Зборник радова Факултета техничких наука, Нови Сад

УДК: 004.9:519.83
ДОИ: https://doi.org/10.24867/35BE08Mirosavic

Систем за адаптацију стратегије у понављеним неодређеним играма нулте суме са два играча

System for Strategy Adaptation in Repeated Two-Player Zero-Sum Uncertain Games

Ђорђе Миросавић, *Факултет техничких наука, Нови Сад*

Студијски програм – ЕЛЕКТРОТЕХНИКА И РАЧУНАРСТВО

Кратак садржај – *Тема овог рада је прављење платформе која учи понашање човека у игри папир-камен-маказе, и сходне томе адаптивно бира стратегију помоћу које га побеђује. У првом делу рада биће приказане стратегије које ће се користити, прављења софтверске и хардверске платформе. У другом делу ће бити приказани резултати експеримената.*
Кључне речи (три до пет): *Теорија игара, папир-камен-маказе, Q учење, ESP32*

Abstract – *The topic of this paper is the creation of a platform that learns human behavior in the game rock-paper-scissors and, accordingly, adaptively selects a strategy to defeat them. The first part of the paper will present the strategies to be used, and the development of the software and hardware platform. The second part will present the results of the experiments.*

Keywords: (three to five): *Game theory, rock-paper-scissors, Q learning, ESP32*

НАПОМЕНА: Овај рад проистекао је из мастер рада чији је ментор био др Милан Рапаић, ред. проф.

1. УВОД

Игра папир камен маказе вуче корене још из периода династије Хан (200 година п.н.е.). Касније је игра стигла у Јапан, где се често називала *sansukumi-ken*, што у преводу значи "игра тројице, од којих се свако плаши једнога". Једна од *sansukumi-ken* игара је *Mushi-ken*, у којој су постојале 3 гестикулације руком, где једна побеђује другу. По изгледу, највише подсећа на тренутну игру папир-камен-маказе.

Уколико стриктно посматрамо проблем са теоријске стране, нема велике разлике који знак ћемо изабрати, пошто увек постоји супротан знак који га може победити. И уколико се примени Нешов еквилибријум на овај проблем, добија се да је најбоља стратегија по ком бисмо бирали потезе та да играмо насумично. И уколико имамо 2 играча који играју насумично, онда је средња добит (награда) коју оба играча могу да добију је једнака нули. Односно, након довољно великог броја партија, број победа, ремија и пораза би био изједна- чен. Међутим, човек није најбољи када је у питању насумичност. Постоји доста студија где је посматрано насумично бирање бројева, и када би се бирао број између 1 и 10, најчешће су били бирани бројеви 3 и 7. Људи заправо не могу да генеришу насумичне бројеве [1]. Такође, у једном од радова су показали да могу успешно да идентификују појединца (преко 95% успешнсти) само на основу 300 цифара које је он изабрао. [2] Слична ствар се уочила и у игри папир камен маказе, где постоји тенденција да се чешће бира камен него остали знакови.[4] Сходно томе, у овом раду ће фокус бити на прављењу агента који учи понашање човека у игри папир камен маказе, покушава да пронађеману у његовој игри и искористи је.

2. ПОСТАВКА ПРОБЛЕМА

Систем који ми посматрамо је игра папир-камен-маказе, у ком човек игра против агента. Акције које имају на располагању су папир, камен и маказе. Стање система можемо да дефинишемо на произвољно начина (уређен пар претходна акција играча - агент, претходних N акција игара, исход претходне партије,...). Награде које могу да добију су победа, нерешено или пораз. Правила игре су да папир побеђује камен, камен побеђује маказе, маказе побеђују папир. У случају да су одигране идентичне акције, резултат је нерешен. Када се игра уживо, оба играча показују акцију у исто време. У нашем случају, агент ће своју акцију испоручити пре него што корисник изабере своју, али ће се оба приказати након корисниковог избора. Агент ће имати на располагању више стратегија на основу којих може да игра. Након сваке акције играча, оне се рангирају и бира се она са највећом вероватноћом за победу у наредној партији. Интеракција са агентом је обезбеђена на 2 начина: *Flask* веб апликација и хардверска платформа базирана на *ESP32* контролеру.

3. Q-УЧЕЊЕ

Једна од стратегија коју користи јесте Q-учење. Уколико је агент у стању s одиграо акцију a, и за њу

добио награду r, ажурирање вредности коришћењем алгоритма Q учења би гласило:

$$Q^{+}(s,a) = (1-\alpha)\cdot Q(s,a) + \alpha\left[r + \gamma \cdot \max_{a'} Q(s',a')\right], \quad \alpha,\gamma \in [0,1] \tag{1}$$

Помоћу параметра учења α можемо да подешавамо колико брзо ће алгоритам реаговати на награде које добија. Први члан је претходна вредност акције, а други представља Белманову једначину. Заједно чине аналогију са филтром првог реда, где уколико желимо стабилно учење агента (са мање шума), параметар α бисмо подесили ближе нули. Након реформулације чланова, израз се чешће јавља у наредној форми:

$$Q^{+}(s,a) = Q(s,a) + \alpha\left[r + \gamma \cdot \max_{a'}(s',a') - Q(s,a)\right], \quad \alpha,\gamma \in [0,1] \tag{2}$$

Ова стратегија је нарочито ефикасна уколико за стање узмемо претходну награду коју је добио, а акцију играча интепретирамо помоћу „смера акције“. Смер акције је „понављање“ уколико су тренутна и претходна акција исте. Посматрајући редослед акција, смер „напред“ је уколико акција папир претходи камену, или из камена следе маказе, или из маказа следи папир. Смер „назад“ је обрнуто. Видећемо у резултатима да је овакво дефинисање акција у значајној мери допринела успешности агента.

3.1. Серијско Q учење

Један од начина како можемо да учење учинимо стабилнијим јесте да ажурирање вредности радимо након N партија. Формирамо листу искустава, односно, чувамо стања, изабране акције и награде. Касније у редоследу чувања података сукцесивно применимо ажурирање табеле.

3.2. Инкрементално Q учење

За разлику од обичног Q учења, овде се ажурирање табеле ради тако што формирамо речник где је кључ уређен пар (s, a) из листе искустава. Уколико запис (s, a) не постоји у табели, вредност $\mathrm{Q}(s,a)^{+}$ је средња вредност свих добијених награда за дато (s, a), тј. $Q_{\text{средње}}(s,a)$. У супротном, вредност је дата као:

$$Q^{+}(s,a) = (1-\alpha)\cdot Q(s,a) + \alpha \cdot Q_{\text{средње}}(s,a) \tag{3}$$

3.3. CAPCA (*eng. State Action Reward State Action*) алгоритам

Белманова једначина узима у обзир највећу вредност акције наредног стања. Док CAPCA бира вредност акције која би била изабрана у том стању. Разлика се примећује уколико је политика одлучивања епсилон-похлепна, с обзиром да неће увек бити изабрана вредност акције са највећом вредности.

$$Q^{+}(s,a) = Q(s,a) + \alpha[r + \gamma \cdot Q(s',a') - Q(s,a)] \tag{4}$$

3.4. Двоструко Q учење

Уколико применимо Јенсенову неједнакост над елементом за највећу Q вредност наредног стања (то можемо да урадимо пошто је функција максимума конвексна функција) добија се:

$$\mathrm{E}\left[\max_{a'} Q(s',a')\right] \geq \max_{a'} \mathrm{E}[Q(s',a')] \tag{5}$$

Ово нам говори да ако се много пута нађемо у истом стању, очекивана (средња) вредност будућег стања у које се прелази ће бити увек већа или једнака највећој очекиваној вредности. То говори да прецељујемо (*eng. overestimating)* колико је добро наредно стање у које ћемо отићи. Стога уводимо 2 табеле, где случајним одабиром бирамо коју ажурирамо након партије.

$$Q_1^{+}(\mathrm{s},\mathrm{a}) = (1-\alpha) * Q_1(s,a) + \alpha\left[r + \gamma \max_{a'} Q_2(s',a')\right] \tag{6}$$

$$Q_2^{+}(\mathrm{s},\mathrm{a}) = (1-\alpha) * Q_2(s,a) + \alpha\left[r + \gamma \max_{a'} Q_1(s',a')\right] \tag{7}$$

Вредност акције $\mathrm{Q}(\mathrm{s},\mathrm{a})$ је средња вредност обе табеле.

3.5. Дубоко Q учење

Једно од ограничења Q учења настаје када систем у ком се налазимо има велики број стања. Када се нађе у новом стању, алгоритам нема искуства о том стању и коју акцију би требало да одигра, тако да враћа насумичну. Уколико се приликом тренинга алгоритам не нађе у неком од стања, онда можемо да имамо неочекивано понашање у њима. Стога се уместо табеле користе 2 неуронске мреже. Једна за одређивање тренутне а друга на наредну вредност стања, које се изједначавају периодично. У раду је коришћена потпуно повезана секвенцијална мрежа са једним скривеним слојем. Надоградњу представља двоструко дубоко Q учење, које има значајан потенцијал у побеђивању човека у сложенијим играма. [3, 5]

4. Остале стратегије

4.1. Марковљева стратегија

Стратегија формира матрице прелаза из стања у акцију коју је изабрао играч. Свако поље матрице представља број појављивања акције у датом стању. Акција може да се бира епсилон-похлепном или политиком базираној на вероватноћи појављивања акција у стању.

4.2. Стратегија тражења секвенци

Посматра се секвенца претходних N акција играча. Проласком кроз историју потеза, бележимо колико пута је играч изабрао коју акцију након дате секвенце. Фактором заборављања више уважавамо скорије изабране акције.

4.3. Остале стратегије

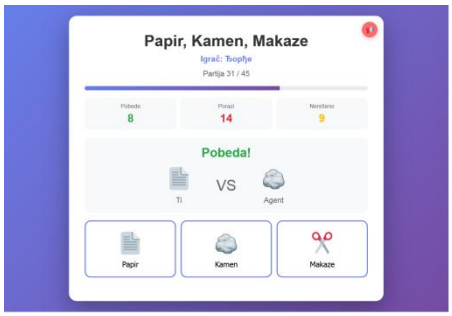
Поред споменутих, још неке од стратегија укључују коришћење неуронске мреже, посматрање

најкоришћеније акције играча, вероватноћа бирања акције, насумична...

За рангирање стратегија коришћен је UBC (*Upper Bound Confidence*) алгоритам. Параметри који се подешавају су фактор заборављања γ, ширина прозора (број посматраних партија) и одабир насумичне стратегије са вероватноћом ε.

5. Аквизиција података

За потребе извођења експеримената и прикупљања података креирана је веб апликација као и хардверска платформа.



Слика 1. Интерфејс веб апликације

Апликација је писана у *Flasku*, док је за бележење података коришћена *NoSQL* база података. Хардверска имплементација је базирана на ESP32 контролеру и SH1107 ОЛЕД екрана коришћењем *MicroPython* језика.

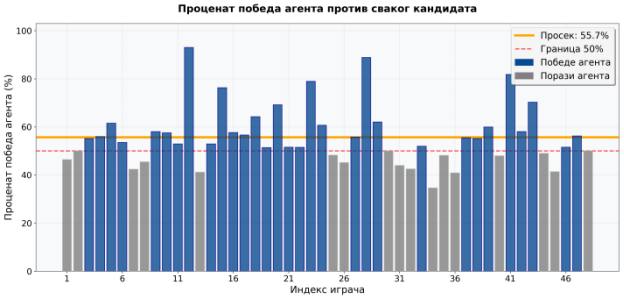
Слика 2. Хардверска платформа

6. Експеримент

У првом експерименту је коришћено 28 стратегија. Од тога је 8 било Q учење (класично и инкрементално). Ту су стања представљала једна и две претходне акције играча, један и два претходна смера акције играча, као и претходна награда играча. Код инкременталног учења, период ажурирања табеле је било 10 и 15 партија. 2 стратегије су биле CAPCA. Једно стање су представљала преходни смер акције, а друго претходна награда. 2 стратегије су биле праћење омиљене акције и смера акције играча, унутар прозора од 10 потеза са фактором заборављања 0.97. Марковљеви процеси су имали 4 стратегије где се пратила претходна акција играча уз фактор заборављања од 0.95, 0.97 и 0.995. Насумична стратегија са вероватноћама је имала 4 стратегије, од којих су 2 пратиле акције, а 2 смер акције играча. И ова случаја је једна имала фактор заборављања 0.95. 7 стратегија је било праћење секвенци, од чега је пола имало фактор заборављања 0.98. Пратиле су акције и смер акције играча, у оквиру прозора ширине 2 и 3 партије. Последња стратегија је била неуронска мрежа са 1 скривеним слојем, 3 неурона унутар њега уз фактором учења 0.04. UBC је имао епсилон-похлепну политику одлучивања са $\varepsilon = 0.05$, фактором заборављања $\gamma = 0.98$, и ширином прозора од 10 партија.
Платформа на којој су играчи играли је била веб апликација. Број партија које је играч требао да одигра је 45. Након тога може да настави игру произвољно дуго с тим да на сваких 30 партија добија питање да ли жели да настави даље. Број кандидата који је учествовао је 48. Укупан број одиграних мечева је 80.

7. Резултати

За почетак посматрамо само први одигран меч играча. Просечан број партија изражен аритметичком средином је 49.06, а медијаном 45.



Слика 3. Приказ исхода мечева

Проценат победа ћемо дефинисати као однос победа и укупног броја победа и пораза. Нерешене резултате нећемо разматрати. На слици 3 смо приказали све мечеве уз забележен проценат победа, а у табели 1 смо изразили исходе свих мечева. Агент је однео победу у 64.5% мечева.

Табела 1. Исход агента у првим одиграним мечевима

Победе	Нерешено	Порази	Укупно мечева
31	3	14	48

Аритметичка средина података свих мечева је $\mu = 55.7$ медијана $\tilde{\mu} = 53.2$, док је стандардна девијација $\sigma = 12.1$. Након уклањања резултата изнад $\mu + 2 * \sigma$, добијамо да је $\mu = 53.56$, $\tilde{\mu} = 52.9$, $\sigma = 9$.

Табела 2. Исход агента у првим одиграним партијама

Победе	Нерешено	Порази	Укупно партија
832	686	712	2230

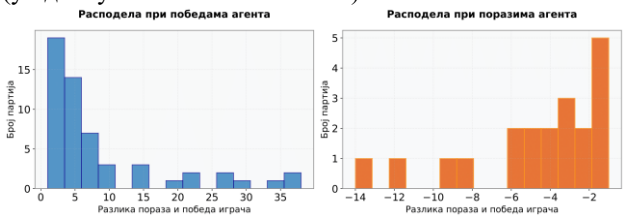
Интервал поверења за средњу вредност на нивоу поверења 95% износи $[50.83\%, 56.3\%]$. Уколико је ниво поверења 99%, онда интервал износи $[49.91\%, 57.22\%]$. Ово значи да уколико имамо велики број мечева трајања 45 партија, агентов проценат победа ће се кретати у датом опсегу. Што значи да би агент практично увек био успешнији од играча.

Посматрајмо сада свих 80 одиграних мечева. Овде је проценат победа био 68.7%.

Табела 3. Исход агента у свим одиграним партијама

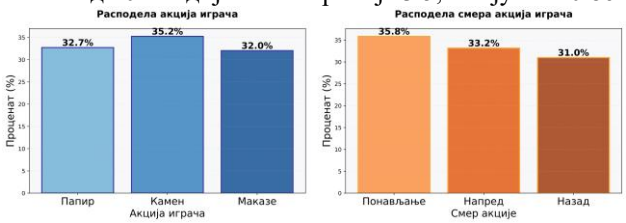
Победе	Нерешено	Порази	Укупно партија
1786	1326	1365	4477

Проценат победа по партијама је већи и износи 56.6% (у односу на 53.8% из табеле 2).



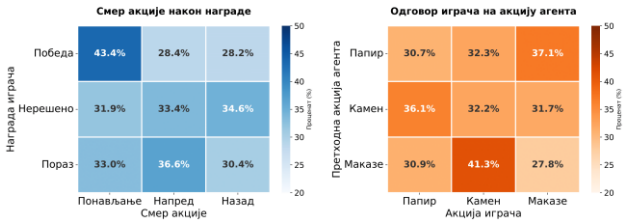
Слика 4. Убедљивост агента при победама и поразима по мечевима

Са слике 4 имамо увид у разлику са којом је агент побеђивао и губио. Када је агент побеђивао, медијана је +5, тј. имао је 5 победа више него пораза, а скјунис метрика је 1.13. Видимо да је постојало пар мечева где је агент имао предност од 20 или више победа, али то се дешавало када је играч више пута продужавао наставак игре. Посматрањем пораза агента, примећујемо да је постојало пар играча који су приметилиману у игри и побеђивали га са разликом већом од 10. Медијана за поразе је 3.5, а скјунис -0.86.



Слика 5. Приказ расподела акција играча у свим партијама

Посматрајући најчешћу акцију коју су играчи бирали, видимо да је то камен, што се поклапа са резултатима осталих великих експеримената који су рађени [4]. Као смер акције, доминантно је понављање претходно одигране акције. Такође, чешће се бира акција која је позиционирана десно односу на претходну акцију, него лево.



Слика 6. Матрице прелаза

На слици 6 смо приказали 2 матрице прелаза. Прва приказује прелаз из награде играча у смер одигране акције. Уколико је играч победио, интересантно је да у 43.4% случајева понавља исту акцију. Матрица десно, приказује одговор играча спрам претходне акције агента. Можемо приметити да је доминантно понашање играње акције која ће победити тренутну акцију агента. Ово је наизраженије код агентове акције Маказе, након које играч у 41.4% бира Камен. У табели 4 смо издвојили стратегије које су донеле највише победа агенту. Праћењем смера акције играча, можемо да приметимо највећи број победа.

Табела 4. Најуспешније стратегије

Назив стратегије	Број победа	Број ремија	Број пораза
Q учење (награда->смер)	152	85	77
Q учење (смер->смер)	148	95	98
Марковљева (награда->смер)	124	51	70
Најкоришћенији смер	123	47	54
Q инкрементално учење (смер->смер)	84	62	57

8. ЗАКЉУЧАК

У овом раду смо показали да је могуће направити систем који ће победити човека у игри папир-камен-маказе. Посматрајући први одигран меч играча, проценат победа агента је 64.5%, док у појединачним партијама 53.5%. Ово је значајан резултат узимајући у обзир да је просечан број одиграних партија (изражен медијаном) био свега 45.

9. ЛИТЕРАТУРА

[1] Figurska M, Stańczyk M, Kulesza K. (2008). *"Humans cannot consciously generate random numbers sequences: Polemic study"*

[2] Schulz, MA., Baier, S., Timmermann, B. (2021). *"A cognitive fingerprint in human random number generation"*

[3] Hasselt, H., Guez, A., & Silver, D. (2015). *"Deep Reinforcement Learning with Double Q-learning"*

[4] Wang, Z., Xu, B. & Zhou, HJ. (2014). *"Social cycling and conditional responses in the Rock-Paper-Scissors game"*

[5] Kumar, D. (2024). *"DDQN: Tackling Overestimation Bias in Deep Reinforcement Learning"*

Кратка биографија:

Ђорђе Миросавић рођен је у Ваљеву 2000. године. Мастер рад на Факултету техничких наука из области Електротехнике и рачунарства одбранио је 2025.год.
Контакт:
mirosavicdjordje@gmail.com