Инженерный вестник Дона, №3 (2026)
ivdon.ru/ru/magazine/archive/n3y2026/10793

Обзор исследований по интеграции эволюционной теории игр и мультиагентного обучения с подкреплением

Чжу Хан, М.С. Селезнева

Московский государственный технический университет им. Н.Э. Баумана, Россия, Москва

Аннотация: В работе представлен обзор исследований по интеграции эволюционной теории игр (ЭТИ) и мультиагентного обучения с подкреплением (Multi-Agent Reinforcement Learning - MARL). Проанализированы основные проблемы MARL и соответствующие преимущества ЭТИ. В результате анализа установлено, что внедрение ЭТИ позволяет эффективно решить проблемы нестабильности, распределения кредитов и частичной наблюдаемости в MARL, обеспечивая стабильную стратегическую конвергенцию и новый путь для групповой оптимизации. Показано, что интеграция ЭТИ и MARL формирует перспективную теоретическую и техническую основу для прорыва в технологиях мультиагентного управления. Наряду с этим, для глубокого слияния двух направлений в будущем предстоит оптимизировать механизмы интеграции, разработать более надежные алгоритмы и укрепить прикладные исследования в сложных гетерогенных системах.
Ключевые слова: эволюционная теория игр, мультиагентное обучение с подкреплением, мультиагентное управление, нестабильность, распределение кредитов, частичная наблюдаемость.

Введение

В последние годы стремительное развитие глубокого обучения способствовало тому, что глубокое обучение с подкреплением продемонстрировало мощные возможности в обработке высокомерных состояний пространства, что в свою очередь породило актуальную исследовательскую тему — мультиагентное обучение с подкреплением (Multi-Agent Reinforcement Learning - MARL) [1,2]. MARL позволяет нескольким агентам, обладающим способностями к самостоятельному восприятию и принятию решений, взаимодействовать и обучаться в общей среде, генерируя глобальное совместное поведение через локальные взаимодействия. Без необходимости в точной системной модели оно предоставляет эффективные решения для крупномасштабных сложных задач, с которыми одиночные агенты не справляются — например, для

транспортного управления, интеллектуальных электросетей, кооперации много роботов, автономного вождения и других сфер.

Начиная с AlphaGo от DeepMind, превосходящего людей в игре с полной информацией, и заканчивая DeepNash, достигшим экспертного уровня в игре Stratego с неполной информацией, теория и практическое применение MARL привели к значительным прорывам [3]. Однако, когда MARL переходит от контролируемой виртуальной среды к открытой и динамичной реальной среде, присущие ей дилеммы, такие как нестабильность, трудности с распределением кредитов и частичная наблюдаемость, становятся все более очевидными, что серьезно ограничивает практический процесс применения технологии мультиагентного управления.

В этом контексте эволюционная теория игр (ЭТИ), зародившаяся в биологии и фокусирующаяся на эволюционном законе взаимодействия рациональных конечных групп особей, стала междисциплинарным и многообещающим способом решения потенциальных проблем MARL. В данной работе основное внимание уделено исследованию ключевых вызовов, с которыми сталкивается MARL в сфере мультиагентного управления, при этом проводится углубленный анализ выполнимости и необходимости решения этих проблем. В заключение анализируется возможность проведения интегрированных исследований, а также предлагаются новые подходы для развития соответствующих технологий.

Ключевые проблемы MARL

Хотя MARL имеет широкие перспективы, его применение в мультиагентном управлении сталкивается с рядом серьезных проблем. Эти проблемы взаимосвязаны и в совокупности ограничивают стабильность, эффективность и практичность алгоритмов.

1. Нестационарность

Нестационарность является одной из ключевых проблем MARL и важным фактором, ограничивающим стабильность мультиагентного управлении [4]. В одиночной агентной постановке динамика окружающей среды стабильна, но в мультиагентной системе политика каждого агента постоянно обновляется в процессе обучения. Синхронное обучение и взаимовлияние агентов нарушают гипотезу Марковского свойства, на которой основаны большинство алгоритмов усиленного обучения для одиночных агентов, приводя к нестабильности процессов оптимизации и обучения.

Современные методы MARL в основном полагаются на градиентный спуск для обновления параметров политики, однако в мультиагентной среде для отдельного агента окружающая среда является высоконестационарной. Эта нестабильность приводит к тому, что агент легко приближается к локальной оптимальности или колеблется и расходится [5].

2. Распределения кредитов

При совместном выполнении задач группой агентов общий доход является результатом совместных действий всех агентов и не может быть напрямую разделен между отдельными агентами. Как точно приписать глобальный успех или неудачу поведению каждого индивидуального агента — то есть проблема распределения кредитов — является ядровым вызовом. Эта неопределенность в распределении кредитов серьезно снижает эффективность обучения и даже приводит к таким неэффективным поведениям агентов, как «паразитизм», разрушая результаты группового сотрудничества [6].

Хотя некоторые улучшенные алгоритмы, например, COMA и MAPPO [7], пытаются смягчить эту проблему с помощью механизмов распределения

кредитов, их вычислительная сложность остается высокой, и трудно гарантировать существование глобального оптимального решения.

COMA пытаются организовать сотрудничество агентов через совместное обучение, но требуют явной моделирования пространства политик каждого агента, что может стать чрезвычайно сложным при работе с крупномасштабными мультиагентными системами [8].

Проблема переобучения политики, вызванная общими функциями преимущества в MAPPO: это наиболее критическая проблема. Все агенты используют одну глобальную функцию преимущества для обновления политики, что вводит их в заблуждение при обучении. Например, пассивный агент может «паразитировать» на отличном выступлении напарников и не выучить правильную политику [9].

3. Частичная наблюдаемость

В большинстве мультиагентных сценариев агенты не могут получать состояние других агентов и связанную с ними информацию, что дополнительно усложняет оптимизацию политик. Частичная наблюдаемость возникает, когда агенты не могут получить полное состояние среды и вынуждены полагаться на ограниченную или зашумленную наблюдаемую информацию. Это ограничение требует разработки алгоритмов, способных выводить скрытую информацию о состоянии или принимать надежные решения в условиях неопределенности [10]. Когда не удается получить достаточного количества динамической информации, эффективность глубокого обучения с подкреплением резко снижается [11].

Помимо проблем со стабильностью, распределением кредитов и частичной наблюдаемостью, MARL также сталкивается с многочисленными проблемами: дилемма разведки и использования усугубляется

нестабильностью окружающей среды, и агенту необходимо сбалансировать индивидуальную оптимизацию и командную разведку, чтобы избежать стратегического вмешательства [12]. Координация связи ограничена полосой пропускания, задержками и помехами, и при разработке эффективного протокола необходимо учитывать получение информации и накладные расходы [13]. Узкое место расширяемости связано с экспоненциальным ростом пространства совместной стратегии, вызванным увеличением числа агентов, что привело к резкому росту вычислительных и коммуникационных затрат [14]. Недостаточная надежность и защищенность делают его уязвимым для враждебных атак и нарушений окружающей среды, а также затрудняют соблюдение требований безопасности, связанных с фактическим развертыванием [15]. Эти проблемы взаимосвязаны, что еще больше затрудняет широкомасштабное применение MARL в сложных сценариях.

Анализ интеграции ЭТИ с мультиагентным усиленным обучением

Ключевые проблемы современного MARL по существу обусловлены отсутствием эффективного описания взаимодействия агентов. Это приводит к взаимному вмешательству в процессы обновления политик, а тренировка часто застревает в колебаниях или субоптимальных решениях. Одновременно механизмы совместной эволюции группового поведения трудно смоделировать в существующих рамках, а проблема распределения кредитов лишена системного учета влияния долгосрочных взаимодействий. Следовательно, настоятельно требуется введение теоретического инструмента, способного описывать динамическую эволюцию политик агентов, чтобы достичь глубокого понимания и управления механизмами формирования совместного поведения мультиагентных систем.

ЭТИ успешно объяснила ряд явлений биологической эволюции. Классическая работа Smith [16] использовала ее для объяснения поведения

животных в борьбах, а также представила концепцию эволюционно стабильной стратегии (Evolutionarily Stable Strategy - ESS). Происходящая из биологии, ЭТИ фокусируется на законов эволюции политик ограниченно рациональных индивидов в долгосрочных взаимодействиях, предлагая целенаправленные решения для вышеупомянутых проблем. Она изучает, как частотное распределение различных поведенческих стратегий в популяции динамически эволюционирует во времени через механизмы выбора, мутации и т.д. Её ядровые концепции — ESS и репликаторная динамика.

В отношении ключевых проблем MARL, рассмотренных ранее, ЭТИ предлагает совершенно другой набор аналитических инструментов и подходов к решению.

1. Нестационарности

ЭТИ через концепцию ESS предоставляет стабильную цель конвергенции политик для мультиагентных систем. ESS — это стратегия, при которой, если большинство индивидов в группе ее используют, любая мутационная стратегия не может вторгнуться, обеспечивая долгосрочную стабильность групповой политики. Внедрение ESS в MARL позволяет смягчить нестационарность окружающей среды, вызванную частыми колебаниями индивидуальных политик, путем ограничения направления эволюции групповых политик, а также повысить стабильность контроля системы.

В работе [17] доказали, что эволюционные стратегии демонстрируют отличные результаты при нестационарности и частичной наблюдаемости. Поскольку они постоянно используют и развивают группу агентов, а не одного агента; оптимизация - это оптимальное распределение нескольких решений, а не одного оптимального решения, поэтому при наличии достаточного количества параметров эволюционная стратеги я обладает большей надежностью, чем модель, обученная методом градиентного спуска.

Исследования [18] показали, что в механизмах эволюционного сотрудничества при повторяемых взаимодействиях в эволюционных и стохастических играх взаимность и обратная связь от окружающей среды могут значительно усилить склонность к сотрудничеству. В мультиагентной среде агенты периодически корректируют свои политики в процессе взаимодействия друг с другом, а эволюционный подход является эффективным способом моделировать эти взаимодействия, тем самым улучшая исследование политик.

2. Распределения кредитов

ЭТИ через матрицу полезности определяет обратную связь по результатам индивидуальных взаимодействий, напрямую связывая индивидуальную выгоду с результатами группового сотрудничества. Это позволяет индивидам точно оценивать ценность собственного поведения через восприятие собственной выгоды от взаимодействий, предоставляя четкую обратную связь для обновления политик и фундаментально оптимизируя механизм распределения кредитов. Кроме того, механизм популяционной эволюции ЭТИ может эффективно снизить сложность пространства политик в крупномасштабных мультиагентных системах, обеспечивая эффективную оптимизацию групповых политик через селекцию популяции и итерацию политик, а также повышая масштабируемость мультиагентного контроля.

В работе [19] установили, что алгоритм Novelty Search может обрабатывать случаи с обманчивой функцией обратной связи и редкими наградами, сохраняя при этом одинаковую масштабируемость.

3. Частичная наблюдаемость

ЭТИ основывается на ключевом предположении «ограниченной рациональности», не требуя от агентов полной информации или сверхъестественной вычислительной способности. Агенты лишь должны корректировать свое поведение на основе локальных наблюдений и эмпирического подражания, что делает теоретический фреймворк естественно соответствующим границам возможностей реальных агентов. Ввиду проблемы, заключающейся в том, что один агент не может наблюдать за глобальным состоянием, динамика распределения групповых стратегий, на которой сосредоточена ЭТИ, обеспечивает механизм агрегирования макроинформации - даже если индивиды владеют только локальной “фрагментированной” информацией, они могут распространять и имитировать стратегии, чтобы неявно вносить изменения в групповое распределение, закодированную и распространяемую информацию об адаптации к окружающей среде.

EMARL [20] сочетает генетический алгоритм с COMA: сначала генетический алгоритм оптимизирует индивидов популяции, а затем дополнительно усиливает градиент политики на оптимизированной популяции. Оценка алгоритма EMARL на задачах кластеризации показала, что он демонстрирует лучшие результаты по сравнению с базовыми алгоритмами MARL.

RACE [21] предлагает новый гибридный фреймворк, внедряющий концепцию совместного представления в интеграцию MARL и эволюционных алгоритмов. RACE делит групповую политику на совместный наблюдательный энкодер и независимые линейные представления политик, максимизируя взаимную информацию между функцией ценности и восприятой ценностью, а также внедряя информацию, связанную с сотрудничеством, и превосходное глобальное состояние в совместное представление. Операции скрещивания и мутации на уровне агентов

выполняются над линейными представлениями для обеспечения стабильной эволюции. В сравнении с FACMAC [22], MATD3 [23] и MERL [24], RACE добился выдающихся результатов на задачах SMAC и MA-MuJoCo.

В работе [25] использовали эволюционные методы обучения в играх для анализа игровой динамики различных подходов к мультиагентному обучению с подкреплением и выявили глубокую связь между ЭТИ и указанными методами. В работе [24] благодаря введению эволюционной динамики исследование стратегий рассматривается процесс групповой эволюции, который позволяет избежать многомерных катастрофных ситуаций при большом количестве агентов и привести децентрализованную систему к стабильному равновесию. В работе [26] предполагается, что фреймворки ЭТИ могут естественным образом моделировать долгосрочные отношения сотрудничества и конкуренции между агентами, способствовать стратегическому разнообразию через механизмы, такие как динамика репликаторов, а также помогать алгоритмам достигать более надежного группового взаимодействия в сложных динамичных средах.

Приведенные выше доводы доказывают, что эволюционная теория игр способна эффективно решать проблемы, с которыми сталкивается MARL. На основе трех ключевых парадигм ЭТИ — перспективы групповой динамики, ограниченной рациональности и эволюционно стабильной стратегии — на рис.1 предлагаются ключевые стадии интеграции ЭТИ и MARL.

Рис. 1. - Ключевые стадии интеграции ЭТИ и MARL

1. Формулировка задачи и определение целей

ЭТИ рассматривает мультиагентную систему в целом как экосистему эволюции стратегий, при этом основная цель заключается в сходимости системы к эволюционно стабильной стратегии. На уровне проектирования структуры игры взаимодействие между агентами можно смоделировать как повторяющуюся игру, а функцию полезности определить, как «приспособленность», характеризующую выживаемость стратегии. Цели проектирования системы должны сосредоточиться на динамической сходимости и устойчивости к возмущениям, а не ограничиваться единственным показателем оптимальности производительности.

2. Представление и инициализация стратегий

Стратегия является материалом для эволюционного процесса, при этом её начальное состояние должно гарантировать достаточное разнообразие. ЭТИ изменяет парадигму традиционного MARL «обучение оптимальной стратегической сети для отдельного агента» на новый модуль «поддержание и итерация популяции стратегий для каждого агента». Данная парадигма

позволяет эффективно решить ключевые проблемы MARL, такие как преждевременная сходимость и недостаточная способность к исследованию, предоставить богатый материал для последующей конкуренции и итеративного обновления стратегий, а также составить ядровую основу интегрированного алгоритма.

3. Оценка и улучшение стратегий

Распространение стратегий моделируется с помощью «динамики репликатора»: вероятность имитации стратегий с высокой приспособленностью в популяции возрастает. Правила обучения заменяются эволюционными правилами: агенты не вычисляют градиенты напрямую, а с определенной вероятностью переключаются на наблюдаемые стратегии с более высокой прибылью, сравнивая собственный выигрыш со средним выигрышем по группе. Это естественным образом решает проблему нестационарности, реализует полностью распределенное обучение и лучше подходит для сценариев с частичной наблюдаемостью.

4. Распределение кредитов

Для распределения кредитов используется относительная приспособленность, и определяется функция преимущества на основе популяции: преимущество сравнивается не с собственными ожиданиями агента, а со средним значением по популяции или с конкретной базовой стратегией. Данный подход находится в прямой связи с ядровой идеей контрфактических базисов, предоставляя естественное решение для проблемы распределения заслуг в MARL на основе показателей работы популяции.

5. Оценка сходимости и устойчивости

ESS принимается как ядровой критерий сходимости интегрированного алгоритма. В процессе обучения модели или после его завершения можно проверить, удовлетворяет ли текущая комбинация стратегий условию ESS, посредством «теста на инвазию» — то есть при введении небольшого количества произвольных новых стратегий, эти новые стратегии не могут распространяться и закрепиться в популяции. По сравнению с традиционным равновесием Нэша, ESS обеспечивает более высокий уровень устойчивости для MARL, который не только способен сопротивляться небольшим мутациям индивидуальных стратегий, но и позволяет системе поддерживать стабильные показатели производительности управления в динамически изменяющейся среде.

Таким образом, ЭТИ открывает многообещающий путь для решения основных проблем мультиагентного управления. Это позволяет не только эффективно решать проблемы нестабильности и стратегической нестабильности, с которыми сталкивается MARL, но и повышать надежность и адаптивность системы, внедряя идею биологической эволюции.

Заключение

Хотя MARL является основным инструментом для реализации группового интеллекта, оно сталкивается с фундаментальными проблемами, такими как нестабильность, распределение кредитов и частичная наблюдаемость, которые ограничивают его применение в сложных сценариях реального мира. Внедрение ЭТИ может не только целенаправленно решить вышеупомянутые основные проблемы, но и обеспечить стабильную стратегическую цель конвергенции, точный механизм обратной связи по доходам и эффективный путь групповой оптимизации для мультиагентного управления.

В будущем, чтобы способствовать глубокой интеграции ЭТИ и MARL, необходимо дополнительно оптимизировать механизм их интеграции, например, разработать более надежный алгоритм MARL, основанный на ЭТИ, для повышения способности алгоритмов к обобщению в динамических системах. В то же время нужно укреплять прикладные исследования в сложных сценариях, таких как гетерогенные мультиагентные системы и гибридные системы совместной работы человека и компьютера. Интеграция ЭТИ и MARL обеспечит новую теоретическую и техническую поддержку для прорывов в технологии мультиагентного управления и будет способствовать ее широкомасштабному применению в более ключевых областях инженерии.

Литература (References)

1. Sun Changyin, Mu Chaoxu. Important scientific problems of multi-agent deep reinforcement learning // Acta Automatica Sinica. 2020. Vol. 46, No. 7. pp. 1301–1312. DOI: 10.16383/j.aas.c200159.

2. Luo Biao, Hu Tian-Meng, Zhou Yu-Hao, Huang Ting-Wen, Yang Chun-Hua, Gui Wei-Hua. Survey on multi-agent reinforcement learning for control and decision-making // Acta Automatica Sinica. 2025. Vol. 51, No. 3. pp. 510−539. DOI: 10.16383/j.aas.c240392.

3. Perolat J., et al. Mastering the game of Stratego with model-free multiagent reinforcement learning // Science. 2022. Vol. 378, No. 6623. pp. 990–996.

4. De La Fuente N., Noguer i Alonso M., Casadellà G. Game Theory and Multi-Agent Reinforcement Learning: From Nash Equilibria to Evolutionary Dynamics // arXiv. 2024. DOI: 10.48550/arXiv.2412.20523.

5. Zhang Rong-Rong, Guo Lei. Controllability of Nash equilibrium in game-based control systems // IEEE Transactions on Automatic Control. 2019. Vol. 64, No. 10. pp. 4180–4187.

6. Vezhnevets A. S., Osindero S., Schaul T., et al. FeUdal networks for hierarchical reinforcement learning // Proceedings of the 34th International Conference on Machine Learning (ICML). Sydney, Australia, 2017. pp. 3540–3549.

7. Lohse O., Pütz N., Hörmann K. Implementing an Online Scheduling Approach for Production with Multi Agent Proximal Policy Optimization (MAPPO) // Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems. 2021. pp. 586–595. DOI: 10.1007/978-3-030-85914-5_62.

8. Foerster J., Farquhar G., Afouras T., Nardelli N., Whiteson S. Counterfactual multi-agent policy gradients // Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18). New Orleans, Louisiana, USA, 2018. pp. 2974–2982.

9. Wang S., Chen W., Hu J., Hu S., Huang L. Noise-Regularized Advantage Value for Multi-Agent Reinforcement Learning // Mathematics. 2022. Vol. 10, No. 15. pp. 2728. DOI: 10.3390/math10152728.

10. Oliehoek F. A., Amato C. A Concise Introduction to Decentralized POMDPs. Springer International Publishing, 2016. – 130 p.

11. Hausknecht M., Stone P. Deep recurrent Q-learning for partially observable MDPs // Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, Texas, USA, 2015. pp. 4012–4018.

12. Hu Junling, Wellman Michael P. Multiagent reinforcement learning: Theoretical framework and an algorithm // Proceedings of the Fifteenth International Conference on Machine Learning. Madison, Wisconsin, USA, 1998. pp. 242–250.

13. Foerster J. N., Assael Y. M., de Freitas N., Whiteson S. Learning to communicate with deep multi-agent reinforcement learning // Proceedings of the

30th International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016. pp. 2145–2153.

14. Bernstein D. S., Givan R., Immerman N., Zilberstein S. The complexity of decentralized control of Markov decision processes // Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence. Stanford, California, USA, 2000. pp. 32–37.

15. Amato C., Bernstein D. S., Zilberstein S. Decentralized control of partially observable Markov decision processes // IEEE Conference on Decision and Control. 2013. pp. 2398–2405.

16. Smith J. M., Price G. R. The logic of animal conflict // Nature. 1973. Vol. 246. pp. 15–18.

17. Liu Zexi, Chen Ben M., Zhou Hongyu, et al. MAPPER: Multi-agent path planning with evolutionary reinforcement learning in mixed dynamic environments // 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas, NV, USA, 2020. pp. 11748–11754.

18. Hilbe C., Šimsa Š., Chatterjee K., Nowak M. A. Evolution of cooperation in stochastic games // Nature. 2018. Vol. 559. pp. 246–249.

19. Conti E., Madhavan V., Such F. P., et al. Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents // Advances in Neural Information Processing Systems 31. Montréal, Canada, 2018. pp. 5032 – 5043.

20. Guo Y., Xie X., Zhao R., Zhu C., Yin J., Long H. Cooperation and competition: Flocking with evolutionary multi-agent reinforcement learning // International Conference on Neural Information Processing. 2022. pp. 271–283.

21. Li P., Hao J., Tang H., Zheng Y., Fu X. Race: Improve multi-agent reinforcement learning with representation asymmetry and collaborative evolution // Proceedings of the 40th International Conference on Machine Learning. Honolulu, Hawaii, USA, 2023. pp. 19490-19503.

22. Peng B., Rashid T., de Witt C. A. S., et al. FACMAC: Factored multi-agent centralised policy gradients // Advances in Neural Information Processing Systems 34. 2021. pp. 12208 – 12221.

23. Laasner R., Du X., Tanikanti A., et al. MatD³: A Database and Online Presentation Package for Research Data Supporting Materials Discovery, Design, and Dissemination // Journal of Open Source Software. 2020. Vol. 5, No. 45. pp. 1945. DOI: 10.21105/joss.01945.

24. Khadka S., Majumdar S., Miret S., McAleer S., Tumer K. Evolutionary reinforcement learning for sample-efficient multiagent coordination // Proceedings of the 37th International Conference on Machine Learning. 2020. Vol. 119. pp. 6651–6660.

25. Bloembergen D., Tuyls K., Hennes D., Kaisers M. Evolutionary dynamics of multi-agent learning: A survey // Journal of Artificial Intelligence Research. 2015. Vol. 53. pp. 659–697.

26. Wang R., Dong Q. Multiagent game decision-making method based on the learning mechanism // Chinese Journal of Engineering. 2024. Vol. 46, No. 7. pp. 1251–1268. DOI: 10.13374/j.issn2095-9389.2023.08.08.003.

Авторы согласны на обработку и хранение персональных данных.

Дата поступления: 6.01.2026

Дата публикации: 24.02.2026