УДК 004.8:004.85:004.896

АДАПТИВНА МУЛЬТИАГЕНТНА СИСТЕМА З ДИНАМІЧНИМ ПЕРЕМИКАННЯМ СТРАТЕГІЙ У ЗМАГАЛЬНО-КООПЕРАТИВНОМУ ІГРОВОМУ СЕРЕДОВИЩІ

Падалкін К.Д.
e-mail: kostiantyn.padalkin@nure.ua
Науковий керівник – к.т.н., проф. Шевченко О.Ю.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

This paper presents an adaptive multi-agent reinforcement learning (MARL) system for a mixed cooperative-competitive game environment. The proposed model introduces the Adaptive Mode Switching (AMS) mechanism, which allows agents to dynamically shift between cooperative and competitive behavior depending on the current environment state. A custom grid-based game environment was developed using the Gymnasium API, featuring partial observability and resource control zones. The AMS meta-signal is integrated into the MADDPG reward function, enabling agents to independently learn when cooperation or competition yields greater benefit without explicit switching rules.

Мультиагентне навчання з підкріпленням (MARL) є активною областю досліджень у сфері штучного інтелекту, що знаходить застосування в робототехніці, управлінні мережами, ігровій індустрії та моделюванні складних систем [1]. Більшість існуючих підходів у мультиагентному навчанні з підкріпленням розглядають або виключно кооперативні сценарії, де агенти максимізують спільну винагороду, або виключно конкурентні, де кожен агент діє у власних інтересах [2]. Проте в реальних середовищах оптимальна стратегія часто передбачає динамічне поєднання обох режимів взаємодії залежно від ситуації. Подібні ідеї самокерованих та еволюційних систем знань розглядались у роботі [3], де запропоновано концепцію Executable Knowledge та Knowledge Computing як основу для автономного управління знаннями у динамічних середовищах.

Проблема, яку розв'язує дана робота, полягає у відсутності механізму, що дозволяє агентам автоматично адаптувати режим поведінки залежно від поточного стану середовища без явного задання правил переключення. Метою дослідження є розробка моделі адаптивної мультиагентної системи, у якій агенти самостійно навчаються обирати між кооперацією та конкуренцією на основі спостережень.

Алгоритм MADDPG (Multi-Agent Deep Deterministic Policy Gradient) [4] є одним із найпоширеніших методів для змішаних середовищ, однак передбачає фіксовану функцію винагороди, що не адаптується в процесі навчання. QMIX [5] ефективно вирішує кооперативні задачі шляхом факторизації спільної функції цінності, проте не підтримує конкурентну

взаємодію між агентами. Середовища MPE (Multi-Particle Environments) з бібліотеки PettingZoo забезпечують зручну платформу для дослідження змішаних сценаріїв, але не містять механізмів адаптивного вибору режиму поведінки. Таким чином, жоден із розглянутих підходів не реалізує навчаємий механізм динамічного переключення між кооперацією та конкуренцією в рамках одного агента.

Для дослідження розроблено ігрове середовище типу «захоплення зон» на основі Gymnasium API. Середовище являє собою сітку розміром NxN (N=10), де розташовані ресурсні зони, контроль яких приносить очки команді, та нейтральні клітини. Дві команди по два агенти діють в умовах часткової спостережуваності: кожен агент бачить лише локальну область навколо себе радіусом r=2. Формально середовище описується кортежем (S, A, T, R, O), де S – простір станів, A – простір дій, T – функція переходу, R – функція винагороди, O – простір спостережень агента.

Ключовим нововведенням запропонованої моделі є механізм Adaptive Mode Switching (AMS). Кожен агент має навчаємий мета-сигнал – скалярне значення в діапазоні від нуля до одиниці, що визначає його поточний режим взаємодії. Близьке до нуля значення відповідає конкурентному режиму, коли агент максимізує власну індивідуальну винагороду. Близьке до одиниці – кооперативному, коли агент орієнтується на максимізацію спільної командної нагороди. Сумарна винагорода агента визначається як зважена сума індивідуальної та командної складових, де ваги задаються саме цим мета-сигналом. Таким чином, агент навчається не лише обирати дії, а й вибирати режим взаємодії з іншими агентами.

Мета-сигнал є додатковим виходом мережі актора і навчається спільно з основною політикою агента в рамках алгоритму MADDPG. Кожен агент реалізовано як мережу актор-критик. Актор отримує на вхід локальне спостереження і генерує дію та мета-сигнал. Критик є централізованим і під час навчання має доступ до станів та дій усіх агентів, що відповідає парадигмі CTDE (Centralized Training, Decentralized Execution). Це дозволяє критику враховувати поведінку всіх учасників при оцінці Q-значень, тоді як виконання залишається децентралізованим – агент приймає рішення лише на основі власних спостережень.

Для оцінки ефективності запропонованої моделі проводиться порівняльний експеримент між трьома конфігураціями: запропонована адаптивна модель AMS з MADDPG, MADDPG з фіксованим кооперативним режимом, MADDPG з фіксованим конкурентним режимом. Основними метриками є сумарна командна винагорода за епізод, відсоток контрольованих ресурсних зон наприкінці епізоду та динаміка мета-сигналу протягом навчання. Навчання проводиться локально: 500 епізодів по 200 кроків, розмір батчу 64, буфер відтворення 100 000 переходів.

Очікується, що адаптивна модель перевершить обидва фіксовані базові рівні в умовах динамічно змінюваного середовища, оскільки агенти

здатні ситуативно обирати оптимальний режим взаємодії. Зокрема, передбачається, що на ранніх стадіях навчання домінуватиме конкурентний режим, а в міру накопичення досвіду агенти навчаться переключатися на кооперацію у сценаріях, де суперник контролює більшість зон. Аналіз динаміки мета-сигналу дозволить виявити закономірності у виборі режиму та підтвердити, що переключення відбувається осмислено, а не випадково.

У роботі запропоновано модель адаптивної мультиагентної системи, що реалізує динамічне переключення між кооперативною та конкурентною поведінкою агентів у змагальному ігровому середовищі. Розроблено оригінальне ігрове середовище «захоплення зон» на базі Gymnasium API з частковою спостережуваністю. Запропонований механізм AMS, інтегрований в алгоритм MADDPG, дозволяє агентам автоматично адаптувати режим взаємодії без явного задання правил переключення. Наукова новизна роботи полягає у введенні навчаємого мета-сигналу, що параметризує функцію винагороди агента і забезпечує безперервний перехід між режимами поведінки. На відміну від існуючих підходів, де режим взаємодії фіксується на етапі проектування, запропонований механізм дозволяє агентам самостійно формувати стратегію на основі поточного контексту. Практична цінність полягає у застосуванні моделі для задач розподіленого управління, де учасники балансують між власними інтересами та колективною ефективністю. Це підкреслює актуальність розробки механізмів, що дозволяють агентам автоматично адаптувати режим поведінки залежно від поточного стану середовища без явного задання правил переключення. Подальші дослідження передбачають розширення моделі на більшу кількість агентів та тестування в середовищах з вищим ступенем невизначеності.

Список використаних джерел:

1. Sutton R. S., Barto A. G., Bach F. Reinforcement Learning: An Introduction. MIT Press, 2018. 552 c.

2. Hernandez-Leal P., Kartal B., Taylor M. E. A survey and critique of multiagent deep reinforcement learning. Autonomous Agents and Multi-Agent Systems. 2019. Т. 33, № 6. С. 750–797. DOI: https://doi.org/10.1007/s10458-019-09421-1.

3. Terziyan V., Shevchenko O., Golovianko M. An introduction to knowledge computing. Eastern-European Journal of Enterprise Technologies. 2014. Vol. 1, no. 2. P. 27–40. DOI: https://doi.org/10.15587/1729-4061.2014.21830.

4. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments / R. Lowe et al. Proceedings of the Conference on Neural Information Processing Systems (NeurIPS). 2017.

5. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning / T. Rashid et al. Proceedings of the International Conference on Machine Learning (ICML). 2018. С. 4295–4304.