WSI, Ćwiczenie 6

Michał Kwarciński

Polecenie:

Proszę zaimplementować algorytm Q-Learning i użyć go do wyznaczenia polityki decyzyjnej dla problemu FrozenLake8x8 (w wersji domyślnej, czyli z włączonym poślizgiem). W problemie chodzi o to, aby agent przedostał się przez zamarznięte jezioro z pozycji 'S' do pozycji 'G' unikając punktów 'H'. Symulator dla tego problemu można pobrać z podanej strony lub napisać własny o takiej samej funkcjonalności.

Oprócz zbadania domyślnego sposobu nagradzania (1 za dojście do celu, 0 w przeciwnym przypadku) proszę zaproponować własny system nagród i kar, po czym porównać osiągane wyniki z wynikami systemu domyślnego.

Za wynik (podczas testowania) uznajemy procent dojść do celu w 1000 prób (10x więcej prób używamy w treningu). W każdej próbie można wykonać maksymalnie 200 akcji.

Badania wpływu parametrów gamma i beta dla funkcji bazowej:

Gamma	Beta	Średni wynik	Najlepsza wartość	Najgorsza wartość	Odchylenie standardowe
0,1	0,1	0,01	**0,04**	0,0	0,01
	0,25	0,01	0,02	0,0	0,01
	0,5	0,01	0,03	0,0	0,01
	0,75	0,01	0,04	0,0	0,01
	0,9	0,02	0,03	0,0	0,01
0,25	0,1	0,01	0,03	0,0	0,01
	0,25	0,01	0,04	0,0	0,01
	0,5	0,01	0,02	0,0	0,01
	0,75	0,01	0,03	0,0	0,01
	0,9	0,0	0,01	0,0	0,0
0,5	0,1	0,01	0,02	0,0	0,01
	0,25	0,01	0,03	0,0	0,01
	0,5	0,0	0,01	0,0	0,0
	0,75	0,0	0,01	0,0	0,0
	0,9	0,01	**0,04**	0,0	0,01
0,75	0,1	0,01	**0,06**	0,0	0,02
	0,25	0,01	0,02	0,0	0,0
	0,5	0,01	0,03	0,0	0,01
	0,75	0,01	0,02	0,0	0,0
	0,9	0,01	0,02	0,0	0,0
0,9	0,1	0,01	0,02	0,0	0,01
	0,25	0,01	0,03	0,0	0,01
	0,5	0,01	0,03	0,0	0,01
	0,75	0,01	**0,04**	0,0	0,01
	0,9	0,01	**0,04**	0,0	0,01

Na podstawie tych wyników nie da się stwierdzić dokładnie, który z nich był najlepszy.

Bazowa funkcja nie doprowadza algorytmu do celu. Można zauważyć większą częstość pojawiania się najlepszych pojedynczych wyników przy wyższych wartościach parametru gamma.

Badania wpływu parametrów epsilon dla funkcji bazowej:

Przeprowadziłem również badanie parametru epsilon – odpowiedzialnego za zwiększenie zdolności algorytmu do eksploracji.

Epsilon	Gamma	Beta	Średni wynik	Najlepsza wartość	Najgorsza wartość	Odchylenie standardowe
0	0,1	0,1	0,01	0,01	0	0
2,5			0,02	0,04	0	0,01
5			0,02	0,04	0,01	0,01
7,5			0,02	0,05	0,01	0,01
10			0,04	0,1	0	0,03
20			0,03	0,05	0,01	0,01
30			0,03	0,1	0,01	0,02
40			0,03	0,07	0,01	0,02
50			0,02	0,05	0	0,02
0	0,1	0,9	0,01	0,02	0	0,01
2,5			0,02	0,06	0	0,02
5			0,03	0,09	0	0,03
7,5			0,02	0,05	0	0,01
10			0,03	0,13	0	0,04
20			0,02	0,1	0	0,03
30			0,05	0,27	0	0,08
40			0,02	0,07	0	0,02
50			0,03	0,1	0,01	0,03
0	0,5	0,5	0,02	0,05	0	0,02
2,5			0,06	0,15	0	0,05
5			0,06	0,14	0,03	0,03
7,5			0,06	0,2	0,01	0,07
10			0,06	0,17	0,02	0,04
20			0,03	0,07	0	0,02
30			0,03	0,07	0	0,02
40			0,03	0,07	0,01	0,02
50			0,08	0,25	0,01	0,07
0	0,9	0,1	0,01	0,01	0	0,01
2,5			**0,6**	**0,74**	**0,24**	**0,15**
5			**0,51**	**0,74**	0,32	0,13
7,5			0,46	**0,66**	0,21	0,15
10			**0,5**	**0,63**	0,23	0,13
20			0,48	**0,76**	0,21	0,16
30			**0,59**	**0,81**	**0,17**	**0,19**
40			0,45	**0,69**	0,06	0,2
50			**0,51**	**0,69**	0,16	0,18
0	0,9	0,75	0,01	0,02	0	0,01
2,5			0,3	0,59	0,15	0,14
5			0,36	**0,8**	0,03	0,2
7,5			0,22	0,51	0	0,17
10			0,25	0,54	0,05	0,17
20			0,22	**0,7**	0,03	0,18
30			0,12	0,39	0	0,11
40			0,12	0,52	0	0,14

50			0,1	0,19	0,01	0,07
0	0,9	0,9	0,01	0,02	0	0
2,5			0,34	**0,68**	0,12	0,17
5			0,21	0,48	0,01	0,16
7,5			0,29	**0,72**	0	0,22
10			0,11	0,28	0	0,1
20			0,09	0,38	0	0,11
30			0,05	0,14	0	0,05
40			0,04	0,12	0	0,04
50			0,03	0,1	0	0,03

Najlepszy średni wyniki: 59%, dla parametrów: epsilon = 30, gamma = 0,9, beta = 0,1.

Najlepszy osiągnięty pojedynczy wynik: 88%, dla paramentów takich samych jak najlepszy średni.

Z testów wynika, że epsilon większy od zera zdecydowanie poprawił działanie algorytmu. W szczególności widać to dla testów przy większych gammach.

Również stworzyłem własną funkcję służącą do uczenie algorytmu.

Dodałem:

- Dużą karę za użycie akcji, która skończyła się śmiercią.
- Niewielką karę za używanie akcji, które nie przybliżają do celu (wysokość kary maleje wraz ze wzrostem bliskości do celu), w celu zachęcenia algorytmu do wybierania akcji kierujących go do oczekiwanego pola.
- Nagrodę dla wszystkich pól znajdujących się na ścieżce, która doprowadziła algorytm do celu (nagroda jest przyznawana jednorazowo - jeśli podczas dochodzenia do celu algorytm znalazł się na danym polu i użył tej samej akcji więcej niż raz nie ma to znaczenia).

Badania wpływu parametrów gamma i beta dla mojej funkcji:

Gamma	Beta	Średni wynik	Najlepsza wartość	Najgorsza wartość	Odchylenie standardowe
0,1	0,1	0,18	0,28	0,07	0,05
	0,25	0,17	0,22	0,05	0,05
	0,5	0,11	0,16	0,02	0,04
	0,75	0,11	0,15	0,02	0,04
	0,9	0,07	0,13	0,01	0,05
0,25	0,1	0,29	0,49	0,12	0,1
	0,25	0,26	0,43	0,14	0,09
	0,5	0,36	0,45	0,26	0,06
	0,75	0,16	0,27	0,07	0,07
	0,9	0,1	0,16	0,04	0,03
0,5	0,1	0,5	0,7	0,25	0,11
	0,25	0,56	0,71	0,39	0,08
	0,5	0,54	0,71	0,4	0,11
	0,75	0,32	0,49	0,09	0,12
	0,9	0,17	0,3	0,09	0,06
0,75	0,1	0,63	**0,84**	0,28	0,16
	0,25	0,67	0,77	0,48	0,08
	0,5	0,63	0,69	0,58	0,04
	0,75	0,56	0,7	0,4	0,1
	0,9	0,26	0,38	0,04	0,1
0,9	0,1	**0,72**	**0,88**	0,39	0,12
	0,25	0,65	**0,88**	0,28	0,2
	0,5	**0,75**	**0,85**	0,67	0,06
	0,75	**0,72**	**0,8**	0,61	0,06
	0,9	0,34	0,5	0,2	0,09

Najlepszy średni wyniki: 75%, dla parametrów: gamma = 0,9, beta = 0,5.

Najlepszy osiągnięty pojedynczy wynik: 88%, dla paramentów: gamma = 0,9, beta = 0,1.

Można zauważyć, że zaproponowana przeze mnie funkcja zapewnia dużo lepszą skuteczność działania algorytmu niż funkcja bazowa. Widać, że najlepsze średnie wyniki, jak i najlepsze pojedyncze, są położone w dolnej części tabeli – tam, gdzie są większe wartości parametru gamma.

Badania wpływu parametrów epsilon dla mojej funkcji:

Epsilon	Gamma	Beta	Średni wynik	Najlepsza wartość	Najgorsza wartość	Odchylenie standardowe
0	0,1	0,1	0,18	0,28	0,03	0,07
2,5			0,1	0,25	0,01	0,08
5			0,17	0,33	0,04	0,09
7,5			0,11	0,24	0,01	0,08
10			0,13	0,33	0,01	0,11
20			0,16	0,42	0,01	0,13
30			0,12	0,5	0,04	0,13
40			0,14	0,39	0,01	0,13
50			0,16	0,33	0,01	0,11
0	0,1	0,9	0,09	0,17	0,01	0,05
2,5			0,07	0,22	0,01	0,06
5			0,07	0,19	0,01	0,05
7,5			0,08	0,23	0,02	0,07
10			0,07	0,15	0,01	0,05
20			0,07	0,15	0,01	0,05
30			0,08	0,15	0,01	0,05
40			0,04	0,08	0,01	0,03
50			0,1	0,19	0,03	0,04
0	0,5	0,5	0,52	0,69	0,36	0,1
2,5			0,27	0,4	0,14	0,07
5			0,2	0,43	0	0,13
7,5			0,32	0,6	0,01	0,16
10			0,15	0,27	0	0,08
20			0,22	0,49	0,02	0,17
30			0,21	0,39	0,03	0,11
40			0,2	0,36	0	0,12
50			0,2	0,3	0,07	0,06
0	**0,9**	**0,1**	**0,73**	**0,82**	**0,54**	**0,08**
2,5			0,56	**0,83**	0,28	0,21
5			0,54	**0,83**	0	0,28
7,5			0,47	**0,85**	0	0,23
10			0,36	0,66	0	0,22
20			0,4	**0,86**	0,09	0,26
30			0,53	0,73	0,29	0,13
40			0,53	**0,9**	0,14	0,22
50			**0,68**	**0,83**	0,46	0,12
0	0,9	0,75	**0,65**	0,79	0,45	0,09
2,5			0,43	0,67	0,2	0,16
5			0,48	0,66	0,16	0,14
7,5			0,41	0,66	0,13	0,15
10			0,41	0,7	0,26	0,14
20			0,29	0,4	0,09	0,09
30			0,29	0,51	0,14	0,1
40			0,33	0,52	0,14	0,1
50			0,28	0,48	0,04	0,14
0	0,9	0,9	0,29	0,39	0,18	0,07
2,5			0,26	0,42	0,1	0,11
5			0,28	0,4	0,14	0,09

7,5			0,22	0,33	0,06	0,07
10			0,22	0,38	0,1	0,08
20			0,19	0,25	0,02	0,06
30			0,16	0,25	0,1	0,05
40			0,15	0,24	0,06	0,07
50			0,09	0,19	0,05	0,04

Przy testach wpływu epsilonu na moją funkcję można zobaczyć ogólny spadek wartości średniego wyniku, w porównaniu z testami z epsilonem = 0. Jest to prawdopodobnie spowodowane dodanym elementem nagradzającym całą ścieżkę, więc także losowo wykonane ruchy są nagradzane w sytuacji, gdy algorytm dojdzie do celu, co powoduje zaburzenia w poprawnym uczeniu.

Wnioski

Do poprawnego działania algorytmu wymagany jest prawidłowy dobór parametrów. Z testów można zauważyć, że przynajmniej dla tego problemu bardziej pożądane były wyższe wartości parametru gamma. Dla bazowej funkcji uczenia dodanie elementu losowego (parametr epsilon odpowiadający za zwiększenie eksploracji przez algorytm) zdecydowanie zwiększyło skuteczność algorytmu, również zwiększa się wtedy odchylenie standardowe (czasem algorytm nauczy się bardzo dobrze a czas słabo, co wynika z losowości).

Wydaje mi się, że czynnikiem kluczowym, przez którego sprawność mojej funkcji w stosunku do bazowej oceny, jest kara za wpadnięcie do dziury. Nagradzanie całej ścieżki, którą algorytm podążał przy dojściu do celu powoduje zmniejszenie umiejętności eksploracyjnych i algorytm skupia się tylko na danej ścieżce, co może skutkować nieznalezieniem potencjalnie lepszej drogi.