Posilované učení a jeho uplatnění v praxi vzdělávání a rozvoje managementu

Horníková Eva, Matěnová Petra, Matouš Petr

Klíčová slova: cca 5 klíčových slov oddělených čárkou

Abstrakt 100-150 slov

Tato práce se zabývá posilovaným učením, které má dvě základní charakteristiky – je založeno na souboru pokusů a omylů a jedinec může získat odměnu až po provedení určité akce. První část práce se zabývá teorií, kde jsou vymezeny základní pojmy, jež jsou nezbytné pro pochopení následujících částí a zároveň je zde nastíněn také vznik tohoto nástroje. V následující kapitole – výzkumné metody a data – je uvedeno využití daného učení v současné praxi. V neposlední řadě se poté práce zabývá podnikovou sférou a využití tohoto typu učení v organizacích a doporučeními autorů pro užití tohoto nástroje.

Úvod 400-600 slov

VO (např.): Prostřednictvím jakých nástrojů jsou uplatňovány principy posilovaného učení při vzdělávání a rozvoji vedoucích pracovníků?

1. Teoretická východiska

Co je posilované učení?

Dle Akhtara (2017) je posilované učení takový druh učení, jež určuje akci jedince v konkrétním okamžiku s cílem maximalizovat jeho odměnu. Jednou z charakteristických vlastností posilovaného učení je, že jedinec může získat odměnu až po provedení akce. Proto musí pokračovat v komunikaci s prostředím za účelem stanovení si optimálního chování pomocí pokusů a omylů. Jako příklad Akhtar (ibid.) uvádí situaci, kdy se dítě učí chodit. Dítě se postaví, ujde pár kroků a za chvíli spadne. A to se opakuje stále dokola a není zde nikdo, kdo by jej učil, jak se to dělá. Je to učení pomocí pokusu a omylu. V životě je mnoho situací, kdy lidé nemají k dispozici detailní instrukce, jak vykonat úkol, nýbrž vykonaný úkol hodnotí a snaží se zlepšovat své chování.

Princip

V Markovově rozhodovacím procesu je jedinec, který má odpovědnost za zvolenou akci. Zároveň komunikuje s prostředím, do kterého je vložen. Jedinec je povinen neustále monitorovat prostředí, jelikož předešlé vykonané akce mohly změnit jeho stav. Poté je udělena odměna vzhledem k poslední vykonané akci. Přechody mezi stavy, akcemi a následnými odměnami se odehrávají sekvenčně a cyklicky. Během procesu má jedinec za úkol maximalizovat získané odměny, jež získá za zvolené akce. Z těchto akcí následně získá strategii pro řešení systému. Čím je strategie lepší, tím rychleji jedinec dosáhne cíle. Lze tedy říct, že posilované učení je naučení jedince vykonávat optimální strategii (Xie, 2018).

Pozitivní posilované učení

Je to učení s dostáním pozitivní odměny, což znamená něco žádoucího, co jedinec dostane po vykonání akce. Akhtar (2017, strana) uvádí příklad, že po pilném učením se žák dostane na místo premianta ve třídě. Poté, co žák zjistí, že jeho učení a dobré výsledky vedly k pozitivní odměně, bude se snažit pokračovat stejným způsobem.

Negativní posilované učení

Učení znamenající dostání negativní odměny nebo něco nežádoucího za vykonání akce. Akhtar (2017, str.) uvádí příklad, když jde jedinec do kina sledovat film a je mu zde velmi zima. Pokračování ve sledování filmu je tedy nekomfortní. Příště jde jedinec do kina znovu a je mu znovu zima. Potřetí, když jde jedinec do kina, obleče si bundu. Touto akcí způsobí, že je negativní element odstraněn. A zde lze vidět, že jedinec “upadl” a zase se “postavil na nohy”.

Historie vzniku

Podle Suttona a Barto (2018, str.) lze historii vzniku posilovaného učení rozdělit na dvě části – první se týká učení na principu pokus a omyl a druhá se snaží najít optimální řízení. Základy tohoto nástroje byly položeny psychologem C. Lloyd Morganem v první polovině minulého století, který se zabýval studií zvířat a zveřejnil metodu pokus-omyl. Následně pak byla zveřejněna Bellmanova rovnice, jež umožnila výpočet optimálního řízení a zároveň o několik let později uvedl Markovův rozhodovací proces (pojmenovaný po ruském matematikovi Andrey Markov). Tento proces se týká sekvence událostí, kde každá závisí na té předchozí. Na zmínku o výhodách metod pokus-omyl, kterou zveřejnil H. Klopf, poté navázali Richard Sutton a Andrew Barto.

Největší průlom v oblasti posilovaného učení zajistil Watkinst (1989 str.), který propojil oblasti metod pokus-omyl a dynamické programování. Dalších úspěchů bylo dosaženo až v roce 2013, kdy londýnská firma DeepMind Techologies vytvořila algoritmus, jež je schopen naučit se hrát hry s využitím neuronové sítě a posilovaného učení. O měsíc později byla firma odkoupena firmou Google a byl

vytvořen Alpha Zero, jež dokázal porazit všechny šampiony ve hře Go. V roce 2018 byla univerzitou v Berkeley vydána studie o užití posilovaného učení pro replikování složitých pohybů člověka. Mnoho vědců tvrdí, že posilované učení je jedním z nejslibnějších nástrojů, které povedou k vytvoření univerzálního inteligentního robotického chování. (Sutton a Barto, 2018)

2. Výzkumné metody a data 300-700

Využití

Samořídící auto

Společnosti jako Toyota a Ford investovaly miliony dolarů do oddělení výzkumu a vývoje v souvislosti s technologií samořídících aut. Služby jako Uber a Lyft v současné době platí lidské řidiče, ale v budoucnu mohou nasadit celou řadu samořídících aut. (Akhtar, 2017) Auta se nepřetržitě během tréninku učí opravovat své jízdní schopnosti metodou pokus-omyl.

Autonomní letecké taxi

Mezitím co se neustále hovoří o samořídících autech, Spojené Arabské Emiráty připravují vypuštění autonomního leteckého taxi. Vzdálenost doletu je asi 30 mil a jeho nosnost je asi 100 kg. Pasažéři mohou k rezervaci letu z určené zóny využít mobilní aplikaci. Ovládání pak probíhá pomocí dotykové obrazovky uvnitř taxi. (Akhtar, 2017)

Autonomní akrobatický vrtulník

Počítačoví experti ze Stanfordu úspěšně vytvořili AI systém, který umožňuje robotický vrtulník učit se provádět obtížné kaskadérské kousky při sledování ostatních vrtulníků při těchto manévrech. Díky tomu vznikla autonomní helikoptéra, která umí samostatně předvést kompletní leteckou show. Autonomní let je obecně považován za velmi náročný kontrolní problém. Díky užití posilovaného učení pro optimalizaci problému jsou optimalizovány modelové a odměnové funkce. (Akhtar, 2017)

Problémy v posilovaném učení

Jedním z problémů posilovaného učení je jeho užití pouze v diskrétních prostorech s diskrétními akcemi. První problém se dá řešit diskretizací jednotlivých stavů (vytvoření intervalů). Druhý problém spojitých akcí lze řešit obdobně. Problémem může být i přílišná velikost stavových prostorů, což poté vede k přílišné velikosti matic, které jsou nezbytné pro tvorbu algoritmu. Ten se poté tedy učí velmi pomalu, nebo vůbec.

Kontext

Dříve se firmy snažily získat konkurenční výhodu zejména pomocí nových výrobních technologií. Nyní tuto výhodu získávají mimo jiné pomocí lidského kapitálu. Ten označuje zkušenosti, znalosti, pokročilé dovednosti, kreativitu a

motivaci jednotlivých zaměstnanců. Tak, aby dodávali vysoce kvalitní produkty a služby. Každý člověk je jiný, a to platí i o zaměstnancích, jejich zkušenosti, znalosti a dovednosti jsou různé a vzájemně se liší jeden od druhého. Zaměstnavatel by proto měl dbát na rozvoj těchto jedinečných schopností a tím zlepšit schopnosti celého lidského kapitálu v organizaci. Může tak učinit prostřednictvím vzdělávacích a rozvojových programů.

Teorie posilovaného učení popisuje fakt, že lidé jsou motivování k určitému chování (nebo k jeho vyhnutí) na základě minulých zkušeností. Takové chování, které je posilováno má tendenci se opakovat. Stejně to funguje i obráceně, chování, které posilováno není, bývá eliminováno.

Jednou z možností, jak využít posilované učení v podnikovém prostředí je jeho zapojení do managementu. Je důležité, aby management pracoval na efektivní zpětné vazbě. Taková zpětná vazba by měla mít pozitivní dopad na zaměstnance a jejich pracovní efektivitu. Není vhodné, aby manager pouze káral své zaměstnance, ale měl by jim poskytnout takové hodnocení, které je posune dál ve vývoji jejich schopností.

Jak již bylo zmíněno dříve posilované učení se dělí na pozitivní a negativní. Při pozitivním posilovaném učení se posiluje dané chování zaměstnance. Když bude zaměstnanec vědět, že za určité chování bude od zaměstnavatele odměněn, bude motivován k jeho opakování. Příkladem odměn mohou být bonusy či zvyšování mzdy, kariérní růst nebo například získání osvědčení po splnění určitého vzdělávacího programu. Takové odměny představují pozitivní dopad na chování jedince. Když bude mít zaměstnanec zadaný úkol a po jeho naplnění bude následovat povýšení, tak se bude aktivně snažit o jeho splnění, jelikož bude pozitivně motivován. I když se to z názvu nemusí zdát, tak i negativní posilované učení posiluje chování. A to tím způsobem, že eliminuje nepříjemné zkušenosti zaměstnance. Zaměstnanec je motivován k podávání dobrého výkonu, protože si je vědom faktu, že když úspěšně nesplní zadaný úkol, nebude zvažováno jeho povýšení. Snaží se tedy eliminovat negativní zkušenost, kterou v tomto příkladu představuje nepovýšení.

Z výše popsaného je patrné, že teorii posilovaného učení užívají zejména vedoucí pracovníci a zaměstnanci. Vedoucí pracovníci poskytují pozitivní zpětnou vazbu a s tím spojené pozitivní posilované učení. Naopak zaměstnanci sami na sobě praktikují negativní posilované učení, kdy se snaží omezit negativní zkušenosti plynoucí z jejich chování.

S posilovaným učením může být spojen problém v podobě pochopení zpětné vazby. Může se stát, že i když zkušený manažer pro slovní zpětnou vazbu použije "Sandwich metodu", tak si z ní zaměstnanec odnese pouze tu negativní část. Taková zpětná vazba poté nemusí mít tížený výsledek. Naopak může na zaměstnance působit demotivujícím způsobem. V nejhorším případě pak může zaměstnanec ztratit motivaci k podávání výkonu, jelikož se cítí nedostatečně ohodnocen. Další problém může představovat situace, kdy jsou na sebe

zaměstnanci při negativním posilovaném učení příliš tvrdí. Snaží se eliminovat příliš mnoho způsobů chování, které ve výsledku nemají až takový vliv na splnění úkolu. V takovém případě může přijít i vyhoření, kdy si sami zaměstnanci budou připadat nedostateční pro danou práci. Budou mít pocit, že dělají vše špatně a že svůj pracovní výkon nezvládají. V obou zmíněných problémech by měl zasáhnout vedoucí pracovník a opět podpořit důvěru pracovníka v sebe samotného a jeho schopnosti.

3. Výsledky 800-1400

- Existuje dostatečná vědecká podpora pro daný nástroj?
- Jaká jsou vaše doporučení stran aplikování daného nástroje?

4. Diskuze 300-700 (interpretace výsledků, náměty pro další výzkum)

Závěr 150-200

Literatura

Akhtar, D. E. S. M. F. (2017). Practical Reinforcement Learning. Packt Publishing.

Najib Akraml Ruicong Xie. Markov Decision Processes (MDPs) - Structuring a Reinforcement Learning Problem, září 2018. (Citováno dne: 12. 11. 2020), Dostupné z: http: //deeplizard.com/learn/video/my207WNoeyA.

Sutton, R. S., Barto, A. G. (2018). Reinforcement Learning: An Introduction. The MIT Press.

Pilát, M. Zpětnovazebné učení. Retrieved November 19, 2020, from https://martinpilat.com/cs/prirodou-inspirovane-algoritmy/zpetnovazebni-uceni