INTELLEKTUAL BOSHQARISHDA QAYTA TAYYORLASH VA MALAKASINI OSHIRISH JARAYONLARINI INTELLEKTUAL BOSHQARISHDA QO‘LLANILADIGAN KO‘P MAQSADLI VA KO‘P AGENTLI TIZIMLARNING MA’LUMOTLARINI O‘QITISH

Boynazarov Otabek Murot o‘g‘li
Ichki ishlar vazirligi Malaka oshirish instituti o‘qituvchi
Toshkent, O‘zbekiston
E-mail: boynazarovotabek0712@mail.ru

Annotatsiya: Maqolada intellektual boshqarishning samarali yondashuvlari xususan bu jarayonni **ko‘p maqsadli** hamda **ko‘p agentli tizimlar** doirasida matematik modellashtirish masalasi tahlil qilinadi. Bu jarayonni mustahkamlovchi o‘rganish (Reinforcement Learning, RL) va Markov qarorlar jarayoni (MQJ) asosida matematik modellashtirish taklif etiladi. Shuningdek, ko‘p agentli tizimlar doirasida o‘qitish jarayonini avtomatlashtirish va boshqaruv qarorlarini optimallashtirish tamoyillari yoritilgan.

Kalit so‘zlar: Mustahkamlovchi o‘rganish, Markov qarorlar jarayoni, ko‘p agentli tizim, intellektual boshqaruv, Bellman tenglamasi, actor-critic usuli, xususiyat gradienti, o‘qitish algoritmlari, optimallashtirish.

Аннотация: В статье анализируются эффективные подходы к интеллектуальному управлению, в частности, математическое моделирование этого процесса в рамках многоцелевых и многоагентных систем. Предлагается математическое моделирование на основе усиливающего обучения (Reinforcement Learning, RL) и Марковского процесса принятия решений (MQJ). Также освещены принципы автоматизации процесса обучения и оптимизации управленческих решений в рамках многоагентных систем.

Ключевые слова: Укрепляющее обучение, Марковский процесс принятия решений, многоагентная система, интеллектуальное управление, уравнение Беллмана, метод actor-critic, градиент свойств, алгоритмы обучения, оптимизация.

Abstract: The article analyzes effective approaches to intelligent control, in particular, the issue of mathematical modeling of this process within multi-purpose and multi-agent systems. Mathematical modeling is proposed based on Reinforcement Learning (RL) and the Markov Decision Process (MQJ). The principles of automation of the learning process and optimization of management decisions within multi-agent systems are also highlighted.

Keywords: Reinforcing learning, Markov decision process, multi-agent system, intelligent control, Bellman equation, actor-critic method, property gradient, learning algorithms, optimization.

Kirish.

Mazkur maqolada xodimlarini qayta tayyorlash va malakasini oshirish jarayonlarini intellektual boshqarishning samarali yondashuvlarini ko‘rib chiqamiz. Xususan, bu jarayonni **ko‘p maqsadli** hamda **ko‘p agentli tizimlar** doirasida matematik modellashtirish masalasi tahlil qilinadi. Qayta tayyorlash va malaka oshirishni samarali boshqarish uchun **o‘qitishni mustahkamlash (Reinforcement Learning, RL)** usullaridan foydalanish maqsadga muvofiq deb topiladi. Mazkur yondashuv orqali xodimlarning individuallashtirilgan ta’lim strategiyalarini shakllantirish va ularni xizmat faoliyatidagi samaradorligi asosida optimallashtirish mumkin bo‘ladi.

Buning uchun mazkur tizim **Markov qarorlar jarayoni (MQJ)** muhiti sifatida modellashtiriladi. MQJ asosida har bir xodimning bilim darajasi, faoliyati, o‘zlashtirish tezligi kabi parametrlar inobatga olingan holda, ularning malakasini oshirish bo‘yicha optimal qarorlar ketma-ketligini aniqlash imkoniyati yaratiladi.

Shunday qilib, maqolada xodimlarini qayta tayyorlash va malakasini oshirish jarayonlarining intellektual boshqaruvi uchun **ko‘p agentli, ko‘p maqsadli tizimlar** asosida **RL va MQJ yondashuvlari yordamida matematik model ishlab chiqish** tamoyillari asoslab beriladi.

Xodimlarini qayta tayyorlash va malakasini oshirish jarayonlarini intellektual boshqarishda Markov qarori jarayoni $M=(S,A,T,y,R)$ korteji sifatida belgilab olamiz. Bunda,

S – tizimning holat fazosi (xodimning bilim darajasi, kasbiy tajribasi ва boshqa omillar yig‘indisi);

A – harakatlar (yoki qarorlar) fazosi (boshqaruv qarorlar: kursga yo‘naltirish, testdan o‘tkazish, bilimini baholash va boshqalar);

$T:S\times A\times S\to[0,1]$ - ehtimollik o‘tish funksiyasi;

$y\in[0,1]$ - chegara omili;

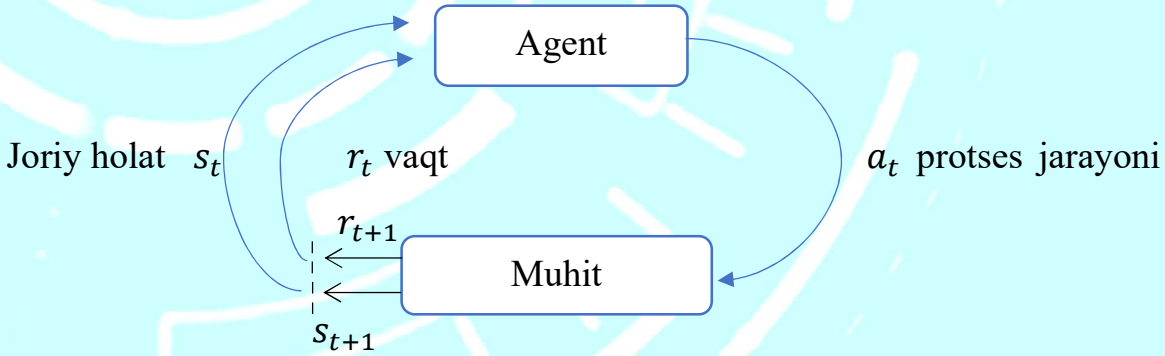
MQJ - ketma-ket qaror qabul qilish jarayonida muhitni S tizimning holat fazosi, A harakatlar (yoki qarorlar) fazosi, muhit dinamikasini kodlaydigan T ehtimollik o‘tish funksiyasi, R baholash funksiyalaridan iborat kortej sifatida modellashtiruvchi matematik tizim bo‘yicha agent tomonidan olingan raqamli qiymatni aniqlaydi. Dinamik dasturlashdan farqli o‘laroq [1], mustahkamlashni o‘rganish sharoitida agentlar muhit modeliga, ya’ni T o‘tish funksiyasiga va R baholash funksiyasiga kirish imkoniga ega emas deb taxmin qilamiz [2]. Xodimlarini qayta tayyorlash va malakasini oshirish jarayonlarini intellektual boshqarish, ya’ni agentning protses jarayoni uning xususiyati $\pi:S\times A\to[0,1]$, bilan belgilanadi. Qarorlar fazosi bajarilgandan so‘ng, muhit harakatlar (yoki qarorlar) fazosi (boshqaruv qarorlar: kursga yo‘naltirish, testdan o‘tkazish, bilimini baholash va boshqalar) bilan birga keyingi holatni qaytaradi. Ushbu ketma-ket o‘zaro ta’sir sxemasi 1-rasmda keltirilgan. Agentning maqsadi kutilgan

ISSN 3030-3907

chegirmali baholash jarayoni, ya'ni kutilgan natijalarni maksimal darajada oshiradigan π xususiyatini topish quyidagi ifoda orqali aniqlanadi:

$$V^{\pi} = E\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| \pi, \mu_0\right]$$

Bu yerda $\mu_0 = \mu_0(s_0)$ boshlang'ich holatlar bo'yicha taqsimot (o'quv jarayoni boshlanishidagi xodimlarning dastlabki tayyorgarlik darajasi); $s_0 \in S, \gamma$ chegirma koeffitsienti va $r_t = R(s_t, a_t, s_{t+1})$ - agent tomonidan t vaqt oralig'ida chora ko'rganlik uchun olingan baholash, $a_t \in A, s_t \in S$ joriy holat (xodimning hozirgi bilim/tajriba darajasi) va $s_{t+1} \in S$. - keyingi jarayonga tizimning holat fazosi o'tish. Agent va muhit o'rtasidagi o'zaro ta'sir quyidagi trayektoriyaga olib keladi: $s_0, a_0, r_0, s_1, a_1, r_{1,\ldots,}, \ldots,$ har bir epizod uchun bo'ladi.



1-rasm. Markov qarori jarayonida agent va muhit o'zaro ta'sirini tavsiflovchi *m* o'quv sikli

Keling, ko'plab *O'M* algoritmlarida ishlatiladigan bir nechta keng tarqalgan elementlarni muhokama qilaylik. s holatining qiymat funksiyasi π xususiyati ostida agentning kutilayotgan natijasini s holatidan boshlab va undan keyin π dan keyin belgilaydi:

$$V^{\pi}(s) = E\left[\sum_{k=0}^{\infty} \gamma^t r_{t+k} \middle| \pi, s_0 = s\right] \quad (1)$$

Bu erda $r_{t+k} = R(s_{t+k}, a_{t+k}, s_{t+k+1})$ agent tomonidan $t+k$ vaqt bosqichida, $a_{t+k} \in A$, $s_{t+k} \in s$ tizimning holat fazosida chora ko'rganlik uchun olinadigan baholash va keyingi $s_{t+k+1} \in s$ holatga o'tish.

Bundan tashqari, harakat-qiymat funksiyasini π xususiyati ostida s tizimning holat fazosida a protses jarayonini amalga oshirishda kutilgan natija sifatida belgilashimiz mumkin:

$$Q^{\pi}(s,a) = E\left[\sum_{t=0}^{\infty} \gamma^t r_{t+k} \middle| \pi, s_t = s, a_t = a\right] \quad (2)$$

Mustahkamlovchi o'rganishda, dinamik dasturlashda ham qo'llaniladigan qiymat funksiyalari bilan qanoatlantiriladigan rekursiv munosabatni hosil qilamiz

ISSN 3030-3907

[3], ya'ni, Bellman tenglamasidan xodimlarini qayta tayyorlash va malakasini oshirish jarayonlarini intellektual boshqarishda foydalanamiz.

$$V^{\pi}(s)=\sum_{a}\pi(s,a)\sum_{s^{t}}T(s,a,s')\left(R(s,a,s')+\gamma V^{\pi}(s')\right) \qquad (3)$$

Bellman tenglamasi shuni ko'rsatadiki, holatning qiymati darhol kutilgan baholashga, kutilayotgan keyingi holatning diskontlangan qiymatiga teng bo'lib, holatning qiymat funktsiyasi va uning davom ettiruvchilarning qiymati o'rtasidagi munosabatni o'rnatadi. Umuman olganda, holatning qiymat funksiyasi - bu barcha mumkin bo'lgan harakatlar *(ularni ko'rib chiqilayotgan π xususiyatiga muvofiq qabul qilish ehtimoli bilan tortilgan)* hamda, keyingi holatlar va baholashlar *(shuningdek, ularga erishish ehtimoli bo'yicha tortilgan muhitga o'tish funksiyasi T bilan aniqlanadi)*. Bellman tenglamasi ko'plab mustahkamlashni o'rganish algoritmlarining asosini tashkil qiladi, ularning ba'zilari ushbu ishda qo'llaniladi va kengaytiriladi. Bir nechta optimal xususiyat mavjud bo'lishi mumkin bo'lsa-da, ularning barchasi, $s\in S$ uchun bir xil optimal holat-qiymat funksiyasi $V*s=\max_{\pi}V^{\pi}(s)$, shuningdek, barcha $s\in S$ va barcha $a\in A$ uchun optimal protses jarayon qiymati funksiyasi $Q*(s,a)=\max_{\pi}Q^{\pi}(s,a)$ ga teng bo'ladi.

Agentli tizimlarga asoslangan qarorlarini baholash

O'M texnikasining umumiy qiymatga asoslangan algoritmlardir. Qiymatga asoslangan algoritmlarda maqsad odatda (2) tenglamada aniqlangan Q harakat-qiymat funksiyasining bahosini topishdir.

Q-learning [4] qiymatga asoslangan RL algoritmi bo'lib, unda qiymat funksiyasi keyingi holatning taxminiy qiymatini yuklash va foydalanish orqali kutilgan uzoq muddatli baholashni optimallashtirish uchun takroriy ravishda yangilanadi. Xususan, s holatidan s' ga o'tgandan so'ng a harakati orqali Q-learning quyidagi yangilanishni amalga oshiradi:

$$Q(s,a)\leftarrow Q(s,a)+\alpha\left[r+\gamma_{s'}^{\max}Q(s',a')-Q(s,a)\right]; \qquad (4)$$

Bu yerda, α - o'rganish darajasi, γ - chegirma omili va r - muhitdan olingan natijasini darhol baholash. *0* ga yaqin bo'lgan γ qiymati "miyopik" (yoki uzoqni ko'rmaydigan) agentni yaratadi, ya'ni darhol baholash qiymatiga ko'proq ahamiyat beradi, *1* ga yaqin qiymatdan foydalanish agentni uzoq muddatli baholashlar uchun ko'proq optimallashtirishga majbur qiladi.

Q-learning amal qilinayotgan xususiyatidan qat'iy nazar, $Q*(s,a)$ optimal protses jarayoni qiymati funksiyasini to'g'ridan-to'g'ri yaqinlashtirish imkonini beradi, chunki u keyingi holatning maksimal protses jarayoni qiymatidan foydalangan holda yuklanadi.

Xususiyat gradienti [3,4] qiymatga asoslangan usullarda bajarilganidek, qiymat funksiyalariga asoslangan xususiyatni bilvosita xulosa qilish o'rniga θ tomonidan parametrlangan π_{θ} xususiyatini to'g'ridan-to'g'ri o'rganadigan

ISSN 3030-3907

mustahkamlovchi o‘rganish algoritmlari oilasidir. Xususiyat gradienti usullari agentning muhit bilan o‘zaro ta’sir qilish tajribasidan *(ya’ni kuzatilgan holatlar, harakatlar va baholashlar)* foydalangan holda θ ga nisbatan maqsadning $J(\theta)$ gradientlarini hisoblab chiqadi va ushbu gradient yo‘nalishida bosqichma-bosqich B parametrlarini yangilaydi:

$$\theta_{t+1} = \theta_t + \alpha \nabla J(\theta). \quad (5)$$

$J(\theta)$ maqsad funksiyasi kutilgan natijalar kabi samaradorlik o‘lchovining ba’zi shakllarini ifodalaydi (3.1 tenglama):

$$J(\theta) = V^{\pi\theta} = E\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| \pi_0, \mu_0\right]. \quad (6)$$

π_0 uning parametrlari bo‘yicha farqlanadigan bo‘lsa, xususiyatni qanday parametrlashtirish kerakligi haqida odatda hech qanday cheklovlar yo‘q.

Xususiyat gradient usullari oilasi ichida Actor-critic usullar deb ataladigan yana bir kuchli o’qitish usullari sinfi mavjud. Ushbu usullar Actor deb ataladigan xususiyatni, shuningdek critic deb ataladigan qiymat funktsiyasini o‘rganadi [3]. Xususiyat gradienti usullari shuning uchun faqat Actor uchun usullar sifatida ham tanilgan. Faqat Actor uchun usullar bilan solishtirganda, tanqidchidan foydalanish odatda gradientlardagi tafovutni kamaytiradi va shuning uchun ko‘pincha barqarorroq xususiyat yangilanishiga erishadi. Yagona maqsadli sozlamalar uchun ikkala sinfda ham bitta, ham ko‘p agentli sozlamalar uchun mashhur zamonaviy usullar mavjud [5,6,7,8,9]. Epizodik vazifalarda xususiyat gradienti teoremasi [3,10] π holat taqsimotining hosilasini talab qilmasdan, xususiyat parametriga nisbatan maqsadning gradientiga mutanosib bo‘lgan quyidagi analitik ifodani taqdim etadi:

$$J(\theta) \propto \sum_s \pi(s) \sum_a Q^\pi(s,a) \nabla \pi(a|s,\theta) \quad (7)$$

Bu iboraning birinchi atamasi maqsadli π xususiyati bo‘yicha yuzaga kelgan holatlar bo‘yicha yig‘indini ifodalaydi. Agar π ga rioya qilinsa, shtatlar shu nisbatda uchraydi, ya’ni gradientni keyinroq quyidagicha yozish mumkin:

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a Q^\pi(s,a) \nabla \pi_\theta(a|s) \quad (8)$$

$$= \underset{\mu(s)}{E}\left[\sum_a Q^\pi(s,a) \nabla \pi_\theta(a|s)\right] \quad (9)$$

$$= \underset{\mu(s)}{E}\left[\sum_a \pi_\theta(a|s) Q^\pi(s,a) \frac{\nabla \pi_\theta(a|s)}{\pi_\theta(a|s)}\right] \quad (10)$$

$$= \underset{\substack{\mu(s) \\ \pi_\theta(a|s)}}{E}\left[Q^\pi(s,a) \frac{\nabla \pi_\theta(a|s)}{\pi_\theta(a|s)}\right] \quad (11)$$

ISSN 3030-3907

$$= \underset{\substack{\mu(s)\\ \pi_\theta(a|s)}}{E}\left[Q^\pi(s,a)\nabla \log \pi_\theta(a|s)\right] \quad (12)$$

Amalda $Q^\pi(s,a)$ harakat-qiymat funksiyasining o‘rganilgan $\hat{Q}^\pi(s,a)$ yaqinlashuvi bilan almashtirilishi mumkin.

Umuman olganda, xususiyat gradient usullari shaklning ifodasini baholashga qaratilgan [11]:

$$\nabla J(\theta) = E\left[\sum_{t=0}^{\infty}\psi_t \nabla \log \pi_\theta(a_t|s_t)\right] \quad (13)$$

ψ_t uchun ikkita umumiy shakl bilan hamda, ushbu ishda ko‘rib chiqilgan:

$\psi_t = \sum_{t=0}^{\infty}\psi^t r_t, i.e$ yani trayektoriyaning umumiy qaytishi

$\psi_t = \hat{Q}^\pi(s_t,a_t)$, ya’ni, harakat-qiymat funksiyasi

Ko‘p agentli holatga o‘tayotganda, har bir agent o‘z qiymatlarini va/yoki xususiyat parametrlarini yangilash uchun taqdim etilgan o‘qitish algoritmlaridan birini mustaqil ravishda qo‘llaydi.

Ko‘p agentli tizimlar ko‘plab real muammolarni modellashtirish uchun tabiiy paradigma sifatida paydo bo‘ldi [12], aqlli tarmoq boshqaruvi [13; 14] transport [15] va IOT [16], chunki ular katta taqsimlangan tizimlar g‘oyasiga juda mos keladi. Ko‘p agentli qaror qabul qilish muammolari ko‘pincha stokastik (yoki Markov) zanjiri (SG) sifatida modellashtirilishi mumkin.

Stoxastik jarayonni quyidagicha aniqlash mumkin:

Stoxastik (Markov) zanjiri $M = (S, A, T, R)$, $n \geq 2$ ta agent bilan, bu yerda:

- S – tizim holati fazosi
- $A = A_1 \times ... \times A_n$ harakatlar (yoki qarorlar) fazosi (boshqaruv qarorlar: kursga yo‘naltirish, testdan o‘tkazish, bilimini baholash va boshqalar) to‘plami; $A_i - i$ agentning protses jarayon to‘plami
- $T : S \times A \times S \to [0,1]$ ehtimollik o‘tish funksiyasi
- $R = R_1 \times ... \times R_n$ muhitni baholashlash funksiyalari, bu yerda
- $R : S \times A \times S \to R$ i agent i ning baholashlash funksiyasidir.

Har bir vaqt oralig‘ida muhit qo‘shma holatni chiqaradi $s = \langle s_1, ..., s_n \rangle \in S$. Agent tomonidan olingan baholash faqat uning o‘z protses jarayonlariga emas, balki muhitdagi barcha agentlarning birgalikdagi protses jarayonlariga bog‘liq.

Agentning protses jarayoni uning i xususiyati bilan belgilanadi: $\pi_i : S \times A_i \to [0,1]$ ya’ni vaziyat berilganda, protses jarayonlar ma’lum bir ehtimollik taqsimotiga ko‘ra tanlanadi. Cheklangan cheksiz holatida agentning maqsadi kutilgan chegirmali uzoq muddatli baholashni maksimal darajada oshiradigan i xususiyatini aniqlaydi:

ISSN 3030-3907

$$V^{\pi i} = E\left[\sum_{t=0}^{\infty} \gamma^t r_i, t \middle| \pi, \mu_0\right] \quad (14)$$

Bu yerda $\pi = (\pi_1, ..., \pi_n)$ - agentlarning qo‘shma xususiyati, μ_0 - boshlang‘ich holatlar bo‘yicha taqsimlash $s_0 \in S$, γ - chegirma koeffitsiyenti va $r_{i,t} = R_i(s_t, a_t, s_{t+1})$ - agent i tomonidan t vaqt oralig‘ida, $a_t \in A$ qo‘shma protses jarayon uchun, $s_t \in S$ holatida va keyingi $s_{t+1} \in S$ holatiga o‘tish uchun olinadigan baholash.

Xodimlarini qayta tayyorlash va malakasini oshirish jarayonlarini intellektual boshqarishda ko‘p agentli tizimlarda o‘qitishni mustahkamlash[3] (O‘M) muhim ahamiyat kasb etadi. Chunki bunday o‘quv muhitlari ko‘pincha yuqori murakkablik va stoxastik xususiyatlar bilan tavsiflanadi. Bu esa oldindan qat’iy dasturlashtirilgan yondashuvlar orqali optimal qaror va harakatlarga erishishni murakkablashtiradi yoki imkonsiz qiladi.

Shu nuqtai nazardan, xodimlarning individuallashtirilgan ta’lim yo‘nalishlarini shakllantirish, o‘zlashtirish tezligi, xizmatdagi faoliyat samaradorligi kabi mezonlar asosida ularga mos strategiyalarni ishlab chiqish zarurati yuzaga keladi. Bunday muhitda samarali boshqaruvga erishish uchun ko‘p agentli O‘M yondashuvlari muhim vosita hisoblanadi.

FOYDALANLGAN ADABIYOTLAR RO‘YXATI:

1. Bellman, R. Dynamic Programming. Princeton, NJ, USA: Princeton University Press. 1957y.
2. Busoniu, L., R. Babuška, B. De Schutter, and D. Ernst Reinforcement learning and dynamic programming using function approximators. CRC press. 2017y.
3. Sutton, R. and A. Barto Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press. 1998y.
4. Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn., 8(3-4):229–256. 1992y.
5. Silver, D., G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller Deterministic policy gradient algorithms. In International conference on machine learning, Pp. 387–395. PMLR. 2014y.
6. Haarnoja, T., A. Zhou, P. Abbeel, and S. Levine Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International Conference on Machine Learning, Pp. 1861– 1870. PMLR. 2018y.
7. Foerster, J., R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch Learning with opponent-learning awareness. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, Pp. 122– 130. 2018y.

ISSN 3030-3907

8. Foerster, J., G. Farquhar, M. Al-Shedivat, T. Rocktäschel, E. Xing, and S. Whiteson Dice: The infinitely differentiable monte carlo estimator. In International Conference on Machine Learning, Pp. 1529–1538. 2018y.
9. Lowe, R., Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch Multi-agent actor-critic for mixed cooperative-competitive environments. In Advances in Neural Information Processing Systems, Pp. 6379–6390. 2017y.
11. Sutton, R. S., D. A. McAllester, S. P. Singh, Y. Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In NIPs, volume 99, Pp. 1057–1063. Citeseer. 1999y.
12. Schulman, J., P. Moritz, S. Levine, M. Jordan, and P. Abbeel High-dimensional continuous control using generalized advantage estimation. In Proceedings of the International Conference on Learning Representations (ICLR). 2016y.
13. Hurtado, C., M. R. Ramirez, A. Alanis, S. O. Vazquez, B. Ramirez, and E. Manrique Towards a multi-agent system for an informative healthcare mobile application. In KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications, Pp. 215–219. Springer. 2018y.
14. Moradi, M. H., S. Razini, and S. M. Hosseinian. State of art of multiagent systems in power engineering: A review. Renewable and Sustainable Energy Reviews, 58:814–824. 2016y.
15. Khan, M. W. and J. Wang. The research on multi-agent system for microgrid control and optimization. Renewable and Sustainable Energy Reviews, 80:1399–1411. 2017y.
16. Hamidi, H. and A. Kamankesh. An approach to intelligent traffic management system using a multi-agent system. International Journal of Intelligent Transportation Systems Research, 16(2):112–124. 2018y.
17. Calvaresi, D., M. Marinoni, A. Sturm, M. Schumacher, and G. Buttazzo. The Challenge of Real-time Multi-agent Systems for Enabling IoT and CPS. In Proceedings of the International Conference on Web Intelligence, WI '17, Pp. 356–364, New York, NY, USA. ACM. 2017y.

ISSN 3030-3907