Uczenie ze wzmocnieniem RL (ang. Reinforcement Learning)

By kama3 kama3
lis 3, 2025

Q-learning / Deep Q-Networks (DQN)

Maszyna uczy się poprzez nagrody i kary; może być używana do optymalizacji strategii konserwacji (np. Predictive Maintenance).

Aby sprostać temu wyzwaniu, alternatywnym podejściem jest połączenie uczenia Q z głębokimi sieciami neuronowymi . To podejście nazywa się głębokim uczeniem Q (Deep Q-Learning – DQL) . Sieci neuronowe w DQL działają jako aproksymator wartości Q dla każdej pary (stan, działanie).

Sieć neuronowa odbiera stan jako sygnał wejściowy i generuje wartości Q dla wszystkich możliwych działań . Poniższy rysunek ilustruje różnicę między uczeniem Q a głębokim uczeniem Q w ocenie wartości Q.

Termin Deep Q-Network odnosi się do sieci neuronowej w architekturze DQL .

DQN uczy się optymalnej strategii, wykorzystując głęboką sieć neuronową do szacowania wartości Q, bufor pamięci do przechowywania danych z przeszłości oraz sieć docelową, aby zapobiec przeszacowaniu wartości Q. Agent stosuje strategię eksploracji epsilon-greedy podczas treningu i wybiera akcję o najwyższej wartości Q podczas testowania.

Zalety:
Bezmodelowy (Model-Free): Nie wymaga wcześniejszej wiedzy o dynamice środowiska, ucząc się bezpośrednio z interakcji.
Gwarantowana zbieżność: Przy wystarczającej eksploracji i skończonej liczbie stanów/akcji, algorytm gwarantuje zbieżność do optymalnej polityki.
Prostota implementacji: Jest stosunkowo prosty do wdrożenia, szczególnie w przypadku małych problemów z użyciem tablicy Q.
Obsługa opóźnionych nagród: Skutecznie radzi sobie z nagrodami, które są odroczone w czasie.

Wady:
Problem "klątwy wymiarowości": Wymaga przechowywania i aktualizowania wartości dla każdej pary stan-akcja (tabela Q), co staje się niewykonalne w dużych lub ciągłych przestrzeniach stanów/akcji.
Wolna konwergencja: Proces uczenia może być bardzo powolny w środowiskach o dużej liczbie stanów i akcji, wymagając wielu epizodów.
Wrażliwość na hiperparametry: Wydajność algorytmu jest bardzo zależna od właściwego doboru współczynników uczenia (alpha), dyskontowania (gamma) i eksploracji (epsilon).
Ograniczony do dyskretnych akcji: Standardowy Q-learning działa tylko w środowiskach z dyskretnym (skończonym) zbiorem akcji.

Read more

Policy Gradient / Actor-Critic

Zaawansowane techniki RL stosowane np. do dynamicznego planowania inspekcji i sterowania konserwacją.

Teoretyczne podstawy algorytmów aktora-krytyka. Algorytm aktora-krytyka (AC) składa się z dwóch części: estymacji funkcji wartości danej polityki oraz aktualizacji (ulepszania) tej polityki. W tej sekcji przedstawiamy analizę teoretyczną, która
pomoże w opracowaniu takiego algorytmu poprzez ewaluację polityki (PE) i gradient polityki (PG).

Zalety:
Obsługa ciągłych i wysokowymiarowych przestrzeni akcji: W przeciwieństwie do metod opartych na wartościach (np. Q-learning), które wymagają iteracji po wszystkich możliwych akcjach, metody PG mogą naturalnie obsługiwać akcje ciągłe (np. sterowanie robotem).
Uczenie stochastycznych polityk: PG uczą się rozkładu prawdopodobieństwa akcji, co ułatwia eksplorację środowiska i radzenie sobie z niepewnością (problem aliasowania percepcyjnego, gdy dwa różne stany wyglądają tak samo).
Bezpośrednia optymalizacja celu: Bezpośrednio dążą do maksymalizacji nagrody, co może prowadzić do lepszych właściwości zbieżności w niektórych problemach.

Wady:
Duża wariancja estymacji gradientu: Algorytmy te (np. REINFORCE) opierają się na próbkowanych trajektoriach (metoda Monte Carlo), co prowadzi do dużej wariancji estymacji gradientu, a w konsekwencji do powolnego i niestabilnego uczenia.
Niska efektywność próbkowania (sample inefficiency): Wymagają dużej ilości interakcji ze środowiskiem (danych) do skutecznego uczenia.
Zbieżność do lokalnego optimum: Jako metody gradientowe, zbiegają się do lokalnych, a nie globalnych maksimów funkcji nagrody.