WSI Zadanie 6 Q-learning

Piotr Lenczewski

Styczeń 2024

1 Opis badanego algorytmu

1.1 Cele algorytmu

Celem algorytmu Q-learning jest nauczenie agenta podejmowania optymalnych decyzji w środowisku, aby osiągnąć maksymalną nagrodę. Agent jest w nim podmiotem, który poprzez podejmowanie decyzji jest karany lub nagradzany, ucząc się w ten sposób optymalnego zachowania. Jest to tzw. uczenie ze wzmocnieniem.

1.2 Strategie eksploracji

1.2.1 Strategia ϵ-zachłanna

Strategia ϵ-zachłanna jest techniką eksploracji w której z prawdopodobieństwem ϵ losujemy akcję. W przeciwnym wyadku wybieramy najkorzystniejszą akcję według aktualnych wartości tablicy Q.

$$\pi(x,a) = \begin{cases} \text{losowa akcja} & \text{dla jeżeli rand()} < \epsilon \\ \arg\max_a Q(x,a) & \text{dla jeżeli rand()} \geq \epsilon \end{cases}$$

gdzie: ϵ - parametr strategii eksploracji, x - stan środowiska, a - akcja.

1.2.2 Strategia oparta na rozkładzie Boltzmanna

Definiujemy strategię opartą na rozkładzie Boltzmanna, gdzie prawdopodobieństwo wyboru akcji a w stanie x jest proporcjonalne do eksponentu z wartości $Q(x,a)$ podzielonej przez temperaturę T:

$$\pi(x,a) = \frac{\exp\left(\frac{Q(x,a)}{T}\right)}{\sum_b \exp\left(\frac{Q(x,b)}{T}\right)}$$

gdzie: $\pi(x,a)$ - prawdopodobieństwo wyboru akcji a w stanie x, $Q(x,a)$ - wartość funkcji Q dla akcji a w stanie x, T - temperatura, kontrolująca wpływ wartości Q na prawdopodobieństwo.

1.3 Pseudokod

- Q - tablica o wielkości (ilość możliwych stanów)x(ilość możliwych decyzji). Wartości tablicy inicjowane zerami symbolizują ocenę danego wyboru i są aktualizowane wraz z działaniem algorytmu,
- e - ilość episodów działania algorytmu,
- x - stan agenta (np. miejsce na planszy),
- g - współczynniki dyskątujący,
- lr - współczynnik uczenia,
- a - akcja,
- r - nagroda,
- stan absorbujący, jest stanem kończącym symulację.

```
begin
    Q <- 0; e <- 0
    while e < emax do
        x_i <- inicjuj stan poczatkowy
        while x_i not in stany absorbujace do
            a_i <- wybierz akcje(x_i, Q_i)
            r_i, x_i+1 <- wykonaj akcje a_i
            cel <- r_i + g * max_a(Q_i(x_i+1, a))
            Q_i+1 <- Q_i + lr * (cel - Q_i(x_i, a_i))
        end
        e <- e + 1
    end
end
```

2 Planowane eksperymenty numeryczne

Mam zamiar zbadać wpływ współczynnika uczenia oraz rodzaju algorytmu eksploracji na działanie algorytmu Q-learning. Wykorzystam do tego środowisko `https://www.gymlibrary.dev/environments/toy_text/taxi/`.

2.1 Założenia początkowe

- Testowane bądą średnia liczba kroków na episod oraz średnia nagroda na episod,
- Trenowanie przeprowadzę na 10000 episodów, a testowanie na 100 episodach,
- Strategie eksploracji będą testowane dla różnych wartości parametrów: ϵ, T, lr.,
- Podobnie współczynnik dyskątujący: g=0.9.

3 Uzyskane wyniki

3.1 Strategia ϵ-zachłanna

	$\epsilon = 0.01$		$\epsilon = 0.1$		$\epsilon = 0.5$	
	avg steps	avg reward	avg steps	avg reward	avg steps	avg reward
lr = 0.1	13.89	6.57	14.31	2.64	28.98	-50.01
lr = 0.5	13.17	7.56	14.17	3.59	29.4	-49.44
lr = 1	13.63	7.01	14.88	2.7	28.92	-47.43

Table 1: Wyniki działania algorytmu dla różnych ϵ i lr

3.2 Strategia oparta na rozkładzie Boltzmanna

	$T = 0.1$		$T = 0.5$		$T = 1$	
	avg steps	avg reward	avg steps	avg reward	avg steps	avg reward
lr = 0.1	13.09	7.91	13.33	7.67	13.69	7.31
lr = 0.5	12.97	8.03	13.47	7.53	18.59	2.41
lr = 1	13.42	7.58	13.49	7.51	18.03	2.97

Table 2: Wyniki działania algorytmu dla różnych T i lr

4 Wnioski

- Optymalna wartość współczynnika uczenia, niezależnie od strategi eksploracji, jest bliska lr=0.5
- Optymalna wartość ϵ jest bliska $\epsilon = 0.01$
- Optymalna wartość T jest bliska $T = 0.1$
- Dla badanych wartości parametrów strategia oparta na rozkładzie Boltzmanna jest wydajniejsza niż strategia ϵ-zachłanna