УДК 004.94, 004.852
Асєєв Владислав Дмитрович, аспірант, Кулаковська Інесса Василівна, к. ф-мат. н,
Чорноморський національний університет імені Петра Могили,

МОДЕЛЮВАННЯ СТРАТЕГІЙ АГЕНТІВ В ІГРАХ З ЛОКАЛЬНОЮ ВЗАЄМОДІЄЮ

Асєєв В.Д., Кулаковська І.В. Моделювання стратегій агентів в іграх з локальною взаємодією. У даній роботі розглянуто актуальну тему оптимальних стратегій в іграх з локальною взаємодією, розглянуто стимулююче навчання багатоагентних систем у ігровій постановці. Метою даної роботи є розгляд методу побудови системи з локальною взаємодією агентів на основі завдання «синхронізації» за допомогою марковської моделі стохастичної гри. Метод дослідження - комп'ютерна програма для моделювання завдання з використанням Q-методу навчання.
Ключові слова: мультиагентна система, стохастична гра, адаптивний ігровий метод, Q-метод.

Асеев В. Д., Кулаковская И. В. Моделирование стратегий агентов в играх с локальным взаимодействием. В данной статье рассматривается актуальная тема оптимальных стратегий в играх с локальным взаимодействием, рассматривается стимулирующее обучение многоагентных систем в разработке игр. Целью данной работы является рассмотрение метода построения системы с локальным взаимодействием агентов на основе задачи «синхронизации» с использованием марковской модели стохастической игры. Метод исследования - компьютерная программа для моделирования задачи с использованием Q-метода обучения.
Ключевые слова: многоагентная система, стохастическая игра, метод адаптивных игр, Q-метод.

Asieiev V.D., Kulakovska I.V. Modeling agents strategies in players with local interaction. In this paper, the current topic of optimal strategies in games with local interaction is considered, the incentive training of multiagent systems in game formulation is considered. The purpose of this work is to consider the method of constructing a system with local interaction of agents based on the task of "synchronization" using the Markov model of the stochastic game. The research method is a computer program for modeling a task using the Q-method of training.
Key words: multiagent system, stochastic game, adaptive gaming method, Q-method.

Постановка наукової проблеми. Порівняно з одноагентними системами структура, функціонування та дослідження методів багатоагентного Q-навчання значно ускладнюються. За рахунок колективної взаємодії агентів стаціонарне середовище переводиться у клас нестаціонарних. Зміна станів середовища та значення виграшів кожного агента залежать від дій інших агентів. У загальному випадку у МАС агент не може досягти максимального виграшу, який дорівнює його виграшу в одноагентній системі. Оптимальні виграші агентів повинні бути збалансованими і відповідати критеріям вигодих, справедливості,рівноваги. Так, замість критерію скалярної максимізації виграшів одноагентної системи, вводяться критерії векторної максимізації виграшів МАС, наприклад, рівноваги за Нешем, оптимальності за Парето або ін.

За умови використання методів Q-навчання МАС відбувається ітераційна побудова системи характеристичних Q-функцій у просторі стан-дія, причому приріст елементів цих функцій здійснюється у напрямку досягнення їх колективної рівноваги [1].

Для побудови МАС необхідно виконати попередні дослідження на основі адекватних математичних моделей, які дають змогу вивчити динаміку системи в умовах невизначеності, побудувати стратегії поведінки агентів, які забезпечують оптимальні техніко-економічні параметри функціонування системи[10]. Враховуючи особливості предметної області, а саме багатоагентність, невизначеність середовища прийняття рішень, антагонізм або конкурентність цілей, комунікативність, координація дій, адаптивність стратегій поведінки агентів, для побудови моделей МАС використаємо математичний апарат теорії стохастичних ігор [2,3]. Розв'язування стохастичної гри полягає у пошуку таких стратегій агентів, які максимізують їх виграші так, щоб забезпечити певний колективний баланс інтересів усіх гравців [9]. Шукати оптимальні стратегії гравців в умовах невизначеності будемо за методом заохочувального навчання.

Метою роботи є побудова ігрової моделі самоорганізації мультиагентних систем для підтримки прийняття рішень в умовах невизначеності. Ця мета досягається розв'язуванням таких задач: розроблення математичної моделі мультиагентної стохастичної гри; розроблення самонавчального методу та алгоритму розв'язування стохастичної гри; розроблення програмних засобів моделювання стохастичної гри; аналіз отриманих результатів та вироблення рекомендацій для їх практичного застосування. Методом дослідження є комп`ютерна програма для моделювання задачі.

Аналіз досліджень. Функціонування МАС [4,5], як правило, здійснюється в умовах апріорної невизначеності інформації про стани середовища прийняття рішень та дії інших агентів. У зв'язку з цим стратегії поведінки агентів повинні бути адаптивними за рахунок здатності агентів до самонавчання [6]. Серед методів навчання в умовах невизначеності практичної привабливості набули методи, які грунтуються на заохоченнях [7,8], оскільки вони не вимагають математичної моделі середовища та забезпечують можливість приймати рішення безпосередньо в процесі навчання. В основу заохочувального навчання покладено механізми рефлекторної поведінки живих організмів з розвинутою нервовою системою. Ефективним методом заохочувального навчання є марківське Q-навчання [1], яке здійснює числову ідентифікацію характеристичної функції динамічної системи у просторі "стан–дія". Як характеристичну функцію переважно використовують функцію сумарної очікуваної винагороди агента.

Використано ідеї статі П.О.Кравця [9] «Ігрова модель самоорганізації мультиагентних систем», яка розглядає основні властивості МАС та зв`язок задачі "імітувати синхронізоване ритмічне світіння колонії комах-світлячків" з МАС. Метою моделювання є визначення умов та механізмів локальної координації агентів, для самоорганізації МАС. Для цього необхідно розв'язати такі задачі: побудувати модель гри, розробити метод та алгоритм для розв'язування та виконати програмне комп'ютерне моделювання для виявлення координації та самоорганізації МАС.

Виклад основного матеріалу й обґрунтування отриманих результатів дослідження. Для виконання експериментів використовуємо обґрунтоване комунікаційне середовище, яке складається з n агентів і l станів, що населяють двовимірний світ з безперервним простором і дискретним часом. Агенти можуть здійснювати фізичні дії в середовищі та комунікаційні дії, які передаються іншим агентам. Вважаємо, що агенти не мають ідентичний простір для дій і спостережень,а обмежені, але вони діють за однією і тією ж стратегією π. Розглядаємо ігри, які є кооперативними (всі агенти повинні максимізувати спільний результат) і конкурентні (агенти мають протилежні цілі). Деякі середовища вимагають явного зв'язку між агентами для досягнення найкращої винагороди, тоді як в інших середовищах агенти можуть виконувати тільки фізичні дії. Інформація про кожне середовище приведена нижче.

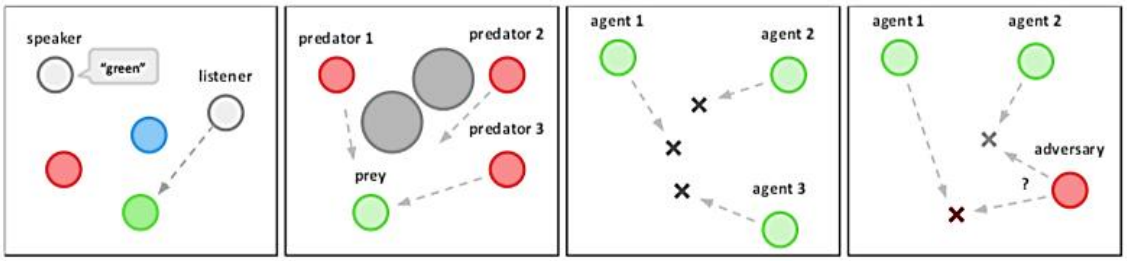

Рис. 1. Ілюстрації експериментального середовища і деякі розглянуті завдання, в тому числі: а) кооперативна комунікація, б) хижак-жертва, в) кооперативна навігація, г) фізичний обман.

Для побудови МАС необхідно виконати попередні дослідження на основі адекватних математичних моделей, які дають змогу вивчити динаміку системи в умовах невизначеності, побудувати стратегії поведінки агентів, які забезпечують оптимальні техніко-економічні параметри функціонування системи. Враховуючи особливості предметної області, а саме багатоагентність, невизначеність середовища прийняття рішень, антагонізм або конкурентність цілей, комунікативність, координація дій, адаптивність стратегій поведінки агентів, для побудови моделей МАС використаємо математичний апарат теорії стохастичних ігор. Розв'язування стохастичної гри полягає у пошуку таких стратегій агентів, які максимізують їх виграші так, щоб забезпечити певний колективний баланс інтересів усіх гравців. Шукати оптимальні стратегії гравців в умовах невизна-ченості будемо за методом заохочувального навчання.

Математична модель стохастичної гри. Середні програші агентів

$$\theta_n^i\left(\left\{\tau, u_n^{Di}\right\}\right) = \frac{1}{n}\sum_t^n \xi_t^i, \forall i \in D \qquad (1)$$

характеризують якість проведення гри у момент часу n .

Метою кожного агента є мінімізація власної функції середніх програшів:

$$\lim_{n\to\infty} \theta_n^i \to \min_{\{u_n^i\}} \forall i \in D. \qquad (2)$$

Задача стохастичної гри полягає в тому, що агенти на основі спостереження поточних програшів $\{\xi_t^i\}$, повинні навчитися вибирати чисті стратегії $\{u_n^i\}$ так, щоб з ходом часу n =1,2,... забезпечити виконання системи критеріїв (2).

Для розв'язування задачі (2) необхідно визначити спосіб формування послідовностей чистих стратегій $\{u_n^i\}$ у часі, які забезпечують виконання умови (4) внаслідок асимптотичної адекватності функцій середніх виграшів (3).

Значення функції θ_n^i середніх програшів гравців наближаються до значень функцій V^i середніх програшів відповідної матричної гри:

$$\forall i \in D \ \lim_{n\to\infty} n^k M\left\{\left[\theta_n^i - V^i(p^{Di^*}\right]^2\right\} < \infty. \qquad (3)$$

Умова доповняльної нежорсткості, зважена елементами векторів змішаних стратегій, описує розв'язки гри як у змішаних, так і у чистих стратегіях:

$$diag(p_i)\left(\nabla_{pi} V_i(p) - e^{N_i} V_i(p)\right) = 0, \forall i \in D, \qquad (4)$$

де *diag(p$_i$)* – квадратна діагональна матриця порядку N_i, побудована з елементів вектора p_i.

Поточні програші агентів визначимо як штраф за порушення просторової та часової координації стратегій:

$$\xi_n^i = \lambda \sum_{s\epsilon D} \frac{\left|u_n^i - u_n^s\right|}{L_i} + (1-\lambda)\left|\overline{u_n^i} - u_{n-1}^i\right| + \mu_n, \qquad (5)$$

де $\xi_n^i \epsilon R^1; \lambda \epsilon [0;1]$ – ваговий коефіцієнт; D_i – множина сусідніх агентів, що відповідає зображеним на рис.2 зв'язкам; $L_i=|Di|$ – кількість сусідніх агентів; u_n^i– чиста стратегія з бінарним значенням; $\overline{u_n^i}$ – інверсне значення чистої стратегії; $\mu_n \sim Normal(0,d)$ – білий гауссівський шум, нормально розподілена випадкова величина з нульовим математичним сподіванням та дисперсією $d > 0$.

Перша складова виразу (5) визначає штраф за порушення просторової (взаємної) координації стратегій гравців у межах підмножини D_i, друга складова – штраф за порушення часової координації у два послідовні моменти часу, а третя складова визначає дію випадкових завад у вигляді білого шуму.

Враховуючи величину періоду $\tau = N_i$ динамічної самоорганізації МАС, формування послідовностей чистих стратегій $\{u_n^i\}$ з потрібними властивостями виконаємо на основі матриць імовірностей переходів між чистими стратегіями агентів:

$$p_n^i = \left[\begin{pmatrix} p_n^i(1.1) & \cdots & p_n^i(1,N_i) \\ \vdots & \ddots & \vdots \\ p_n^i(N_i,1) & \cdots & p_n^i(1N_i,N_i) \end{pmatrix}\right], \forall i \in D. \qquad (6)$$

Рядки матриці p_n^iє змішаними стратегіями і-го гравця, якщо він вибрав чисту стратегію $u_n^i \in U^i$. Елементи $p_n^i(j.k)$ рядків є умовними імовірностями вибору чистих стратегій залежно від поточного варіанта дії u_n^i та отриманого програшу ξ_n^i. Вважатимемо, що вибрані чисті стратегії відповідають станам агента. Тоді p_n^i (6) є матрицею імовірностей зміни станів агента.

Гра розпочинається з ненавчених змішаних стратегій $p_n^i(j.k) = \frac{1}{N_i}$ j,k = 1...N_i. Для адаптивного формування розподілу випадкових стратегій, який мінімізує функції середніх програшів (1) усіх гравців, імовірність вибору стратегій з меншими програшами повинна зростати у часі *n =1,2,...* .

Враховуючи (4), отримаємо такий рекурентний метод зміни векторів змішаних стратегій:

$$p_{n+1}^i = \pi_{\varepsilon_{n+1}}\left\{p_n^i(u_n) - \gamma_n \xi_n^i\left[e(u_n^i) - p_n^i(u_n)\right]\right\}, \forall i \in D, \qquad (7)$$

де $p_n^i(u_n)$ – змішана стратегія *i*-го гравця у стані $u_n \in U^i$; $\pi_{\varepsilon_{n+1}}$ – оператор проектування на одиничний є –симплекс $S_\varepsilon^{N_i} \subseteq S_*^{N_i}$ [11], який є підмножиною одиничного симплексу $S_*^{N_i}$;$\gamma_n > 0$ – монотонно спадна послідовність додатних величин, яка регулює величину кроку методу; ε_n>0 – монотонно спадна послідовність додатних величин, яка регулює швидкість розширення є -симплексу.

Дослідження збіжності методу (11) виконаємо у класі монотонних послідовностей $\{\gamma_n\}$ та $\{\varepsilon_n\}$ виду

$$\gamma_n = \gamma(n+\alpha)^{-\alpha}; \alpha > 0; \varepsilon_n = \varepsilon(n+\beta)^{-\beta}; \beta > 0. \qquad (8)$$

Збіжність методу (11) спостерігається:

1) з ймовірністю *1*, якщо $\alpha \epsilon (0.5;1]$; $\beta > 0$;

2) у середньоквадратичному, якщо $\alpha \epsilon (0,1]$; $\beta > 0$.

Метод (7) забезпечує адаптивний вибір агентами чистих стратегій завдяки динамічній перебудові змішаних стратегій на основі опрацювання поточних програшів.

На підставі поточного розподілу імовірностей $p_n^i(u_n{}^i)$ агент здійснює випадковий вибір чистої стратегії $\forall i \in D$.

$$u_n^i = \{\frac{u^i(l)}{l} = arg\min_l \sum_{k=1}^{l} p_n^i(j,k) > \omega(j, l = 1..N_i)\}, \quad (9)$$

де *wє[0,1]* – випадкова величина з рівномірним розподілом.

Отже, якщо в момент часу *n* агент перебуває у стані in u_n^iто на основі змішаної стратегії $p_n^i(u_n)$ він вибирає чисту стратегію $u_n^{i'}$згідно з (3), за що до моменту часу *n +1* отримує поточний програш ξ_n^i який використовує для обчислення змішаної стратегії $p_{n+1}^i(u_n)$ згідно з (7), після чого переходить у новий стан $u_{n+1}^i = u_n^{i'}$

Оцінювання ефективності ігрової самоорганізації МАС виконаємо за такими показниками:

1)функція середніх втрат або ціна гри:

$$\theta_n = \frac{1}{L}\sum_{i=1}^{L}\theta_n^i, \text{ де } L = |D| \text{ – кількість гравців;}$$

2)коефіцієнт просторової координації стратегій гравців:

$$K_n = \frac{1}{nL}\sum_{t=1}^{n}\sum_{i=1}^{L}\chi(\sum_{s\epsilon D_i}|u_t^i - u_t^s|=0), \text{ де } \chi\epsilon\{0,1\}\text{– індикаторна функція події;}$$

Моделювання прикладу. Для прикладу розглянемо стохастичну ігрову модель самоорганізації комах-світлячків (fireflies) із родини Lampyridae, які ведуть нічний спосіб життя у тропічних регіонах світу. Самці цих комах для приваблювання самок запускають механізм люмінесцентного випромінювання свого черевця. Самоорганізація проявляється у виникненні явища ритмічного синхронізованого світіння усієї колонії самців.

Моделювання поведінки світлячків виконаємо за допомогою стохастичної гри агентів, кожен з яких може перебувати в одному із двох станів $u_n^i\epsilon\{0,1\}$, де ***0*** позначає відсутність, а ***1*** – наявність світіння.

Виконаємо розв'язування стохастичної гри двох агентів з двома чистими стратегіями у середовищі з двома станами. Матриці середніх виграшів такої гри подано у таблиці.

Таблиця 1. Матриці виграшів гравців

стани	стратегії	1 агент		2 агент	
s_1	-	$\pi_2(s_1,u_2[0])$	$\pi_2(s_1,u_2[1])$	$\pi_2(s_1,u_2[0])$	$\pi_2(s_1,u_2[1])$
	$\pi_1(s_1,u_1[0])$	0.5	0.2	0.4	0.1
	$\pi_1(s_1,u_1[1])$	0.6	0.7	0.1	0.9
s_2	-	$\pi_2(s_2,u_2[0])$	$\pi_2(s_2,u_2[1])$	$\pi_2(s_2,u_2[0])$	$\pi_2(s_2,u_2[1])$
	$\pi_1(s_2,u_1[0])$	0.9	0.2	0.4	0.6
	$\pi_1(s_2,u_1[1])$	0.2	0.9	0.6	0.8

Кожен агент може спостерігати стани сусідніх агентів та змінювати власний стан так, щоб у діях бути максимально подібним на своїх сусідів. Структура зв'язків між агентами зображена на рис. 2.

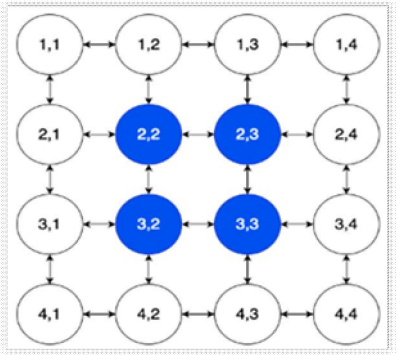

Рис. 2. Модель стохастичної гри "Світлячки"

Регулярна структура гри задається кількістю агентів $L=m*m$, $m\geq 2$, , підмножинами сусідніх агентів D_i та кількістю чистих стратегій $N_i=N=2$, $i=1..L$.

Динаміка процесу самоорганізації складається з просторової та часової координації стратегій агентів. Просторова координація полягає у дотриманні співвідношення стратегій агентів у локально визначених областях D_i, N_i так, як це зображено на рис. 2. Часова координація визначається дотриманням співвідношення стратегій агентів на проміжку часу $\tau = 2$.

В ігровій термінології просторова координація полягатиме у виборі однакових значень чистих стратегій гравців у фіксовані моменти часу (агенти намагаються повторювати дії один одного), а часова координація – у зміні бінарних стратегій на протилежні значення у два послідовні моменти часу. Результатом самоорганізації агентів є інверсна зміна матриць бінарних чистих стратегій $[0]_{m*m}$-$[1]_{m*m}$-$[0]_{m*m}$-$[1]_{m*m}$ - у часі, що моделює ритмічне світіння колонії світлячків.

Така зміна забезпечується матрицями навчених змішаних стратегій $p^{i^*}(u_n^i)=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$агентів з однаковими початковими станами.

Алгоритм розв'язування стохастичної гри.

1. Задати початкові значення параметрів:
   - $n = 0$ – початковий момент часу;
   - $L = |D|$ – кількість гравців;
   - N – кількість чистих стратегій гравців;
   - $U^i = (u^i[1], u^i[2], \dots, u^i[N])$, $i=1..L$ – вектори чистих стратегій гравців;
   - $p^i_0 = (1/N,\dots,1/N)$, $i=1..L$ – початкові змішані стратегії гравців;
   - $\gamma > 0$ – параметр кроку навчання;
   - $\alpha \in (0,1]$ – порядок кроку навчання;
   - ε – параметр ε -симплекса;
   - $\beta > 0$ – порядок швидкості розширення ε -симплекса;
   - $d > 0$ – дисперсія завад;
   - $max\ n$ – максимальна кількість кроків методу.

2. Вибрати варіанти дій $u^i_n \in U^i$, $i=1..L$ згідно з (9).

3. Отримати значення поточних програшів ξ^i_n, $i=1..L$ згідно з (5). Поточні значення гауссівського білого шуму обчислюють за формулою:

де $\omega \in [0,1]$ – дійсне випадкове число з рівномірним законом розподілу.

4. Обчислити значення параметрів γ_n, ε_n згідно з (8).

5. Обчислити елементи векторів змішаних стратегій p^i_n, $i=1..L$ згідно з (7).

6. Обчислити характеристики якості прийняття рішень Z_n (10), K_n (11).

7. Задати наступний момент часу $n := n+1$.

8. Якщо $n<n_{max}$, то перейти на крок 2, інакше – кінець.

Завдяки локальній координації стратегії агентів даний розв`язок забезпечує самоорганізацію МАС „світляків". Кожен гравець спостерігає дії сусідніх і отримує власні програші через не співпадіння, що заставляє його динамічно обирати стратегії з меншими штрафами. Динамічне обирання стратегій перетворює локально скоординовані дії гравців на глобальну координацію гри, коли колектив гравців поводяться як цілісний організм, дивись рис.3.

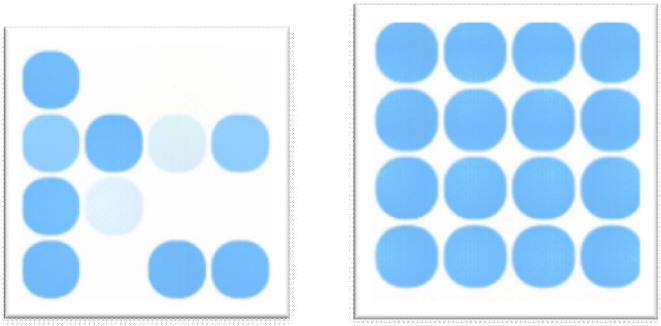

Рис. 3. а) в процесі навчання, б)самоорганізовані агенти

Висновки дослідження. Розроблена ігрова модель забезпечує динамічну самоорганізацію МАС, яка проявляється у ритмічній зміні чистих стратегій агентів, що імітує світлові ефекти колонії комах-світлячків.Характерною особливістю розглянутої ігрової самоорганізації є локально обумовлений збір інформації про стратегії поведінки сусідніх агентів, який у результаті навчання приводить до глобальної координації стратегій усіх агентів.

Генерування послідовностей чистих стратегій з потрібними властивостями забезпечується випадковим розподілом, побудованим на динамічних змішаних стратегіях гравців. Обчислення змішаних стратегій здійснюється за адаптивним рекурентним методом, отриманим на основі стохастичної апроксимації умови доповняльної нежорсткості, яка описує колективні розв'язки гри, що задовольняють умову рівноваги за Нешем.

Ефективність ігрової самоорганізації стратегій МАС вивчали за допомогою функцій середніх програшів, коефіцієнтів координації та норми відхилення динамічних змішаних стратегій від оптимальних значень. Спадання функції середніх програшів і функції відхилення змішаних стратегій, зростання коефіцієнтів координації свідчать про збіжність ігрового методу та входження МАС у режим самоорганізації. Повторення значень характеристик гри у різних експериментах з унікальними послідовностями випадкових величин підтверджує достовірність отриманих результатів.

1. Watkins, C.J.C.H. Q-Learning / C.J.C.H. Watkins, P. Dayan // Machine Learning. – Kluwer Academic Publishers, Boston. – 1992. – No. 8. – PP. 279–292.
2. Fudenberg, D. The Theory of Learning in Games / D. Fudenberg, D.K. Levine. – Cambridge, MA: MIT Press, 1998. – 292 pp.
3. Hu, J. Nash Q-learning for general-sum stochastic games / J. Hu, M. P. Wellman // Machine Learning Research. – 2003. – No. 4. – PP. 1039 – 1069.
4. Wooldridge M. An Introduction to Multiagent Systems / M. Wooldridge. – John Wiley & Sons, 2002. – 366 pp.
5. Weiss, G. Multiagent Systems. A Modern Approach to Distributed Artificial Intelligence / G. Weiss, editor. – Springer Verlag, Berlin, 1996. – 643 pp.
6. Назин А.В. Адаптивный выбор вариантов: Рекуррентные алгоритмы / А.В. Назин, А.С. Позняк. – М.: Наука, 1986. – 288 с.
7. Kaelbling, Leslie. Reinforcement learning: A survey / Leslie Kaelbling, Michael L. Littman, Andrew W. Moore. Journal of Artificial Intelligence Research. – 1996. – No. 4. – PP. 237–285.
8. Sutton, R. S. Reinforcement Learning: An Introduction / Richard S. Sutton, Andrew G. Barto. – MIT Press, 1998. – 322 pp.
9. Кравець П. О. Ігрова модель самоорганізації мультиагентних систем / П. О. Кравець // Вісник Національного університету "Львівська політехніка". Серія: Інформаційні системи та мережі : збірник наукових праць. – 2015. – № 829. – С. 161–176.
10. Musiyenko M., Zhuravska I., Kulakovska I., Kulakovska A. Simulation the behavior of robot sub swarm in spatial corridors. 2016 IEEE 36th ELNANO. April 19-21, 2016. Kyiv, Ukraine. Page(s) 382-387. DOI: 10.1109/ELNANO.2016.7493090 http://ieeexplore.ieee.org/document/7493090/