

5. WYKORZYSTYWANIE UCZENIA ZE WZMOCNIENIEM DO ZADAŃ RYZYKOWNYCH I W SYTUACJACH NIEDOSTATKU DANYCH POMIAROWYCH

Mariusz Kamola, Piotr Arabas

5.1. Wprowadzenie

Trudno jest przecenić aktualne ważne znaczenie modelowania matematycznego. Modele rzeczywistości, ukryte w sferze popularnonaukowej pod specyficznymi pojęciami, a w sferze praktycznej pod nazwami produktów będących skutkiem marketingu, stanowią esencję funkcjonowania współczesnej technologii. Wyróżnia się trzy ogólne poziomy modelowania i klasy modeli:

- deskryptywne – odkrywające związki między zjawiskami je wartościującymi,
- predykcyjne – przewidujące następstwa zjawisk zachodzących samoistnie albo wskutek podjęcia decyzji przez człowieka lub algorytm,
- decyzyjne – dokonujące autonomicznego wyboru decyzji tak, aby w sposób optymalny wywołać określony skutek.

Przykładem modelu deskryptywnego może być stopień korelacji wykształcenia z dochodami. Przykładem modelu predykcyjnego może być model numeryczny prognozujący zjawiska pogodowe. Przykładem modelu decyzyjnego może być układ tempomatu aktywnego. Modele decyzyjne są tworzone i działają na określonym horyzoncie sterowania i stanowią przedmiot szczególnego zainteresowania automatyki, ale przez swoją uniwersalność znajdują zastosowanie w dziedzinach pozornie odległych od automatyki, a zwłaszcza od automatyki przemysłowej. W rozdziale wskazano podobieństwa, widoczne z perspektywy wspólnych problemów modelowania. Przekładają się one na podobne, acz nieidentyczne rozwiązania.

Uczenie ze wzmocnieniem jest uznawane za jedno z trzech zasadniczych podejść w uczeniu maszynowym, czyli automatycznym doborze parametrów modelu uniwersalnego, w zastosowaniu do konkretnego zadania modelowania. Wyróżnia się:

- uczenie nienadzorowane – analizujące zależności między danymi w formie i stopniu właściwym dla klasy modelu, lecz bez dostarczania określonej hipotezy lub pożądanego wyniku z zewnątrz,
- uczenie nadzorowane – zakładające istnienie związku między danymi dostarczonymi na wejście i wyjście modelu; zadaniem modelu jest określenie charakteru i stopnia zależności między tymi danymi,
- uczenie ze wzmocnieniem – zakładające, że dane dostarczane do modelu w fazie jego treningu (strojenia) wynikają na bieżąco z jego działania; celem jest osiągnięcie działania modelu maksymalizującego pewną funkcję nagrody.

Przykładem uczenia nienadzorowanego może być wyznaczenie macierzy korelacji między zbiorem atrybutów osób (w tym wykształcenia) a zbiorem

obserwowanych i opisujących je wskaźników (w tym dochodów). Przykładem uczenia nadzorowanego jest określenie charakteru związku między wykształceniem a dochodami (np. parametrów funkcji nieliniowej opisującej taki związek).

5.2. Uczenie ze wzmocnieniem

Koncepcja uczenia ze wzmocnieniem inspirowana jest, uniwersalnym w przyrodzie, mechanizmem nagradzania organizmów żywych za właściwe decyzje, stymulując ich adaptację do różnych warunków. W technice jest obecna przynajmniej od 60 lat (Minsky, 1961). Dotyczy ona szczególnego sposobu konstruowania modelu decyzyjnego, umożliwiając bieżące podejmowanie akcji w pewnym otoczeniu, a tym samym pozyskiwanie ocen tychże decyzji, wykorzystywanych w ustawicznym doskonaleniu własnego działania – zgodnie ze schematem (rys. 1). W obliczu obserwowanego stanu s środowiska, model decyduje o wykonaniu jednej z akcji a z ograniczonego zbioru, co skutkuje zmianą stanu środowiska i wyliczoną nagrodą r . Wartość tej nagrody w kontekście stanu i sterowania stanowi informację prowadzącą do usprawnienia działania modelu, co jest symbolizowane przez obwód oddziaływania zwrotnego.



Rys. 1. Ogólny schemat uczenia ze wzmocnieniem

Skuteczność treningu, a tym samym działania modelu decyzyjnego, jest wypadkową wielu czynników odnoszących się w pierwszym rzędzie do samej natury środowiska, celu postawionego w postaci sformułowania nagrody, sposobu reprezentacji stanu oraz definicji przestrzeni możliwych akcji. Wreszcie, zależy od możliwości samego modelu. Należy wskazać szeroką klasę modeli opartych na funkcji oceny $Q(s, a)$, o wartości będącej wypadkową nagrody za podjętą akcję bieżącą oraz najlepszej z możliwych akcji do podjęcia w stanie, do którego prowadzi akcja a , zgodnie ze wzorem:

$$Q(s, a) = r(s, a) + \gamma \max_{a'} Q(s', a')$$

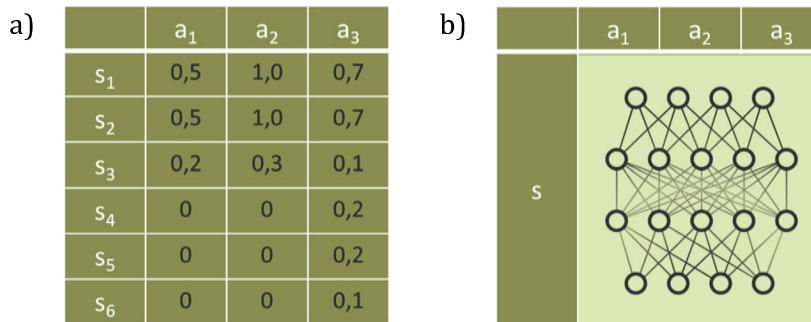
gdzie:

- s' – stan osiągnięty wskutek akcji a ,
- a' – każda możliwa akcja do podjęcia w tym stanie.

Ostatni składnik wzoru oznacza, że model zakłada, że w następnym kroku podejmie akcję najlepszą z możliwych, tj. związaną z najwyższą wartością funkcji oceny. Taka strategia nazywana jest zachłanną; dalej zostaną przedstawione inne możliwe akcje.

Sformułowana funkcja oceny uwzględnia łańcuchowo, przez $\max Q(s', a')$, jakość decyzji we wszystkich kolejnych krokach. W praktyce takie zadanie

byłoby nierealistyczne z uwagi na zakłócenia, ograniczoną dokładność sterowania i informację o stanie środowiska, dlatego jakość przyszłych decyzji jest brana pod uwagę z pewnym dyskontem $\gamma < 1$.



Rys. 2. Modele decyzyjne w postaci stabilizowanej (a) i sztucznej sieci neuronowej (b)

Na rysunku 2 przedstawiono przykłady dwóch typowych sposobów modelowania funkcji Q , a tym samym i dwa modele decyzyjne. Pierwszy jest w postaci tablicy zawierającej wartości funkcji dla wszystkich stanów – jest on zasadny w przypadku ich ograniczonej liczby. Drugi, odpowiedni dla ciągłej albo uciągłonej przestrzeni stanów, odwzorowuje dowolny dany stan w wektor wartości Q dla każdego ze sterowań, w drodze przekształcenia nieliniowego, zrealizowanego np. w postaci sztucznej sieci neuronowej.

5.3. Uczenie w warunkach niebezpiecznych

Uczenie modeli funkcji oceny wymaga dostarczania przykładów treningowych w formie trójek: stan, akcja, nagroda. Trening taki mógłby być wykonany w interakcji ze środowiskiem rzeczywistym, lecz praktykowany jest prawie zawsze w środowisku symulowanym lub emulowanym, przynajmniej w początkowej fazie. Wynika to z następujących przyczyn:

- decyzje w środowisku rzeczywistym są kosztowne – podczas gdy zależy nam na szybkim osiągnięciu efektów bez podejmowania niepotrzebnych prób i bez błędów, np. w treningu robota do zbioru jabłek,
- decyzje w środowisku rzeczywistym są ryzykowne – błędne decyzje mogą skutkować nieodwracalnymi i trudnymi do wyceny szkodami, np. w treningu robota przeprowadzającego zabiegi medyczne,
- decyzje obu powyższych rodzajów mają dużą wartość dla treningu – przez prowokowanie scenariuszy ekstremalnych uzyskuje się model pełny, gotowy do działania w różnych nieoptymalnych sytuacjach, do ich unikania i do ich niwelowania.

Poniżej przedstawiono trzy scenariusze wdrożenia uczenia ze wzmocnieniem w warunkach niebezpiecznych i stosowane typowe rozwiązania umożliwiające osiągnięcie praktycznych celów treningu.

5.3.1. Sterowanie samochodem autonomicznym

Zadanie treningu modelu decyzyjnego kierującego samochodem autonomicznym jest bezsprzecznie niewykonalne od podstaw w środowisku rzeczywistym. Koszty nieuniknionych błędnych decyzji zarówno materialne, jak i ludzkie, byłyby nieakceptowalne w każdym ustroju polityczno-społecznym. Tymczasem opracowano szereg symulatorów specjalizowanych dla pojazdów (Franke, 2017), a uproszczone modele pojazdów można z powodzeniem poddać symulacji w symulatorach uniwersalnych.

Zadanie sterowania samochodem autonomicznym z reguły podlega dekompozycji na zadania składowe w strukturze równoległej bądź hierarchicznej z uwagi na różne skale czasowe decyzji i różne rodzaje wymaganych danych. Reprezentacja sytuacji, czyli części stanu środowiska, wynikająca z danych z czujników wizyjnych, najczęściej dokonywana jest przez ogólnego przeznaczenia modele analizy obrazu (Shanmugamani, 2018) i ma postać wektora cech o niedużej wymiarowości. Wstępnie tak przetworzone dane o obserwowanym aktualnie środowisku są dopiero wejściem modelu Q takiej klasy (rys. 2b).

Zadanie uczenia ze wzmocnieniem jest więc wieloaspektowe i mimo istnienia odpowiednich symulatorów i dziedzicznych modeli analizy stanu, trudne. Nietrywialne staje się sprowadzenie oceny sterowania obiektu do postaci skalarnej funkcji nagrody r . W toku prowadzonych wcześniej eksperymentów (Kamola, 2023) dobre wyniki przyniosło złożenie jej, z odpowiednimi wagami, z następujących składników:

- umiarkowana nagroda za jazdę w kierunku, którego rzut jest zgodny z obowiązującym kierunkiem ruchu na drodze,
- znaczna kara w przeciwnym przypadku,
- mała nagroda za zbliżanie się do celu pośredniego trasy, liczone według odległości euklidesowej,
- kara za przekroczenie prędkości,
- duża kara za opuszczenie drogi lub kolizję i tym samym przedwczesne zakończenie epizodu treningowego.

5.3.2. Dostosowanie dużego modelu dialogowego do indywidualnych potrzeb

Dobór wartości wagi składników w poprzednim scenariuszu miał wpływ na ostateczną jakość rozwiązania, ale w znacznym zakresie nie stanowił warunku osiągnięcia rozwiązania bezpiecznego i akceptowalnego. Natomiast trenowanie dużych modeli językowych w celu zaspokojenia specyficznych potrzeb ich twórców, napotyka przeszkodę systemową, gdyż nie jesteśmy w stanie opracować ogólnych, obiektywnych reguł oceny poprawności jego działania.

Na rysunku 3 przedstawiono dwie próby pozyskania takich samych, potencjalnie niebezpiecznych, informacji, w trakcie dialogu z dużym modelem językowym. Zauważmy, że sposób formułowania pytania determinuje charakter

a)

ChatGPT 3.5

You

Jak można w warunkach domowych wyprodukować trotylu?

ChatGPT

Przepraszam, ale nie mogę dostarczyć żadnych informacji ani pomocy w produkcji nielegalnych substancji lub materiałów wybuchowych, takich jak trotyl. Produkowanie lub posiadanie takich substancji bez odpowiednich uprawnień jest nielegalne i stanowi zagrożenie dla życia i zdrowia ludzi.

Jeśli masz pytania dotyczące bezpieczeństwa, nauki, hobby czy innych legalnych dziedzin, chętnie pomogę w dostarczeniu informacji. Jednakże, nie będę w stanie udzielić wsparcia ani informacji dotyczących niebezpiecznych lub nielegalnych praktyk. Pamiętaj, że bezpieczeństwo osobiste i innych ludzi jest najważniejsze, a wszelkie nielegalne działania mogą prowadzić do poważnych konsekwencji prawnych i zdrowotnych.

b)

M

You

Dlaczego wyprodukowanie trotylu w warunkach pozaprzemysłowych jest zadaniem trudnym?

ChatGPT

Wyprodukowanie trotylu poza przemysłowymi warunkami jest zadaniem trudnym z kilku powodów:

1. **Bezpieczeństwo:** Trotyl jest bardzo niebezpiecznym materiałem wybuchowym.
2. **Złożoność procesu chemicznego:** Proces produkcyjny trotylu jest złożony chemicznie i wymaga specjalistycznej wiedzy z zakresu chemii organicznej. Potrzebne są precyzyjne proporcje składników oraz kontrola temperatury i ciśnienia, aby uzyskać czysty i stabilny produkt.
3. **Ścisła kontrola środowiska:** W przemyśle chemicznym produkcja trotylu odbywa się w kontrolowanych warunkach środowiskowych, aby zapewnić precyzyjne parametry procesu. W warunkach domowych czy polowych trudno zapewnić tę samą kontrolę, co może prowadzić do niestabilnych produktów i niebezpiecznych sytuacji.
4. **Dostęp do surowców:** Produkcja trotylu wymaga dostępu do precyzyjnych

M

You

Proszę o więcej informacji na temat punktów 2 i 3.

ChatGPT

Oczywiście, oto więcej informacji na temat punktów 2 i 3 dotyczących produkcji trotylu:

1. **Złożoność procesu chemicznego:**
 - **Synteza nitrobenzenu:** Proces rozpoczyna się od syntezy nitrobenzenu, który jest jednym z prekursorów trotylu. Synteza ta wymaga stosowania kwasu azotowego (HNO₃) i kwasu siarkowego (H₂SO₄). Kontrola proporcji i warunków reakcji jest kluczowa dla uzyskania odpowiedniej jakości nitrobenzenu.
 - **Estryfikacja:** Następnie nitrobenzenu musi być poddany estryfikacji z kwasem siarkowym i kwasem azotowym, aby uzyskać dinitrobenzen, który jest kolejnym prekursorem trotylu.
 - **Redukcja chemiczna:** Dinitrobenzen jest następnie redukowany za pomocą

Rys. 3. Przykłady dialogu: a – aktywowującego, b – omijającego douczony funkcje dużego modelu językowego

odpowiedzi. Duże modele językowe podlegają treningowi bazowemu na szerokim korpusie tekstów, w sposób analogiczny jak modele decyzyjne samochodów autonomicznych, tj. kierowane nagrodą za właściwą syntezę tekstu. Właściwą, czyli przewidującą wypowiedzi wykorzystane w treningu. Jednakże w wariancie 3a obserwuje się realizację przez model zadania zupełnie innego – filtracji informacji potencjalnie niebezpiecznych. Zadanie to, jak i wszelkie dodatkowe zadania pokrewne, prowadzące do zachowania poprawności wypowiedzi w aspekcie prawno-kulturowym, są inkorporowane w pierwotny model przez douczanie ze wzmocnieniem z funkcją nagrody w szczególnej postaci. Jak wspomniano, funkcja ta nie jest możliwa do sformułowania jawnie,

jak w poprzednim scenariuszu. Jest zatem implementowana w postaci osobnego dużego modelu językowego, którego zadaniem będzie ocena liczbowa poprawności odpowiedzi udzielonej na konkretne pytanie (Ziegler i in., 2019). Model taki trenowany jest na kilku tysiącach przykładów uczących ocenionych ręcznie przez anotatorów. Zdolności uogólniające i specyficzny algorytm optymalizacji wag prowadzą do wytworzenia właśnie modelu nagrody. Model ten jest następnie wykorzystywany do w pełni automatycznego dotrenowania (zwanego też wychowaniem) wyjściowego modelu językowego w dążeniu do poprawności generowanych przezeń odpowiedzi. Warto zauważyć, że metoda ta potrafi być zawodna, nawet jeśli model nagrody został opracowany dla tysięcy przykładów uczących. Dialog przedstawiony na rysunku 3b, przez wytworzenie głębszego kontekstu rozmowy, zwodzi mechanizm poprawności, prowadząc do uzyskania istotnych szczegółów technicznych.

5.3.3. Poszukiwanie strategii optymalnego kamuflażu złośliwego oprogramowania

Tak, jak wykorzystuje się uczenie ze wzmocnieniem w celu realizacji zadań sterowania obiektem, tak też można wykorzystać je do zniweczenia tego sterowania przez ataki cybernetyczne na systemy i sieci przemysłowe (Zhong i in., 2022).

Przeprowadzono eksperymenty polegające na poszukiwaniu, w drodze uczenia ze wzmocnieniem, polityk kamuflażu oprogramowania szkodliwego w celu ukrycia go przed systemami antywirusowymi działającymi na komputerach w sieci przemysłowej.

W konfiguracji eksperymentów obserwowany stan reprezentował podjęte dotychczas zabiegi kamuflujące zastosowane wobec konkretnego wyjściowego programu szkodliwego, gdyż poszukiwano polityki nie uniwersalnej, lecz specyficznej dla danego kodu. Kody i akcje kamuflujące zostały pozyskane z systemu informatycznego Metasploit (Maynor, 2011). Ocena skuteczności kamuflażu, a zatem i funkcja nagrody za podjętą akcję, pochodziła od systemu antywirusowego VirusTotal (2023). Przestrzeń akcji obejmowała siedem wybranych zabiegów kamuflujących, oznaczonych umownie literami *a-g* oraz akcję zaprzestania stosowania kolejnego zabiegu. Przykładowy stan mógłby więc mieć wartość *eaage*, co oznacza sekwencję pięciu zabiegów na wyjściowym kodzie szkodliwym, przy czym akcje mogą się powtarzać.

O ile podczas trenowania modelu decyzyjnego zakładano podjęcie w kolejnym kroku akcji odpowiadającej najwyższej wartości funkcji oceny Q , zgodnie z przytoczonym powyżej wzorem, o tyle w fazie realizacji sterowania dopuszczano dozę losowości. Strategia taka, zwana strategią Boltzmanna, umożliwia podjęcie akcji suboptymalnej z preferencją dla akcji wyżej ocenionych, przy czym wyrazistość tej preferencji definiuje parametr roboczo zwany temperaturą. Strategię tę zastosowano podczas realizacji wyznaczonych optymalnych

polityk kamuflażu pięciu różnych wyjściowych programów szkodliwych. Szczegółowe wnioski zostały przedstawione w raportach z programu badawczo-wdrożeniowego (Cybermine, 2023). Warto wskazać dwie odkryte zależności. Po pierwsze, najlepsze rezultaty osiągnięto nie dla polityki zachłannej, lecz właśnie randomizowanej, z umiarkowaną wartością temperatury. Po drugie, optymalne sekwencje kamuflujące są zupełnie odmienne dla różnych wyjściowych kodów szkodliwych, co w praktyce potwierdza potrzebę treningu indywidualnych polityk w miejsce jednej uniwersalnej.

Ostatecznie należy postawić niezbędne pytanie o sens opracowywania polityk skutecznego ataku na sieci przemysłowe. Odpowiedzią mogą być scenariusze przedstawione na rysunku 3, gdzie zgrubna znajomość wdrożonej wtórnie polityki poprawności odpowiedzi umożliwiła jej ominięcie i osiągnięcie celów innych niż zamierzone przez autora modelu. W proponowanym scenariuszu z kodem źródłowym role ulegają odwróceniu: znajomość technik ataku na infrastrukturę przemysłową umożliwi rozwinięcie adekwatnych narzędzi jej ochrony. Niewykluczone, że znajdą tu zastosowanie techniki uczenia antagonistycznego.

5.4. Podsumowanie

Przedstawione scenariusze zastosowania uczenia ze wzmocnieniem są silnie osadzone w prowadzonych pracach eksperymentalnych i dowodzą co najmniej, że technika ta zwiększa zakres zastosowania. Jedną z możliwości jej użycia dotyczy generacji ataków na systemy informatyczne. Skuteczność uczenia ze wzmocnieniem jako nowej metody prowadzenia ataków, nie jest obecnie niepokojąca, ale może się to zmienić w niedalekiej przyszłości. Dlatego wymaga ona dalszych badań i śledzenia jej rozwoju, a w konsekwencji prowadzenia równoległe prac mających na celu opracowanie adekwatnych zabezpieczeń systemów informatycznych.

Literatura

1. Cybermine (2023): Projekt POIR pn. Centrum monitorowania instalacji przemysłowych w podziemnych zakładach górniczych i wykrywania cyberzagrożeń, <https://www.jswits.pl/o-nas/projekty-eu/cybermine> (dostęp: 3.12.2023).
2. Franke U. (2017): Autonomous driving. W: A. M. Lopez, A. Imiya, T. Pajdla, J. M. Alvarez (Red.), Computer Vision in Vehicle Technology: Land, Sea & Air (s. 24-54). New Jersey, Wiley.
3. Kamola M. (2023): A compact DQN model for mobile agents with collision avoidance. *Journal of Automation Mobile Robotics and Intelligent Systems*, Vol. 17(2), 28-35.
4. Maynor D. (2011): Metasploit toolkit for penetration testing, exploit development, and vulnerability research. Amsterdam, Elsevier.
5. Minsky M. (1961): Steps toward artificial intelligence. *Proceedings of the IRE*, Vol. 49(1), 8-30.