https://doi.org/10.37528/FTTE/9788673954455/POSTEL.2021.026

PRIMENA UČENJA POTKREPLJIVANJEM U PROTOKOLIMA RUTIRANJA ZA DINAMIČKE BEŽIČNE *AD HOC* MREŽE

Nenad Jevtić, Marija Malnar, Pavle Bugarčić
Univerzitet u Beogradu - Saobraćajni fakultet
n.jevtic@sf.bg.ac.rs, m.malnar@sf.bg.ac.rs, p.bugarcic@sf.bg.ac.rs

Rezime: *Rutiranje paketa podataka u dinamičkim bežičnim ad hoc mrežama veoma je zahtevan proces zbog brzih promena u topologiji mreže, što može prouzrokovati česte prekide veza između čvorova, a samim tim i veliki stepen izgubljenih paketa. Jedan od načina da se ovaj problem prevaziđe jeste primena veštačke inteligencije u protokolima rutiranja, kako bi se proces rutiranja prilagodio dinamičkoj prirodi ovih mreža. Veoma značajna oblast veštačke inteligencije koja se, u poslednje vreme, sve više primenjuje u dinamičkim bežičnim ad hoc mrežama je mašinsko učenje (machine learning), a posebno se ističe tip mašinskog učenja pod nazivom učenje potkrepljivanjem (reinforcement learning). U ovom radu predstavljen je pregled aktuelnih rezultata u primeni učenja potkrepljivanjem u protokolima rutiranja za različite dinamičke bežične ad hoc mreže. Protokoli rutiranja najpre su klasifikovani prema tipu mreže na protokole za VANET (Vehicular Ad-hoc Networks) i FANET (Flying Ad-hoc Networks) mreže, a zatim i prema nekim drugim karakteristikama i specifičnostima.*

Ključne reči: *učenje potkrepljivanjem, rutiranje, VANET, FANET*

1. Uvod

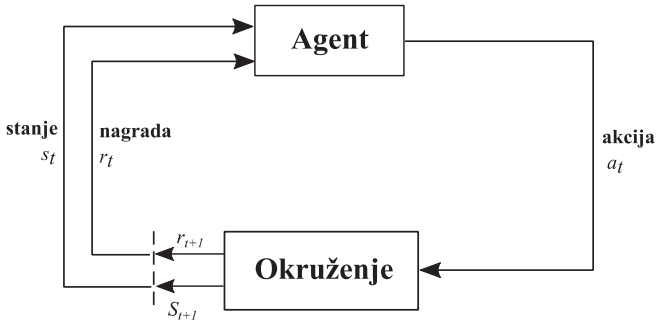
Bežične *ad hoc* mreže (*Wireless Ad hoc Networks*, WANET) predstavljaju mreže koje nemaju fiksnu infrastrukturu, kao što su ruteri ili pristupne tačke, već svaki čvor učestvuje u rutiranju podataka. Najznačajnije kategorije dinamičkih WANET mreža su mobilne *ad hoc* mreže (*Mobile Ahoc Networks*, MANET), *ad hoc* mreže za vozila (*Vehicular Ad hoc Networks*, VANET) i *ad hoc* mreže za letelice (*Flying Ad hoc Networks*, FANET). Proces izbora optimalne putanje kod ovih mreža veoma je kompleksan, zbog stalnih promena u topologiji mreže. To je posebno izraženo kod visoko dinamičkih mreža, kao što su VANET i FANET. Uključivanjem veštačke inteligencije u proces odabira optimalne putanje, moguće je prevazići ovaj problem. Jedna od najznačajnijih oblasti veštačke inteligencije koja se primenjuje u protokolima rutiranja za dinamičke WANET mreže je mašinsko učenje (*machine learning*, ML), a posebno se ističe tip mašinskog učenja pod nazivom učenje potkrepljivanjem (*reinforcement learning*, RL). Ovaj tip mašinskog učenja prati promene u mreži kroz stalnu interakciju sa okruženjem i u skladu

sa aktuelnim stanjem mreže pomaže u izboru optimalne putanje do odredišta, pa je veoma pogodan za mreže kod kojih se topologija često menja.

Rad je organizovan na sledeći način. U drugom poglavlju predstavljeni su osnovni principi učenja potkrepljivanjem, a ukratko su opisani i podtipovi ove vrste mašinskog učenja, koji se najčešće koriste u protokolima rutiranja kod dinamičkih WANET mreža. U trećem poglavlju je dat pregled aktuelnih radova u kojima je primenjeno učenje potkrepljivanjem za unapređenje protokola rutiranja u VANET i FANET mrežama. U četvrtom poglavlju su data zaključna razmatranja.

2. Osnovni principi učenja potkrepljivanjem

Učenje potkrepljivanjem predstavlja najzastupljeniji tip mašinskog učenja u protokolima rutiranja za dinamičke bežične *ad hoc* mreže. Ovaj tip učenja je opisan u [1] i podrazumeva učenje kroz stalnu interakciju sa okruženjem radi postizanja određenog cilja. Kako bi se opisao ovaj proces, neophodno je najpre definisati najznačajnije elemente u procesu učenja. Donosilac odluka u procesu učenja potkrepljivanjem naziva se agent. Sve ono što okružuje agenta i sa čim on vrši interkonekciju, naziva se okruženje. U svakom diskretnom trenutku t, okruženje može da se nađe u određenom stanju s_t, koje pripada ograničenom skupu mogućih stanja S. Agent donosi odluku o preduzimanju određene akcije a_t, iz ograničenog skupa akcija A, koje su dostupne agentu u stanju s_t, a okruženje odgovara na preduzetu akciju povratnom informacijom ka agentu. Ta povratna informacija sadrži novo stanje u kom se našlo okruženje s_{t+1}, kao i numeričku nagradu za preduzetu akciju r_{t+1}. Na ovaj način okruženje potkrepljuje agenta znanjem o korisnosti akcija koje preduzima. Agent tokom vremena pokušava da maksimizira nagradu kroz optimizaciju izbora mogućih akcija. Šematski prikaz interakcije agenta sa okruženjem prikazan je na slici 1.



Slika 1. Interakcija agenta i okruženja u procesu učenja potkrepljivanjem

Ukoliko se posmatra jedna bežična *ad hoc* mreža, proces učenja potkrepljivanjem može da se modeluje na sledeći način. Svaki čvor u mreži koji šalje pakete podataka predstavlja agenta učenja, dok celokupna mreža predstavlja okruženje. Slanje paketa podataka prema jednom od susednih čvorova predstavlja potencijalnu akciju koju agent može da preduzme. S obzirom da svaki čvor ima ograničen skup susednih čvorova kojima može da pošalje pakete, ovaj skup suseda predstavlja skup mogućih akcija koje čvor može da preduzme. Povratne informacije koje dobija čvor koji šalje podatke predstavljaju

nagradu za preduzetu akciju. Nagrada može da zavisi od brojnih uticajnih faktora, a neki od primera uticajnih faktora koji utiču na nagradu navedeni su u narednom poglavlju.

Jedan od najjednostavnijih algoritama učenja potkrepljivanjem je *Q-Learning* (QL) [2]. Kod ovog algoritma svaki agent održava tabelu Q-vrednosti, koje se odnose na korisnost preduzimanja određene akcije u određenom stanju, na osnovu kojih donosi odluke o budućim akcijama. Svaki element ove tabele računa se na osnovu sledeće formule:

$$Q(s_t, a_t) \leftarrow (1 - \alpha) * Q(s_t, a_t) + \alpha * (R + \gamma * max_a(Q(s_t, a_t), a) \qquad (1)$$

U prethodnoj formuli R predstavlja nagradu za preduzetu akciju u odgovarajućem stanju, α predstavlja stopu učenja koja utiče na brzinu učenja i može uzeti vrednost u opsegu [0,1], γ predstavlja diskontni faktor koji određuje važnost budućih nagrada i takođe može uzeti vrednost u opsegu [0,1], dok je $max_a(Q(s_t, a_t), a)$ maksimalna moguća Q-vrednost koju agent može ostvariti preduzimanjem akcije a, iz skupa mogućih akcija A, u stanju s_t. Agent mora da napravi ravnotežu između eksploatacije stečenog znanja i istraživanja okruženja, koje je neophodno kako bi agent ažurirao svoje znanje na osnovu promena stanja okruženja. Tako će agent sa određenom verovatnoćom ε preduzimati akciju sa najvećom Q-vrednošću, dok će sa verovatnoćom (1- ε) preduzimati slučajno izabranu akciju iz skupa mogućih akcija. Ovo se naziva ε-pohlepna (*ε-greedy*) politika, koja se uglavnom koristi kod QL algoritma.

Kako bi se poboljšao proces učenja, u radu [3] je uveden koncept dubokog učenja potkrepljivanjem (*Deep Reinforcement Learning*, DRL). Kako bi se aproksimirale optimalne Q-vrednosti, ovaj koncept koristi duboku neuronsku mrežu (*Deep Q-Network*, DQN). Učenje potkrepljivanjem je često nestabilno, ili čak divergira, kada se koristi nelinearni aproksimator funkcija (*nonlinear function approximator*), kao što je neuronska mreža, za određivanje Q-vrednosti. Kako bi se prevazišle ove nestabilnosti, uvedene su dve nove ideje. Prvo, uvodi se biološki inspirisan mehanizam nazvan ponavljanje iskustva (*experience replay*). Ovim mehanizmom se u skup $D = \{e_1, ..., e_t\}$ skladište podaci $e_t = (s_t, a_t, r_t, s_{t+1})$ koje je prikupio agent u vremenskom trenutku t. Tokom učenja, ažuriraju se vrednosti nasumično izabranog elemnta e_i iz skupa D, čime se smanjuju korelacije među podacima i poboljšavaju performanse u poređenju sa prethodnim algoritmima učenja potkrepljivanjem. Druga ideja je uvođenje ciljne Q-mreže, odnosno prilagođavanje Q-vrednosti prema ciljnim vrednostima koje se samo periodično ažuriraju, čime se smanjuje korelacija sa ciljnim vrednostima. Drugim rečima, trenutne Q-vrednosti se kopiraju u ciljnu Q-mrežu na svakih C koraka u procesu treninga, čime se povećava stabilnost učenja.

Kako bi se dalje poboljšale performanse i povećala stabilnost učenja potkrepljivanjem, u [4] je predložen koncept duelnog dubokog učenja potkrepljivanjem (*Dueling Deep Reinforcement Learning*, DDQL). Ovaj koncept podrazumeva korišćenje duelnih dubokih neuronskih mreža (*Dueling Deep Q-Network*, DDQN) za određivanje optimalnih Q-vrednosti. Osnovna ideja DDQN mreža je da nije potrebno uvek računati vrednosti preduzimanja svake dostupne akcije. Zbog toga se DDQN mrežna arhitektura može podeliti na dve glavne komponente: funkciju vrednosti (*value function*) i funkciju prednosti (*advantage function*). Funkcija vrednosti treba da predstavi koliko je dobro biti u određenom stanju, a funkcija prednosti meri relativnu važnost određene akcije u poređenju sa ostalim akcijama. Nakon posebnog proračuna ove dve funkcije, rezultati ovih fukcija se kombinuju kako bi se dobila konačna Q-vrednost.

Još jedan tip učenja potkrepljivanjem koji se koristi u protokolima rutiranja za dinamičke bežične *ad hoc* mreže je SARSA- λ [5]. Karakteristično za ovaj algoritam je uvođenje koeficijenta slabljenja (λ) u proračun ciljnih Q-vrednosti, čime se povećava efikasnost njihovog ažuriranja.

3. Protokoli rutiranja za VANET i FANET mreže bazirani na učenju potkrepljivanjem

U ovom poglavlju dat je pregled aktuelnih radova u kojima je primenjeno učenje potkrepljivanjem za unapređenje protokola rutiranja u dinamičkim WANET mrežama. Fokus je stavljen na radove u periodu od 2018-2021. godine, kako bi se obuhvatio skup najaktuelnijih istraživanja iz ove oblasti. U tabeli 1 je prikazana kategorizacija protokola rutiranja na osnovu tipa mreže u kojoj se predloženi protokol primenjuje, zatim na osnovu tipa učenja potkrepljivanjem koje je primenjeno u protokolu i najzad na osnovu eventualne kombinacije učenja potkrepljivanjem sa nekom drugom tehnikom u odgovarajućem protokolu rutiranja.

Kada je reč o primeni učenja potkrepljivanjem u dinamičkim WANET mrežama, duži niz godina su objavljivani radovi koji se odnose na primenu učenja potkrepljivanjem isključivo u MANET mrežama. Zatim su se u prethodnoj deceniji pojavila istraživanja u kojima su predloženi protokoli rutiranja na bazi učenja potkrepljivanjem za VANET mreže, a tek su se u poslednjih nekoliko godina u literaturi pojavili radovi koji opisuju protokole rutiranja na bazi učenja potkrepljivanjem za FANET mreže. Upravo je velika ekspanzija protokola za VANET i FANET mreže razlog zbog kog je osnovni fokus ovog istraživanja upravo na protokolima rutiranja u tim mrežama. Zbog toga su protokoli klasifikovani prema tipu mreže na one koji se primenjuju u VANET i one koji se primenjuju u FANET mrežama. Iako je još uvek prisutan veći broj protokola koji se odnose na VANET mreže, sa sve bržim razvojem bespilotnih letelica iz godine u godinu raste broj protokola koji se odnose na FANET mreže.

Tip učenja potkrepljivanjem koji se najčešće primenjuje u navedenim protokolima je QL, koji predstavlja najjednostavniji tip učenja potkrepljivanjem. U nekoliko radova primenjen je DRL tip, koji predstavlja nešto složeniji oblik učenja potkrepljivanjem, dok je par istaživača primenilo unapređeni oblik DRL pod nazivom DDRL. U jednom radu je primenjen SARSA-λ algoritam, koji se razlikuje od QL algoritma u načinu računanja Q-vrednosti, a može se reći da takođe spada u jednostavnije tipove učenja potkrepljivanjem. U određenim protokolima se pored učenja potkrepljivanjem koriste i druge tehnike kao što su *blockchain* i *fuzzy logic*, dok kod nekih protokola ulogu agenta koji donosi odluke u procesu učenja potkrepljivanjem ima SDN (*Software-Defined Networking*) kontroler.

U tabeli 2a prikazana je komparativna analiza protokola rutiranja baziranih na učenju potkrepljivanjem u VANET mrežama, dok je u tabeli 2b predstavljena ista analiza za protokole u FANET mrežama. Analiza je sprovedena na osnovu uticajnih faktora koji određuju vrednost funkcije nagrade u procesu učenja potkrepljivanjem, kao i na osnovu posmatranih performansi u simulacionoj analizi predloženih protokola. Takođe, navedeni su i simulatori koji su korišćeni za evaluaciju performansi predloženih protokola rutiranja.

Tabela 1. Kategorizacija protokola rutiranja baziranih na učenju potkrepljivanjem na osnovu tipa mreže, tipa učenja i kombinacije mašinskog učenja sa drugim tehnikama

Ref.	God.	Tip mreže	Tip učenja				Kombinacija sa drugim tehnikama		
			QL	DRL	DDRL	SARSA-λ	SDN	*Blockchain*	*Fuzzy logic*
[6]	2018.	VANET			✓		✓		
[7]	2018.	VANET		✓			✓		
[8]	2020.	VANET				✓			
[9]	2020.	VANET	✓				✓		
[10]	2019.	VANET			✓		✓	✓	
[11]	2018.	VANET	✓						✓
[12]	2018.	VANET	✓						
[13]	2019.	VANET	✓						
[14]	2018.	VANET	✓					✓	
[15]	2018.	VANET	✓						
[16]	2018.	VANET	✓						
[17]	2021.	VANET		✓					
[18]	2020.	VANET	✓						
[19]	2018.	VANET	✓						
[20]	2019.	VANET		✓			✓		
[21]	2019.	VANET		✓					
[22]	2020.	VANET	✓						
[23]	2021.	VANET		✓					
[24]	2020.	FANET		✓					✓
[25]	2021.	FANET	✓						
[26]	2018.	FANET	✓						
[27]	2020.	FANET	✓						
[28]	2021.	FANET	✓						
[29]	2021.	FANET	✓						
[30]	2020.	FANET	✓						✓
[31]	2020.	FANET	✓						
[32]	2020.	FANET	✓						
[33]	2020.	FANET	✓						

Vrednost funkcije nagrade u najvećoj meri utiče na ishod učenja potkrepljivanjem, pa je iz tog razloga neophodno pažljivo odabrati uticajne faktore koji će definisati njenu vrednost. U različitim istraživanjima korišćeni su različiti uticajni faktori, u zavisnosti od osnovnog cilja optimizacije procesa rutiranja. Neki od najčešćih uticajnih faktora su pouzdanost i kvalitet veze ka potencijalnom sledećem čvoru na putanji, pouzdanost potencijalnog sledećeg čvora, broj skokova koji su potrebni da bi paket stigao do odredišta, dostupni propusni opseg veze, ostvareni protok paketa, kašnjenje, itd. Često je veoma bitno da li je sledeći čvor ujedno i odredišni, odnosno da li sledeći čvor zna putanju do odredišnog čvora. Ukoliko je cilj protokola optimizacija potrošnje energije, gubitak energije će biti važan uticajni faktor pri određivanju vrednosti funkcije nagrade. Sa druge strane, ukoliko je naglasak na zaštiti od neželjenih uticaja sa strane, bitni uticajni faktori biće reputacija sledećeg čvora na putanji i detekcija ometača u blizini tog čvora.

Tabela 2a. Komparativna analiza protokola rutiranja baziranih na učenju potkrepljivanjem u VANET mrežama

Ref.	Uticajni faktori na nagradu	Posmatrane performanse	Simulator
[6]	pouzdanost susednog čvora	gubitak paketa, protok, kašnjenje	TnesorFlow, OPNET
[7]	pouzdanost susednog čvora	gubitak paketa, protok, stepen uspešno isporučenih paketa	TnesorFlow, OPNET
[8]	broj skokova, korisnost veze, propusni opseg	stepen uspešno isporučenih paketa, broj skokova	Python
[9]	rastojanje od odredišta, kašnjenje	kašnjenje, protok	NS-3
[10]	protok	gubitak paketa, protok	TnesorFlow, Phyton
[11]	da li je čvor prvi sused, broj skokova, isplativost akcije, kvalitet veze	stepen uspešno isporučenih paketa, broj kolizija MAC okvira, kašnjenje, protok	NS-2
[12]	da li je poruka isporučena *grid*-u sa odredišnim čvorom	stepen uspešno isporučenih paketa, broj skokova, kašnjenje, broj prosleđivanja, protok	nije naglašeno
[13]	tip i destinacija kontrolnih paketa	stepen uspešno isporučenih paketa, vreme od slanja zahteva do prijema odgovora, overhed	NS-3
[14]	reputacija i isplativost akcije odgovarajućeg čvora	stepen uspešno isporučenih paketa, reputacija, korisnost	nije naglašeno
[15]	direktna veza sa odredištem, odnosno broj skokova i proteklo vreme od poslednje konekcije	stepen uspešno isporučenih paketa, kašnjenje	ONE
[16]	broj skokova, pouzdanost veze, propusni opseg	stepen uspešno isporučenih paketa, kašnjenje, prosečna dužina putanje, overhed	NS-2
[17]	razni neželjeni efekti	/	nisu vršene simulacije
[18]	opterećenje relejnog čvora i zagušenje zemaljske mreže	stepen uspešno isporučenih paketa, iskorišćenost mreže, kašnjenje	OPNET
[19]	da li je kontrolni paket stigao od čvora pošiljaoca	stepen uspešno isporučenih paketa, kašnjenje, broj skokova, overhed	QualNet
[20]	broj skokova, kvalitet veze	protok, broj gejtvej čvorova	nije naglašeno
[21]	uspešnost slanja paketa	stepen uspešno isporučenih paketa, kašnjenje, overhed	NS-2
[22]	kvalitet veze, vreme isteka trajanja veze, kašnjenje	stepen uspešno isporučenih paketa, kašnjenje	QualNet
[23]	gubitak energije, brzina prenosa	potrošnja energije, gubitak paketa, ukupno vreme prenosa, verovatnoća prekida komunikacije	nije naglašeno

Evaluacija performansi predloženih protokola vršena je u različitim simulacionim okruženjima, a neka od najčešće korišćenih su NS-3, NS-2, OPNET, Python, QualNet, MATLAB, itd. U simulacijama su posmatrane različite mrežne performanse, zavisno od toga šta je bio cilj optimizacije. Najčešće posmatrane performanse su stepen uspešno isporučenih paketa (*packet delivery ratio*), gubitak paketa (*packet loss*), kašnjenje s kraja

na kraj (*E2E delay*), protok (*throughput*), dostupni propusni opseg (*bandwidth*), broj skokova (*hop count*), overhed (*overhead*), korisnost (*utility*) itd. Kod FANET mreža su posmatrane i performanse koje su posebno bitne za ovaj tip mreža, kao što su konektivnost veze (*link connectivity*), potrošnja energije (*energy consumption*) i status leta (*flight status*). Kod svih predloženih protokola primetno je značajno poboljšanje posmatranih performansi, u poređenju sa performansama koje se postižu korišćenjem tradicionalnih protokola rutiranja. Ovo je pokazatelj efikasnosti primene učenja potkrepljivanjem, posebno u mrežama u kojima se topologija često menja i u kojima se čvorovi kreću velikom brzinom.

Tabela 2b. Komparativna analiza protokola rutiranja baziranih na učenju potkrepljivanjem u FANET mrežama

Ref.	Uticajni faktori na nagradu	Posmatrane performanse	Simulator
[24]	optimalnost susednog čvora	broj skokova, konektivnost veze	MATLAB
[25]	tip sledećeg čvora, kašnjenje, brzina čvora, potrošnja energije	stepen uspešno isporučenih paketa, kašnjenje, potrošnja energije, životni vek mreže, overhed	MATLAB
[26]	verovatnoća uspešnog prenosa paketa do sledećeg čvora	stepen uspešnosti uspostavljanja putanje, prosečni životni vek putanje, broj skokova, stepen uspešno isporučenih paketa uspešnost isporuke bez ponovnog slanja, prosečno kašnjenje	MATLAB, NS-2
[27]	da li je detektovan ometač	tačnost, uspešnost isporuke, broj skokova, broj iteracija za postizanje konvergencije, kumulativna nagrada	NS-3
[28]	vreme isteka trajanja veze, promene u skupu suseda koji prosleđuju pakete	stepen uspešno isporučenih paketa, kašnjenje	OMNeT++I NETMA-NET
[29]	da li veza vodi ka odredištu, da li je veza lokalni minimum	stepen uspešno isporučenih paketa, kašnjenje, džiter	WSNet
[30]	broj skokova, vreme uspešne isporuke paketa	broj skokova, stanje energije, status leta, brzina prenosa	nije naglašeno
[31]	proporcija poruka osetljivih na kašnjenje u redu za isporuku jednog čvora, kvalitet veze	kašnjenje, protok, gubitak paketa	NS-3
[32]	da li veza vodi ka odredištu, da li je lokalni minimum, kašnjenje, potrošnja energije	prosečno i maksimalno kašnjenje, stepen prispeća paketa, potrošnja energije	WSNet
[33]	uspešan prenos paketa	potrošnja energije, broj prekida veze, životni vek mreže	MATLAB

4. Zaključak

U ovom radu je dat pregled najaktuelnijih protokola rutiranja za VANET i FANET mreže, baziranih na učenju potkrepljivanjem. Protokoli su klasifikovani na osnovu tipa mreže, tipa učenja i eventualne kombinacije učenja potkrepljivanjem sa nekim drugim

tehnikama. Takođe je prikazana i komparativna analiza protokola rutiranja na osnovu uticajnih faktora koji određuju vrednost nagrade u postupku učenja potkrepljivanjem, kao i na osnovu posmatranih performansi na osnovu kojih je izvršena evaluacija predloženih protokola.

Zbog ekspanzije primene mašinskog učenja u WANET mrežama, planirano je dalje proširenje ovog istraživanja. U okviru daljeg istraživanja moguće je uključiti protokole koji se odnose na primenu učenja potkrepljivanjem u rutiranju kod MANET mreža, kao i primene učenja potkrepljivanjem koje se ne odnose konkretno na rutiranje paketa, kao što je određivanje optimalne pozicije i trajektorije čvorova u mreži.

Literatura

[1] R. Sutton, A. Barto, *Reinforcement Learning: An Introduction, second edition*, Cambridge, Massachusetts, MIT Press, 2018.

[2] M. L. Littman, "Reinforcement learning improves behaviour from evaluative feedback", *Nature*, vol. 521, pp. 445-451, May 2015. DOI: 10.1038/nature14540

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski et al., "Human Level Control Through Deep Reinforcement Learning", *Nature*, vol. 518, no. 7540, pp. 529–533, February 2015. DOI: 10.1038/nature14236

[4] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot and N. De Freitas, "Dueling network architectures for deep reinforcement learning", in *Proc. Int. Conf. Mach. Learn.*, pp. 1-9, June 2016.

[5] X. Bi, D. Gao, and M. Yang, "A reinforcement learning-based routing protocol for clustered EV-VANET", in *Proc. IEEE 5th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, pp. 1769–1773, June 2020. DOI: 10.1109/ITOEC49072.2020.9141805

[6] D. Zhang, F. R. Yu, R. Yang and H. Tang, "A Deep Reinforcement Learning-based Trust Management Scheme for Software-defined Vehicular Networks", in *Proc. 8th ACM Symp. Design Anal. Intell. Veh. Netw. Appl. (DIVANet)*, pp. 1-7, October 2018. DOI: 10.1145/3272036.3272037

[7] D. Zhang, F. R. Yu and R. Yang, "A Machine Learning Approach for Software-defined Vehicular Ad Hoc Networks with Trust Management", in *Proc. IEEE GLOBECOM*, pp. 1-6, December 2018. DOI: 10.1109/GLOCOM.2018.8647426

[8] X. Bi, D. Gao, and M. Yang, "A Reinforcement Learning-Based Routing Protocol for Clustered EV-VANET", in *Proc. IEEE 5th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, pp. 1769–1773, Jun 2020. DOI: 10.1109/ACCESS.2021.3058388

[9] A. Nahar and D. Das, "Adaptive Reinforcement Routing in Software Defined Vehicular Networks", in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, pp. 2118–2123, Jun. 2020. DOI: 10.1109/IWCMC48107.2020.9148237

[10] D. Zhang, F. R. Yu, and R. Yang, "Blockchain-Based Distributed Software-defined Vehicular Networks: A Dueling Deep Q-Learning Approach", *IEEE Trans. Cognitive Comm. Networking*, vol. 5, no. 4, pp. 1086–1100, December 2019. DOI: 10.1109/TCCN.2019.2944399

[11] C. Wu, T. Yoshinaga, Y. Ji, and Y. Zhang, "Computational Intelligence Inspired Data Delivery for Vehicle-to-roadside Communications", *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12038–12048, December 2018. DOI: 10.1109/TVT.2018.2871606

[12] F. Li, X. Song, H. Chen, X. Li, and Y. Wang, “Hierarchical Routing for Vehicular Ad Hoc Networks via Reinforcement Learning”, *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1852–1865, February 2019. DOI: 10.1109/TVT.2018.2887282

[13] X. Ji, W. Xu, C. Zhang, T. Yun, G. Zhang, X. Wang, Y. Wang, and B. Liu, “Keep forwarding path freshest in VANET via applying reinforcement learning”, ’ in *Proc. IEEE 1st Int. Workshop Netw. Meets Intell. Computations (NMIC)*, pp. 13–18, July 2019. DOI: 10.1109/NMIC.2019.00008

[14] C. Dai, X. Xiao, Y. Ding, L. Xiao, Y. Tang, and S. Zhou, “Learning Based Security for VANET with Blockchain”, in *2018 IEEE International Conference on Communication Systems (ICCS)*, pp. 210–215, December 2018. DOI: 10.1109/ICCS.2018.8689228

[15] C. Wu, T. Yoshinaga, D. Bayar, and Y. Ji, “Learning for adaptive anycast in vehicular delay tolerant networks”, *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, pp. 1379–1388, May 2019. DOI: 10.1007/s12652-018-0819-y

[16] D. Zhang, T. Zhang, and X. Liu, “Novel self-adaptive routing service algorithm for application in VANET”, *Applied Intelligence*, vol. 49, no. 5, pp. 1866–1879, December 2018. DOI: 10.1007/s10489-018-1368-y

[17] U. Ahmed, J. C. W. Lin, and G. Srivastava, (2021). “Privacy-Preserving Deep Reinforcement Learning in Vehicle AdHoc Networks”, *IEEE Consumer Electronics Magazine*, June 2021. DOI: 10.1109/MCE.2021.3088408

[18] B. S. Roh, M. H. Han, J. H. Ham, and K. Il Kim, “Q-LBR: Q-learning based load balancing routing for UAV-assisted VANET”, *Sensors*, vol. 20, no. 19, pp. 1–17, October 2020. DOI: 10.3390/s20195685

[19] J. Wu, M. Fang, and X. Li. “Reinforcement Learning Based Mobility Adaptive Routing for Vehicular Ad-Hoc Networks”, *Wireless Personal Communications*, vol. 101, pp. 2143–2171, May 2018. DOI: 10.1007/s11277-018-5809-z

[20] Y. Yang, R. Zhao, X. Wei, “Research on Data Distribution for VANET Based on Deep Reinforcement Learning”, in *2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, pp. 484-487, October 2019. DOI: 10.1109/AIAM48774.2019.00102

[21] M. Saravanan, P. Ganeshkumar, “Routing using reinforcement learning in vehicular ad hoc networks”, *Computational Intelligence*, vol. 36, no. 2, pp. 682–697, January 2020. DOI: 10.1111/coin.12261

[22] J. Wu, M. Fang, H. Li, and X. Li, “RSU-Assisted Traffic-Aware Routing Based on Reinforcement Learning for Urban Vanets”, *IEEE Access*, vol. 8, pp. 5733–5748, January 2020. DOI: 10.1109/ACCESS.2020.2963850

[23] S. Ye, L. Xu, and X. Li, “Vehicle-Mounted Self-Organizing Network Routing Algorithm Based on Deep Reinforcement Learning”, *Wireless Communications and Mobile Computing*, vol. 2021, July 2021. DOI: 10.1155/2021/9934585

[24] C. He, S. Liu, and S. Han, “A Fuzzy Logic Reinforcement Learning-Based Routing Algorithm For Flying Ad Hoc Networks”, in *2020 International Conference on Computing, Networking and Communications (ICNC)*, pp. 987–991, February 2020. DOI: 10.1109/ICNC47757.2020.9049705

[25] M. Y. Arafat and S. Moh, “A Q-Learning-Based Topology-Aware Routing Protocol for Flying Ad Hoc Networks”, *IEEE Internet of Things Journal*, pp. 1–1, June 2021. DOI: 10.1109/JIOT.2021.3089759

[26] Z. Zheng, A. K. Sangaiah, and T. Wang, "Adaptive Communication Protocols in Flying Ad Hoc Network", *IEEE Communications Magazine*, vol. 56, no. 1, pp. 136–142, January 2018. DOI: 10.1109/MCOM.2017.1700323
[27] N. I. Mowla, N. H. Tran, I. Doh, and K. Chae, "AFRL: Adaptive Federated Reinforcement Learning for Intelligent Jamming Defense in FANET", *Journal of Communications and Networks*, vol. 22, no. 3, pp. 244–258, June 2020. DOI: 10.1109/JCN.2020.000015
[28] B. Sliwa, C. Schüler, M. Patchou, and C. Wietfeld, "PARRoT: Predictive Ad-hoc Routing Fueled by Reinforcement Learning and Trajectory Knowledge", in *2021 IEEE 93rd Vehicular Technology Conference (VTCSpring)*, pp. 1-7, April 2021. DOI: 10.1109/VTC2021-Spring51267.2021.9448959
[29] L. A. L. F. Da Costa, R. Kunst and E. P. De Freitas, "Q-FANET: Improved Q-learning based routing protocol for FANETs", *Computer Networks*, vol. 198, October 2021. DOI: 10.1016/j.comnet.2021.108379
[30] Q. Yang, S. J. Jang, and S. J. Yoo, "Q-Learning-Based Fuzzy Logic for Multi-objective Routing Algorithm in Flying Ad Hoc Networks", *Wireless Personal Communications*, pp. 1–24, January 2020. DOI: 10.1007/s11277-020-07181-w
[31] J. Li, M. Chen, "QMPS: Q-learning based Message Prioritizing and Scheduling Algorithm for Flying Ad hoc Networks", in *2020 International Conference on Networking and Network Applications (NaNA)*, pp. 265-270, December 2020. DOI: 10.1109/NaNA51271.2020.00052
[32] J. Liu, Q. Wang, C. He, K. Jaffrès-Runser, Y. Xu, Z. Li, and Y. Xu, "QMR: Q-learning based Multi-objective optimization Routing protocol for Flying Ad Hoc Networks", *Computer Communications*, vol. 150, pp. 304–316, January 2020. DOI: 10.1016/j.comcom.2019.11.011
[33] M. Khan, K.L. Yau, "Route Selection in 5G-based Flying Ad-hoc Networks using Reinforcement Learning", in *2020 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, pp. 23-28, August 2020. DOI: 10.1109/ICCSCE50387.2020.9204944

Abstract: *Data packet routing in dynamic wireless ad hoc networks is a very demanding process due to rapid changes in the network topology, which can cause frequent interruptions of connections between nodes, and thus a high degree of lost packets. One way to overcome this problem is to apply artificial intelligence in routing protocols, in order to adapt the routing process to the dynamic nature of these networks. A very important area of artificial intelligence, which has recently been increasingly used in dynamic wireless ad hoc networks, is machine learning, especially the type of machine learning called reinforcement learning. This paper presents an overview of current results in the application of reinforcement learning in routing protocols for different dynamic wireless ad hoc networks. Routing protocols are first classified according to the type of networks into protocols for VANET (Vehicular Ad hoc Networks) and FANET (Flying Ad hoc Networks) networks, and then according to some other characteristics and specifics.*

Keywords: *reinforcement learning, routing, VANET, FANET*

APPLICATION OF REINFORCEMENT LEARNING IN ROUTING PROTOCOLS FOR DYNAMIC WIRELESS AD HOC NETWORKS

Nenad Jevtić, Marija Malnar, Pavle Bugarčić