Učenie posilňovaním a interná motivácia

Matej Pecháč

Centrum pre kognitívnu vedu: KAI FMFI, Univerzita Komenského v Bratislave
Mlynská dolina, 84248 Bratislava
Email: matej.pechac@gmail.com

Abstrakt

Učenie posilňovaním s využitím hlbokých neurónových sietí ukázalo, že poskytuje užitočný nástroj pre učenie agentov v spojitom aj diskrétnom prostredí. Avšak agenty sú stále limitované v učení sa úloh, ktoré im musí zadať dizajnér daného experimentu. Nedisponujú mechanizmami na generovanie vlastných cieľov a úloh, ktorými by mohli zväčšovať svoj repertoár znalosti o prostredí. Je možné sa inšpirovať psychologickými štúdiami o ľudskom vývine, obzvlášť o aktívnom učení motivovanom zvedavosťou, ktoré prekonáva typické prístupy, kde cieľom je zvládnuť iba jednu úlohu. Učenie založené na internej motivácii môže rozšíriť učenie posilňovaním zaujímavým spôsobom, ktorý umožní agentom napodobňovať inteligentné chovanie s neustálym rozvojom. V článku uvedieme niekoľko spôsobov ako pristupovať k internej motivácii v tomto kontexte.

1 Úvod

Jedným z hlavných problémov pri učení posilňovaním je efektívne preskúmanie prostredia a nájdenie takých stavov alebo udalostí, ktoré agentovi poskytnú najvyššiu odmenu. Skúmanie prostredia zabezpečujú rôzne techniky (ϵ-greedy, Bolzmannova metóda, "stíhacia"metóda), ktoré však viac, či menej náhodne vyberajú akcie, ktorými sa dostáva agent do nových stavov. Pre komplexné prostredie s veľkým počtom stavov sú však časovo aj výpočtovo náročné a v praxi neefektívne. Prístupy internej motivácie ponúkajú sadu metód, ktoré dokážu zefektívniť skúmanie prostredia a nasmerovať agenta do málo navštívených či neznámych častí stavového priestoru.

2 Učenie posilňovaním

Učenie posilňovaním (Sutton a Barto, 1998) je oblasť strojového učenia zameraná na učenie agenta pomocou interakcie s prostredím, v ktorom sa nachádza. Agent vyberá akcie zo svojho repertoára, vykonáva ich a následne pozoruje nový stav. Prostredie, v ktorom agent pôsobí, je často formalizované ako Markovov rozhodovací proces. Ten definuje množinu stavového priestoru $\mathcal{S}$, priestoru akcií $\mathcal{A}$, prechodovú funkciu medzi stavmi $\mathcal{T}$ a funkciu odmeny $\mathcal{R}$ a faktor zľavy γ. Zároveň s novým stavom dostáva agent z prostredia aj odmenu r. Hlavným cieľom agenta je nájsť také pravidlá π, ktoré maximalizujú očakávanú odmenu $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$. Na riešenie tohto problému bolo skonštruovaných niekoľko algoritmov, kde medzi najznámejšie určite patrí Temporal difference (TD) algoritmus (Sutton a Barto, 1998), SARSA (Rummery a Niranjan, 1994), či Q-learning (Watkins a Dayan, 1992). V súčasnej dobe sa často používajú Deep Deterministic Policy Gradient (DDPG), Asynchronous Advantage Actor-Critic algoritmus (A3C) , Trust Region Policy Optimization (TRPO), Proximal Policy Optimization (PPO) , Continuous Actor-Critic Learning Automaton (CACLA) a iné v závislosti od skúmanej domény. Referencie kvôli rozsahu neuvádzame.

3 Interná motivácia

Motivácia predstavuje komplexný psychologický fenomén, do ktorého spadá prekvapenie, zaznamenanie niečoho nového, nesúlad či výzva. Preto ju popisuje veľké množstvo teórií, z ktorých niektoré spomenieme. Teória potrieb je založená na tvrdení, že ľudia hľadajú možnosti ako zabezpečiť svoje potreby, skúmať svoje prostredie a hľadajú spôsoby kontrolovať ho. Teória kognitívnej disonancie vysvetľuje motiváciu ako redukciu rozdielu medzi nadobudnutou skúsenosťou a očakávaným výsledkom, ktorý je výstupom našich vnútorných kognitívnych štruktúr. Teória "flow" pripisuje najväčšiu motiváciu riešeniu problémov, ktoré majú optimálnou obtiažnosť vzhľadom na schopnosti človeka. Referencie kvôli rozsahu opäť neuvádzame. Poznatkami týchto teórii sa je možné inšpirovať a využiť ich pri zavedení motivácie u umelých agentov. Pri formalizácii motivácie je možné ju rozdeliť na externú r^{extr} a internú r^{intr}. Externá motivácia má zdroj mimo agenta, čiže odmena prichádza z prostredia a je vždy viazaná na špecifický cieľ v danom prostredí. Interná motivácia je generovaná priamo v štruktúrach agenta na základe nejakej udalosti napr. pozorovaného stavu. Signál z modulu, ktorý modeluje internú motiváciu, sa pridáva k odmene z vonkajšieho prostredia a je modifikovaný parametrom β. Odmena r_t, ktorú agent dostane po vykonaní akcie v čase t môže mať potom tvar

$r_t = (1-\beta)r_t^{\text{extr}} + \beta r_t^{\text{intr}}$. Na základe skoršej práce Oudeyer a Kaplan (2009) sa dajú prístupy k internej motivácii rozdeliť do troch kategórií: *znalostné prístupy*, *kompetenčné prístupy* a *morfologické prístupy*. Každý z nich stručne charakterizujeme v nasledujúcich podkapitolách.

3.1 Znalostné prístupy

Tieto metódy sa zameriavajú na rozširovanie agentovej znalosti o prostredí a odmeňujú také stavy, ktoré agent nepredpovedal svojimi vnútornými štruktúrami modelujúcimi dynamiku prostredia. To vedie agenta k skúmaniu prostredia a tvorbe jeho čo najpresnejšieho modelu. Metóda založená na *neistote* generuje väčšiu motiváciu pre stav s_t s nízkou pravdepodobnosťou pozorovania $p(s_t)$ a môžeme ju definovať ako $r_t^{\text{intr}} = C \cdot (1 - p(s_t))$, kde C je kalibračná konštanta. Ďalšia metóda je založená na *informačnom zisku* a odmeňuje pokles neistoty vyjadrenej entropiou toho istého stavu v dvoch časových okamihoch $r_t^{\text{intr}} = C \cdot (H(s_t, t-1) - H(s_t, t))$ Metóda založená na odmene *podobnosti* naopak motivuje agenta k návšteve stavov s vysokou pravdepodobnosťou pozorovania $r_t^{\text{intr}} = C \cdot p(s_t)$. Úspešne použité boli metódy motivácie podľa počtu návštev stavu s použitím tzv. "pseudo počtu"(Ostrovski a spol., 2017), ktoré je možné aplikovať na spojité či komplexné prostredia a predikčné metódy (napr. Pathak a spol. (2017)), ktoré definujú motiváciu ako rozdiel medzi predpoveďou vnútorného modelu a pozorovaním $r_t^{\text{intr}} = C \cdot \|\hat{s}_{t+1} - s_{t+1}\|_2^2$.

3.2 Kompetenčné prístupy

Tieto prístupy používajú mieru kompetencie agenta dosiahnuť cieľ, ktorý si sám vygeneruje a vedú agenta k osvojeniu sady zručností. Mieru kompetencie možno formálne definovať ako $l_a(g_k, t_g) = \|\tilde{g}_k(t_g) - g_k(t_g)\|$ kde $g_k(t_g)$ je dosiahnutý a $\tilde{g}_k(t_g)$ vygenerovaný cieľ v epoche t_g. Na základe nej možno motiváciu založiť na *maximalizácii nekompetencie*, potom je agent motivovaný k učeniu úloh, ktoré mu idú najhoršie. Formálne ju možno zapísať ako $r_t^{\text{intr}} = C \cdot l_a(g_k, t_g)$. Pri *maximalizácii postupu v kompetencii* vychádza motivácia zo zmeny miery kompetencie agenta za časové obdobie θ: $r_t^{\text{intr}} = C \cdot [l_a(g_k, t_g - \theta) - l_a(g_k, t_g)]$ Podrobné empirické porovnanie rôznych metód možno nájsť v práci Santucci a spol. (2013).

3.3 Morfologické prístupy

Kategória morfologických metód je založená na vlastnostiach pozorovaného stavu prostredia. Agent v tomto prípade nemá žiaden špeciálny modul. Cieľom takto motivovaného agenta môže byť pozorovanie stavov s určitými vlastnosťami ako je *stabilita* či *variácia*. Na základe toho možno motivovať agenta k vykonávaniu takých akcií, ktoré vedú k malým zmenám v pozorovaných stavoch za posledných T krokov, či naopak, vedú k čo najväčším zmenám $r_t^{\text{intr}} = C \cdot \|s_t - \langle s \rangle_T\|$.

4 Diskusia

Ponúkli sme krátky prehľad rôznych prístupov a ich metód modelovania internej motivácie, ktoré sú vhodné pre učenie posilňovaním. Je možné nimi dosiahnuť efektívne skúmanie prostredia a osvojovanie si nových zručností, čo môže u agenta zabezpečiť neustály rozvoj a kontinuálne učenie bez potreby zásahu dizajnéra. Myslíme si, že vhodnou cestou ďalšieho rozvoja môže byť vytvorenie komplexnejších modulov, ktoré by zahŕňali niekoľko submodulov generujúcich internú motiváciu na základe rôznych metód, čím by sa správanie agenta mohlo priblížiť ku správaniu človeka, ktorý je taktiež motivovaný rôznymi podnetmi v závislosti od kontextu.

Poďakovanie

Tento príspevok bol podporený grantovou agentúrou VEGA v rámci projektu 1/0796/18.

Literatúra

Ostrovski, G., Bellemare, M. G., van den Oord, A. a Munos, R. (2017). Count-based exploration with neural density models. V *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, str. 2721–2730. JMLR. org.

Oudeyer, P.-Y. a Kaplan, F. (2009). What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurorobotics*, 1:6.

Pathak, D., Agrawal, P., Efros, A. A. a Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. *CoRR*, abs/1705.05363.

Rummery, G. A. a Niranjan, M. (1994). *On-line Q-learning using connectionist systems*, vol. 37. University of Cambridge, Department of Engineering Cambridge, England.

Santucci, V. G., Baldassarre, G. a Mirolli, M. (2013). Which is the best intrinsic motivation signal for learning multiple skills? *Frontiers in Neurorobotics*, 7:22.

Sutton, R. S. a Barto, A. G. (1998). *Introduction to Reinforcement Learning*, vol. 135. MIT press Cambridge.

Watkins, C. J. a Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3-4):279–292.