right]$$

strategia optymalna

$$\pi^\star = \text{argmax}_\pi \, R_T$$

Proces decyzyjny Markowa (MDP)

- dla modelu stochastycznego ze stanem w pełni obserwowalnym $\pi : x \rightarrow u$
- optymalizacja zachłanna $T = 1$

$$\pi_1 = \text{argmax}_u \, r(x, u)$$

z całkowitą przyszłą nagrodą

$$V_1(x) = \gamma \max_u r(x, u)$$

- czas skończony T

$$\pi_T = \text{argmax}_u \left[r(x, u) + \int V_{T-1}(x') p(x'|u, x) dx' \right]$$

z całkowitą przyszłą nagrodą

$$V_T(x) = \gamma \max_u \left[r(x, u) + \int V_{T-1}(x') p(x'|u, x) dx' \right]$$

Algorytm MDP

korzystając z równania Bellmana

iteracja wartości MDP

for all x do
$\hat{V}(x) = r_{min}$
endfor
repeat until convergence
for all x

$$\hat{V}(x) = \max_u \left[r(x, u) + \gamma \int \hat{V}(x') p(x'|u, x) dx' \right]$$

endfor
endrepeat
return $\hat{V}$

W przypadku dyskretnym: pętla po wszystkich stanach

oczekiwana suma nagród to aktualna nagroda i zdyskontowana suma nagród po wybraniu najlepszej akcji

POMDP

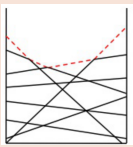
- stan nie jest obserwowalny, więc jest zastąpiony przez b (*belief*)
- funcja wartości

$$V_T(x) = \gamma \max_u \left[r(b, u) + \int V_{T-1}(b') p(b'|u, b) db' \right]$$

- strategia optymalna

$$\pi_T = \text{argmax}_u \left[r(b, u) + \int V_{T-1}(b') p(b'|u, b) db' \right]$$

- kluczem do efektywnej implementacji jest eliminacja nieużytecznych stanów



Przybliżone POMDP

Złożoność standardowych POMDP sprawia, że nie nadają się do praktycznego zastosowania w robotyce

Metody przybliżone

- QMPD – algorytm pomiędzy MDP i POMDP; traktuje stan po pierwszej iteracji jako znany
- Augmented MDP (ADP) – przekonanie jest reprezentowane przez parametry rozkładu (np. stan i entropię)
- Monte Carlo MDP (MC-MDP) – analogiczny do filtru cząsteczkowego

Eksploracja

Przyrost informacji

- Oczekiwana informacja $E[-\log p]$
- Entropia dystrybucji prawdopodobieństwa $p(x)$

$$H_p(x) = -\int p(x)\log p(x)dx \quad \text{lub} -\sum_x p(x)\log p(x)$$

- Przekonanie $B(b,z,u)$
- Entropia warunkowa

$$H_b(x'|z,u) = -\int B(b,z,u)(x')\log B(b,z,u)(x')dx'$$

$$\text{lub: } -\sum_x p(x)\log p(x)$$

Algorytm eksploracji Monte Carlo

set $\rho_u = 0$ for all u
for i=1 to N do
 sample $x \sim b(x)$
 for all u do
  sample $x' \sim p(x'|x,u)$
  sample $z \sim p(z|x')$
  $b' = BayesFilter(b,z,u)$
  $\rho_u = \rho_u + r(x,u) - \alpha H_{b'}(x')$
 endfor
endfor
return $argmax_u\ \rho_u$

Pytania kontrolne

1. Na czym polega zadanie eksploracji?
2. Jak formalnie definiuje się eksplorację?
3. Jakie metody wykorzystywane są w eksploracji? Jaka jest idea ich działania?