Применение обучения с подкреплением в многопользовательских стратегических играх

Мисник А.С.
misnik.as@phystech.edu

Центр когнитивного моделирования МФТИ

2025 – Научно-исследовательский семинар

Мотивация

- Получение структурного представления эталонных стратегий в обучении с подкреплением
- Обобщение полученного результата - формирование набора человекочитаемых правил
- Масштабирование алгоритма ВКФ на многоагентные среды и среды со сложной структурой состояния

Классическая постановка задачи RL

S - пространство состояний среды, A - пространство действий
$P(s'|s,a)$ - вероятности переходов состояний при выполнении действия
$R(s,a)$ - награда за выполнение действия a в состоянии s
$\gamma \in [0,1]$ — коэффициент дисконтирования
Цель агента в классическом RL - нахождение оптимальной стратегии $\pi(a|s)$, определяющей выбор действий в состояниях:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \right]$$

Постановка задачи в рамках работы

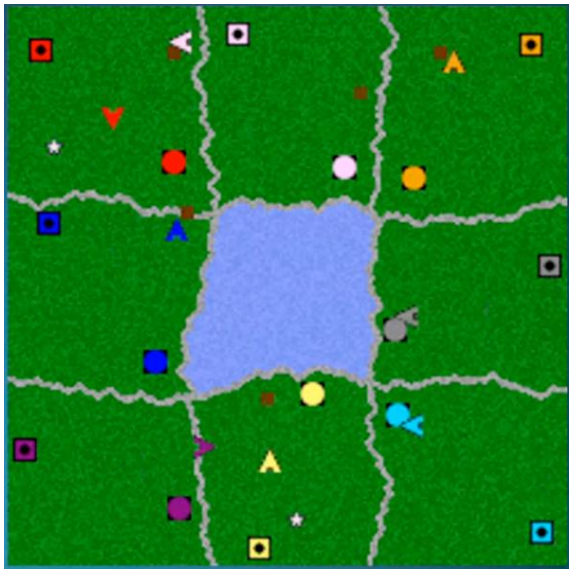
Пусть дополнительно задано отображение признакового описания состояний $F : S \to \{0,1\}^n$, которое каждому состоянию $s \in S$ ставит в соответствие бинарный вектор признаков $F(s) = (f_1(s), f_2(s), ..., f_n(s))$, где $f_i : S \to \{0,1\}$
Цель агента в модели задачи - порождение множества гипотез H $(\forall h_i \in H : h_i = \{F_i, A_i\}, F_i \subset \{0,1\}^n)$, где каждая гипотеза должна удовлетворять свойству согласованности:

$$\forall s \in S : F(s) \in F_h \Rightarrow A_h \sim \pi(\cdot|s)$$

где $\pi(\cdot|s)$ — оптимальная стратегия агента в состоянии s

Модель рассматриваемой игры

1. Игра состоит из тактов, все агенты совершают действия одновременно
2. На каждом из тактов с некоторой вероятностью появляется ресурс, за переработку которого агент получает награду
3. После переработки ресурса появляются отходы, которые возможно утилизировать - их количество влияет на вероятность появления новых ресурсов
4. Итоговое вознаграждение - количество ресурсов, переработанных за время игры



Подзадачи работы

- Разбор нейросетевых подходов и их особенностей - DONE
- Создание утилиты для визуализации последовательности фреймов - DONE
- Интеграция ВКФ в среду игры - DONE
- Эксперименты с порождением гипотез и сравнение результатов - IN PROGRESS

Нейросетевые эксперименты (VDN)

VDN[1] - алгоритм кооперативного обучения с подкреплением, основанный на разложении общей награды в сумму индивидуальных наград.

Проблемы рассмотренного подхода:

- VDN действует в предположении, что все агенты управляются пользователем
- Сформированные стратегии являются сугубо кооперативными

[1] P Sunehag et al., "Value-decomposition networks for cooperative multi-agent learning", 2017

Нейросетевые эксперименты (MAPPO)

MAPPO[2] - адаптация алгоритма Proximal Policy Optimization для многоагентного обучения.

Особенности:

- Требует больше итераций обучения (и хуже параллелится) в сравнении с VDN
- Требует тонкой настройки гиперпараметров

[2] C Yu et al., "The surprising effectiveness of ppo in cooperative multi-agent games", 2022

Алгоритм ВКФ

Шаг алгоритма:

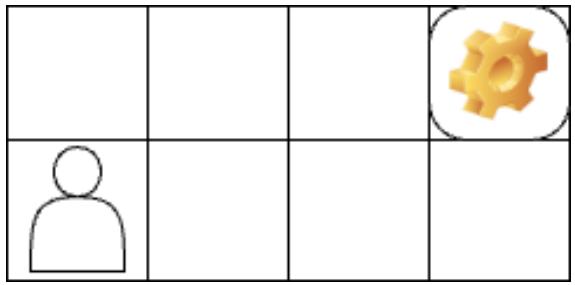
1. Породить гипотезу-кандидата с использованием цепи Маркова
2. Для каждого из состояний, затрагиваемых в гипотезе, проверить оптимальность действия (используя результат симуляции MCTS для этого состояния)
3. Подтвердить гипотезу, если действие оптимально для всех затрагиваемых состояний

Процесс алгоритма:

1. Породить N подтвержденных гипотез, используя шаги алгоритма
2. Вычислить эмпирический риск применения порожденных гипотез
3. Повторить шаги 1 и 2 с увеличенными N, пока эмпирический риск не перестанет уменьшаться

Преимущество ВКФ

Сбор структурной информации о состояниях позволяет агенту мыслить в терминах "сценариев", которые однозначно разбиваются в последовательность действий (алгоритмами на графах).



Пример сценария, который разбивается в последовательность (UP R R R PICK)

Проблемы ВКФ

- Сложность формирования набора бинарных параметров состояний среды, достаточных для обучения
- Существование состояний, не покрытых гипотезами
- Классическая реализация алгоритма не подразумевает параллельного обучения агентов

Адаптация ВКФ для сложных сред

- Возможность формирования стартовых подсказок для ВКФ с помощью данных, предобработанных LLM
- Расширение Encoder для ВКФ - формирование вспомогательных признаков состояния на основе ответа LLM о данных состояния
- Изоляция Experience Buffer между параллельно обучающимися агентами для возможности многоагентном обучении

Спасибо за внимание!

misnik.as@phystech.edu