POLSKA AKADEMIA NAUK
Instytut Badań Systemowych

WSPOMAGANIE DECYZJI
SYSTEMY EKSPERCKIE

pod redakcją

Romana Kulikowskiego i Lucyny Bogdan

Warszawa 1995

WSPOMAGANIE DECYZJI

SYSTEMY EKSPERCKIE

pod redakcją

Romana Kulikowskiego i Lucyny Bogdan

Warszawa 1995

Wydano z wykorzystaniem dotacji
KOMITETU BADAŃ NAUKOWYCH

Materiały konferencji: "Analiza Decyzyjna, Systemy Eksperckie, Zastosowania Systemów Komputerowych", Warszawa, 25-27 maja 1994r.

Komitet Programowy Konferencji:

Andrzej Ameljańczyk, Zdzisław Bubnicki, Wiesław Grudzewski, Olgierd Hryniewicz, Janusz Kacprzyk, Lech Kruś, Roman Kulikowski (przewodniczący), Kazimierz Mańczak, Ireneusz Nykowski, Zdzisław Pawlak, Roman Słowiński, Andrzej Straszak, Andrzej Weryński, Andrzej Wierzbicki.

Wykonano z oryginałów tekstowych dostarczonych przez autorów

ISBN 83-85847-85-5

Uczenie się sekwencyjnego podejmowania decyzji w oparciu o opóźnione nagrody*

Paweł Cichosz Jan J. Mulawka
Instytut Podstaw Elektroniki
Politechniki Warszawskiej

Abstract

Prezentowana praca poświęcona jest interesującemu paradygmatowi uczenia się maszyn, znanemu jako uczenie się ze wzmocnieniem. W odróżnieniu od lepiej znanych systemów uczących się z nadzorem, system uczący się ze wzmocnieniem otrzymuje informację trenującą o charakterze wartościującym, która nie specyfikuje wprost wymaganych w określonych sytuacjach decyzji systemu. Jego zadaniem jest nauczenie się strategii decyzyjnej maksymalizującej otrzymywane nagrody w długim horyzoncie czasowym. Artykuł stanowi zwięzłe wprowadzenie do dziedziny uczenia się ze wzmocnieniem i wykorzystywanych w niej algorytmów.

1 Wprowadzenie

Jednym z kryteriów, według których można porządkować metody uczenia się maszyn, jest kryterium rodzaju dostępnej informacji trenującej. W tradycyjnym *uczeniu się z nadzorem* informacja ta mniej lub bardziej bezpośrednio określa pożądane decyzje systemu w danych warunkach wejściowych. Może ona mieć np. postać par wektorów trenujących przy uczeniu się odwzorowań, przykładów pozytywnych i negatywnych przy uczeniu się pojęć, tablic decyzyjnych przy uczeniu się reguł decyzyjnych, etc. Niestety, w złożonych problemach świata rzeczywistego wiedza niezbędna do dostarczenia systemowi uczącemu się takiej informacji może być niedostępna, trudna do zdobycia lub sformułowania, bądź mało wiarygodna. Uzasadnia to celowość zwrócenia uwagi na metody uczenia się bazujące na wartościującej informacji trenującej: metody *uczenia się ze wzmocnieniem* (ang. *reinforcement learning*).

*Autorzy dziękują za wsparcie ich pracy przez Komitet Badań Naukowych w ramach grantu nr 8 S503 019 05.

Brak miejsca zmusza nas do ograniczenia się w tym artykule do zagadnień najbardziej podstawowych. Jego dalszy ciąg zorganizowany jest następująco: paragraf 2 zawiera omówienie paradygmatu uczenia się ze wzmocnieniem; w paragrafie 3 zwięźle przedstawiono algorytmy wykorzystywane do rozwiązania kluczowego dla uczenia się ze wzmocnieniem problemu temporalnego przypisania zasługi; paragraf 4 poświęcony jest końcowej dyskusji.

2 Uczenie się ze wzmocnieniem

W uczeniu się ze wzmocnieniem rozważamy uczący się system, nazywany dalej dla wygody *agentem*, oraz inny system, który nazywa się *środowiskiem*. W każdej chwili dyskretnego czasu agent obserwuje aktualny *stan* środowiska i wykonuje pewną *akcję*, zgodnie ze swoją *strategią decyzyjną*. Następnie otrzymuje wartość *wzmocnienia*, zwaną także nagrodą (karą) lub wypłatą, która jest pewną miarą jakości wykonanej akcji, oraz następuje zmiana stanu środowiska. Zarówno wartości wzmocnienia, jak i zmiany stanów mogą być w ogólnym przypadku stochastyczne, przy czym ani odpowiednie wartości oczekiwane nagród, ani rozkłady prawdopodobieństw zmian stanów nie są znane. Zadaniem agenta jest nauczenie się strategii decyzyjnej (tj. przyporządkowania stanom środowiska akcji) prowadzącej do maksymalizacji wartości wzmocnienia otrzymywanych *w długim horyzoncie czasowym*.

2.1 Opóźnienie wzmocnienia

Typowo kryterium do optymalizacji przez system uczący się ze wzmocnieniem formalizuje się jako wartość oczekiwaną *zdyskontowanej* całkowitej sumy wzmocnienia, tj.

$$\mathbf{E}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right], \tag{1}$$

gdzie r_t jest wartością wzmocnienia w kroku czasu t, a $\gamma \in [0,1]$ jest tzw. *współczynnikiem dyskontowania* (ang. *discount factor*), determinującym względną ważność wartości wzmocnienia bliskich i odległych w czasie. Aby maksymalizować kryterium wyrażone sumą (1) dla $\gamma > 0$, agent musi brać pod uwagę nie natychmiastowe, lecz długoterminowe konsekwencje swoich akcji. Jest to związane z występującym (potencjalnie) zjawiskiem *opóźnienia wzmocnienia* [4].

2.2 Procesy Markowa i programowanie dynamiczne

Modelem matematycznym tak rozumianego zadania uczenia się ze wzmocnieniem jest tzw. *proces decyzyjny Markowa*. W przypadku, gdy parametry środowiska

(tj. prawdopodobieństwa zmian stanów oraz oczekiwane wartości wzmocnienia dla wszystkich par stanākcja) są w pełni znane, optymalna strategia dla takiego procesu może zostać znaleziona z wykorzystaniem metod stochastycznego programowania dynamicznego [3]. W złożonych problemach taka wiedza jest rzadko dostępna i stąd wynika potrzeba algorytmów uczenia się ze wzmocnieniem. Warto wspomnieć, że algorytmy omawiane w paragrafie 3 są blisko spokrewnione z pewnymi metodami programowania dynamicznego [1, 6, 2].

2.3 Problem przypisania zasługi

Podstawowy problem, jaki musi zostać rozwiązany przez system uczący się ze wzmocnieniem, jest znany jako *problem przypisania zasługi* (ang. *credit assignment*). Sutton [4] wprowadził szeroko przyjętą dekompozycję tego problemu na problem *temporalnego przypisania zasługi* i *strukturalnego przypisania zasługi*. Pierwszy z nich polega na przypisaniu „zasługi" za uzyskane przez system dochody poszczególnym akcjom, podjętym być może wiele kroków wcześniej, zanim dochody były zaobserwowane. Drugi oznacza wykorzystanie zasługi przypisanej w ten sposób akcjom do odpowiedniej modyfikacji wiedzy systemu, reprezentowanej w pewien sposób w jego strukturze. W tej pracy skupimy się na problemie temporalnego przypisania zasługi, ponieważ strukturalne przypisanie zasługi może być sprowadzone do zagadnień metod reprezentacji wiedzy, generalizacji etc., które są szerzej znane.

3 Temporalne przypisanie zasługi

Ze względu na brak miejsca, dwa podstawowe algorytmy temporalnego przypisania zasługi, *AHC* [4] i *Q-learning* [6, 7], będą tu przedstawione bardzo pobieżnie. Zainteresowanych Czytelników odsyłamy do oryginalnych publikacji twórców tych algorytmów lub pracy [2]. Szczególnie godne uwagi, a nie omówione tutaj, są powiązania obu algorytmów z tzw. metodami *różnic czasowych* (ang. *temporal differences*) — klasą metod uczenia się predykcji w wieloetapowych problemach predykcyjnych znaną jako TD(λ) [5]. Cichosz [2] dyskutuje te powiązania oraz metody ich efektywnej praktycznej implementacji.

Oba algorytmy dokonują oszacowania pewnych funkcji, określonych na przestrzeni stanów lub par stanākcja, stanowiących aktualną wiedzę systemu. Uczenie się polega na ich modyfikacji pod wpływem kolejnych doświadczeń: w każdym kroku czasu następuje uaktualnianie wartości funkcji dla ostatnio wykonanej akcji i stanu, w którym została wykonana, z wykorzystaniem otrzymanej wartości wzmocnienia i obserwacji następnego stanu. Przy prezentacji algorytmów używamy notacji

$$uaktualnij^{\alpha}(f,\ p_0, p_1, \ldots, p_{n-1},\ \Delta)$$

do zapisania operacji uaktualnienia wartości pewnej n-argumentowej funkcji f dla wartości argumentów $p_0, p_1, \ldots, p_{n-1}$ z wykorzystaniem wartości błędu Δ, tak aby stała się ona bliższa $f(p_0, p_1, \ldots, p_{n-1}) + \Delta$, w stopniu kontrolowanym przez współczynnik szybkości uczenia α. Ponadto, używamy symboli x_t, a_t i r_t do oznaczenia odpowiednio stanu, akcji i wzmocnienia w chwili czasu t.

3.1 Algorytm AHC

Algorytm AHC wykorzystuje dwie funkcje: *funkcję wartościowania* V i *funkcję strategii* π. Funkcja wartościowania wartościuje każdy stan środowiska, przypisując mu szacowaną wartość zdyskontowanej sumy wzmocnienia, jakie będzie otrzymane przez agenta rozpoczynającego działanie w tym stanie i wybierającego akcje zgodnie z aktualną strategią. Funkcja strategii przyporządkowuje każdej parze stanākcja (x, a) liczbę rzeczywistą reprezentującą względną korzystność wykonania akcji a w stanie x. Wartości funkcji strategii dla danego stanu i wszystkich akcji są używane do wyboru akcji do wykonania w tym stanie, według pewnego (zazwyczaj probabilistycznego) mechanizmu selekcji.

Obie funkcje są uaktualniane w każdym kroku czasu t zgodnie z następującymi regułami:

$$uaktualnij^{\alpha}(V,\ x_t,\ r_t + \gamma V_t(x_{t+1}) - V_t(x_t));$$

$$uaktualnij^{\beta}(\pi,\ x_t, a_t,\ r_t + \gamma V_t(x_{t+1}) - V_t(x_t)).$$

Zgodnie z pierwszą regułą, wartościowanie stanu x_t powinno stać się bliższe sumie natychmiastowego wzmocnienia otrzymanego w tym stanie r_t i (zdyskontowanego) wartościowania następnego stanu x_{t+1}. Reguła dla funkcji strategii oznacza, że ocena korzystności wykonania akcji jest zwiększana bądź zmniejszana w zależności od tego, czy jej długoterminowe konsekwencje wydają się (na podstawie obserwacji wzmocnienia i następnego stanu) być odpowiednio lepsze bądź gorsze, niż wcześniej oczekiwane.

3.2 Algorytm Q-learning

Algorytm Q-learning uczy się jednej funkcji, nazywanej *Q-funkcją*. Każdej parze stanākcja (x, a) przyporządkowuje ona tzw. *Q-wartość* $Q(x, a)$, która jest oszacowaniem zdyskontowanej sumy wzmocnienia, jakie będzie otrzymane przez agenta rozpoczynającego działanie w stanie x wykonaniem akcji a i następnie wybierającego akcje o maksymalnych Q-wartościach. Q-funkcja zastępuje funkcję strategii i jest używana do wyboru akcji, ale zastępuje także funkcję wartościowania, gdyż jest używana również do oceny użyteczności stanów.

Reguła uaktualniania Q-funkcji ma postać:

$$uaktualnij^{\alpha}(Q,\ x_t, a_t,\ r_t + \gamma \max_a Q_t(x_{t+1}, a) - Q_t(x_t, a_t)).$$

Można ją intuicyjnie interpretować w sposób zbliżony do interpretacji reguł poprzedniego algorytmu, przyjmując, że maksymalna Q-wartość w danym stanie stanowi miarę jego użyteczności i zastępuje wartość funkcji V używanej przez AHC.

4 Konkluzja

Prezentowany paradygmat uczenia się jest niezwykle pojemny. Aby sformułować w jego terminach konkretny problem do rozwiązania, należy określić zbiór rozpoznawanych przez agenta stanów środowiska i jego repertuar akcji oraz funkcję wzmocnienia, która stanowi właściwą specyfikację zadania. Metody uczenia się ze wzmocnieniem są szczególnie przydatne do problemów, w których nie posiadamy żadnej wiedzy na temat tego, jakie akcje i w jakich stanach wykonywane prowadzą do pożądanego celu, albo wiedza, którą posiadamy, jest niepewna lub niepełna, problemów wymagających adaptacji do zmieniających się warunków oraz problemów, w których występują zakłócenia i niepewność. Typowe dziedziny zastosowań to programy grające w gry, automatyczne sterowanie i robotyka. Obiecujące, choć nie zweryfikowane dotąd praktycznie, wydają się perspektywy zastosowań metod uczenia się ze wzmocnieniem (lub pewnych ich modyfikacji) m.in. do problemów rozpoznawania mowy i wspomagania decyzji ekonomicznych.

Analizowanie i weryfikowanie praktyczne zastosowań omawianych metod w nowych dziedzinach jest jednym ważnym obszarem dla przyszłych prac. Z drugiej strony, w uczeniu się ze wzmocnieniem pozostaje szereg otwartych problemów [2], które wymagają rozwiązania zanim zastosowania do problemów o dużej skali złożoności staną się możliwe. Czyni to z uczenia się ze wzmocnieniem atrakcyjną i obiecującą dziedzinę badań.

References

[1] A.G. Barto, R.S. Sutton i C.J.C.H. Watkins. Learning and sequential decision making. M. Gabriel i J. Moore (eds.), *Learning and Computational Neuroscience*. The MIT Press, 1990.

[2] P. Cichosz. *Reinforcement Learning Algorithms Based on the Methods of Temporal Differences*. Master's thesis, Warsaw University of Technology, Institute of Computer Science, 1994.

[3] S. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, New York, 1983.

[4] R.S. Sutton. *Temporal Credit Assignment in Reinforcement Learning*. PhD thesis, University of Massachusetts, Department of Computer and Information Science, 1984.

[5] R.S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

[6] C.J.C.H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, 1989.

[7] C.J.C.H. Watkins i P. Dayan. Technical note: Q-learning. *Machine Learning*, 8:279–292, 1992.

ISBN 83-85847-85-5

W celu uzyskania bliższych informacji i zakupu dodatkowych egzemplarzy prosimy o kontakt
z Instytutem Badań Systemowych PAN
ul. Newelska 6, 01-447 Warszawa
tel. 36-19-01 w. 241 e-mail: kotuszew@ibspan.waw.pl