

РЕШЁТОЧНОЕ ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ

Д.В. Виноградов (*vinogradov.d.w@gmail.com*)

Федеральный исследовательский центр
«Информатика и управление» РАН, Москва

Работа посвящена изложению теоретико-решеточного подхода к порождению правил, образующих (суб-)оптимальную стратегию в парадигме обучения с подкреплением. Мы аргументируем за возврат к использованию метода Монте-Карло с поиском по дереву. Вероятностно-комбинаторный формальный метод, основанный на теории решеток, устраняет основной недостаток метода Монте-Карло – отсутствие обобщающей способности. Будут обсуждаться проблемы широко применяемых сейчас нейросетевых подходов и указаны достоинства метода Монте-Карло. Наконец, с использованием теории категорий будет представлена строгая формализация предлагаемого подхода.

Ключевые слова: обучение с подкреплением, метод Монте-Карло, решётки, теория категорий.

Введение

Обучение с подкреплением (см., например, [Саттон и др., 2020]) является одной из парадигм машинного обучения, нацеленной на порождение (суб-)оптимальной стратегии поведения агента (агентов в многоагентной постановке) в вероятностно изменяемой среде на основании награды (и наказаний). Стандартной математической моделью для обучения с подкреплением является Марковский процесс принятия решений – Markov Decision Process. В зависимости от выбора действий агента и случайного ответного поведения среды возникает дерево возможных траекторий.

Основными проблемами для возможности эффективного нахождения (суб-)оптимальной стратегии являются

1. Накопительный характер оценок качества действий в промежуточных состояниях, когда окончательная оценка возникает только по окончании всей траектории.
2. Экспоненциально большое число состояний в дереве траекторий.

С 2015 года в обучении с подкреплением доминирующим является нейросетевой подход, когда нейростеть того или иного вида аппроксимирует или оценки действий (с последующим применением жадного алгоритма выбора действия в наблюдаемом состоянии), или даже напрямую функцию выбора действий.

Ранее для оценок качества действий обучающегося агента использовалась классическая техника Монте-Карло, основанная на поиске в дереве (МКПД) – Monte Carlo Tree Search [Świechowski et al., 2023]. В ней для оценки значения действия в текущем состоянии мы вычисляем награды агента посредством усреднения вознаграждений для некоторого (статистически значимого) количества частичных траекторий, запущенных из текущего состояния с помощью выбранного действия.

Смена парадигмы с МКПД на глубокие нейросети произошла в 2014 году под влиянием впечатляющих успехов систем AlphaGo и AlphaStar фирмы DeepMind.

Теоретическим обоснованием этого перехода является отсутствие у МКПД обобщающей способности, когда для оценки ранее ненаблюдавшегося действия необходимо заново запускать поиск по дереву. В то же время нейросети, как универсальные аппроксиматоры, всегда обеспечивают оценку (возможно, очень плохую) для любого действия!

Однако МКПД имеет и свои неоспоримые достоинства:

1. Он обеспечивает контролируемый уровень дисперсии оценки в каждой оцениваемой точке, а не только оптимизацию единственной глобальной функции потерь, используемой для обучения нейросети.
2. Он работает напрямую с состояниями и действиями агента, а не с их векторными представлениями в скрытых слоях нейросети.

Для добавления обобщающей способности к МКПД предлагается использовать вероятностно-комбинаторный формальный метод (ВКФ-метод) (см., например, [Виноградов, 2022]).

Нейросетевой подход, доминирующий в настоящее время, не лишён существенных недостатков, среди которых следует упомянуть:

1. Невозможность предварительно учитывать правила, сформулированные экспертами.
2. Трудность передачи знаний от одной модели другой.
3. Плохая объясняемость принимаемых решений.
4. Наличие “галлюцинаций” нейросетей.

У любого варианта обучения с подкреплением имеются известные трудности:

1. Нестационарность распределения ответов среды (она может изменять вероятности переходов с течением времени).
2. Возможная экспоненциальная малость числа траекторий с большими наградами.

3. Нахождение баланса между исследованием качества выбранного действия и поиском более успешного альтернативного действия в текущем состоянии.

Наиболее ярко нестационарность проявляется при обучении стратегии игры с полной информацией, когда вероятности перехода в следующее состояние игры определяются как выбором действия учащегося, так и его оппонента. Чтобы справиться с первой трудностью в обучении с подкреплением обычно предполагают, что время изменения распределения ответов оппонента значительно превосходит время перемешивания цепей Маркова, соответствующих применяемым алгоритмам. Мы тоже будем делать это допущение.

Для учёта второй проблемы переходят на точку зрения вероятно-приближенно корректного (ВПК-) обучения – PAC-learning – Л. Вэльянта [Valiant, 1984].

Наконец, третью трудность мы исключим из обсуждения (хотя имеют-ся некоторые теоретические результаты), ограничившись случаем автономного обучения, когда этапы обучения и применения обученной стратегии разнесены во времени.

1. Элементы теории обучения с подкреплением

В обучении с подкреплением (ОсП) обучающийся, называемый *агентом*, взаимодействует с окружением. *Окружение* может находиться в одном из своих состояний $s \in S$, наблюдаемых игроком. В каждом из состояний агент может совершить одно из нескольких *действий* $a \in A$, зависящих от текущего состояния.

Например, для игры состояниями окружения являются только вершины дерева игры, соответствующие состояниям, в которых решение принимает обучающийся.

Так как поведение среды стохастично (например, для игры оппонент может использовать рэндомизованную стратегию), то переход к новому наблюдаемому состоянию имеет условное вероятностное распределение $P_a(s, s') = \mathbf{P}[S_{t+1} = s' | S_t = s, A_t = a]$ оказаться в состоянии s' при выборе (допустимого) действия a в состоянии s . Однако это распределение нам неизвестно. Более того, указанная выше трудность (1) говорит о том, что это распределение может быть нестационарным. В дальнейшем мы будем явно предполагать его стационарность.

В ОсП агент получает *непосредственное вознаграждение* $r_a(s, s') \in \mathbf{R}^1$ за попадание в состояние s' при выборе действия a в состоянии s . Конечно, вознаграждение может быть отрицательным: правильно назвать его в этом случае *наказанием*. В играх вознаграждение обычно получается только в финальных состояниях. Это условие упрощает как формулы, так и вычисления.

Целью агента в ОсП является выбор оптимальной политики $\pi: S \rightarrow \Delta^A$ – вероятностной стратегии выбора действий для каждого состояния, в котором он может находиться. Оптимальность вычисляется относительно (вообще говоря, дисконтированной со скоростью дисконта $\gamma > 0$) накопленной награды

$$V^\pi(s) = \mathbf{E}[\sum \gamma^t \cdot \Pi r_a(s', s'') \cdot \pi(a|s') \cdot \mathbf{P}[S_{j+1}=s'' | S_j=s', A_j=a] | S_0=s].$$

Дисконтирование вводилось для доказательства существования (даже детерминированной) оптимальной стратегии посредством неподвижной точки сжимающего оператора. Переходом $\gamma \rightarrow 1$ существование детерминированной оптимальной стратегии доказывается и в этом предельном случае. В дальнейшем, мы ограничимся случаем без дисконта ($\gamma = 1$). Суммирование идет по всем траекториям $s, \dots, s_n, s_{n+1}, \dots$ до остановки и всем последовательностям действий a_0, \dots, a_n, a_{n+1} , умножение – по j от 0 до $t-1$.

Для выбора оптимального действия в обучении с подкреплением рассматривают функцию Q-values

$$Q^\pi(s_0, a) = \mathbf{E}[\sum \Pi r_a(s', s'') \cdot \pi(a|s') \cdot \mathbf{P}[S_{j+1}=s'' | S_j=s', A_j=a] \cdot \mathbf{P}_a(s_0, s_1) | S_0=s_0, A_0=a].$$

Она соответствует выбору действия a в начальном состоянии s_0 , а затем применению политики π . Здесь умножение идет по j от 1 до $t-1$.

Неизвестность распределения вероятностей переходов заставляет заменять точное значение функции $Q(s_0, a)$ на ее приближение. Первоначально применялся метод Монте-Карло, после 2015 года – аппроксимация нейросетью.

2. Категория решёток и сходство

Мы лишь напомним ключевые понятия про категорию полных полу-решеток. Для более подробного знакомства читатель отсылается к статье [Виноградов, 2021].

Категория **Set** множеств и отображений между ними допускает эндифунктор $2^\wedge: \mathbf{Set} \rightarrow \mathbf{Set}$, который отображает множество X во множество-степень $2^\wedge X = \{A: A \subseteq X\}$, а отображение $f: X \rightarrow Y$ в отображение $2^\wedge f: 2^\wedge X \rightarrow 2^\wedge Y$, где $2^\wedge f(A) = \{f(x): x \in A\} \subseteq Y$, для любого $A \subseteq X$.

Имеется естественное преобразование $\cup: 2^\wedge \cdot 2^\wedge \rightarrow 2^\wedge$, которое отображает каждое семейство $S \subseteq 2^\wedge X$ подмножеств множества X в их объединение $\cup S = \cup \{A: A \in S\} \subseteq 2^\wedge X$.

Существует естественное преобразование $\{\cdot\}: I_{\mathbf{Set}} \rightarrow 2^\wedge$ тождественного функтора $I_{\mathbf{Set}}$ в функтор 2^\wedge , которое отображают каждый $x \in X$ в одноэлементное подмножество $\{x\} \subseteq 2^\wedge X$.

Тройка $\langle 2^\wedge, \{\cdot\}, \cup \rangle$ задаёт монаду в категории **Set**, для которой можно определить категорию полных (полу-)решёток **Lat** как множество пар $\langle L, \wedge \rangle$, где объект (множество) L называется *носителем решётки*, а морфизм $\wedge: 2^\wedge L \rightarrow L$ называется *сходством*, причём должны выполняться тождества $\wedge \cdot \cup = \wedge \cdot 2^\wedge \wedge$ и $\wedge \cdot \{\cdot\} = \text{id}$.

Лемма 1 из статьи [Виноградов, 2021] утверждает, что элементами этой категории являются полные полурешётки и только они.

Эта категория алгебр допускает свободные алгебры. Они соответствуют образу сопряженного функтора $G: \mathbf{Set} \rightarrow \mathbf{Lat}$ к забывающему функтору $F: \mathbf{Lat} \rightarrow \mathbf{Set}$, который сопоставляет полной полурешётке $\langle L, \wedge \rangle$ ее носитель L .

Теорема 1 из работы [Виноградов, 2021] доставляет явную конструкцию свободной алгебры над множеством X : ее носителем является множество-степень $2^\wedge X$, а сходством является объединение $\cup: 2^\wedge \cdot 2^\wedge X \rightarrow 2^\wedge X$.

Свободность образа $G: \mathbf{Set} \rightarrow \mathbf{Lat}$ (сопряженность с функтором $F: \mathbf{Lat} \rightarrow \mathbf{Set}$) означает наличие биекции между отображениями множеств $S \rightarrow L = F\langle L, \wedge \rangle$ и отображениями полурешеток $G(S) = \langle 2^\wedge S, \cup \rangle \rightarrow \langle L, \wedge \rangle$.

Этот факт имеет два важных следствия! Во-первых, при работе с решётками важно учитывать множества родителей сходства, а не только структурное описание. На этой идее базируется Анализ формальных понятий (АФП) [Ganter et al., 1999] – современный раздел теории решёток. Во-вторых, если имеется некоторое описание объектов (в нашем случае состояний среды обучения с подкреплением), на котором задана операция сходства, то она может быть распространена на всё множество-степень $2^\wedge S$.

3. Общая схема решёточного обучения с подкреплением

Из фундаментальной теоремы АФП следует, что любую полную решётку можно породить как сходства из списка битовых строк – формального контекста.

Чтобы применить ВКФ-метод обучения, мы будем предполагать, что состояния игры допускают кодирование бинарными признаками f_j ($j=1, \dots, n$) так, что сходство между состояниями представляется побитовым умножением. Так как побитовое умножение является базовой операцией современных процессоров, это приводит к значительному ускорению вычислений, которое осуществляется компилятором.

Группируя оценки действий в разных состояниях, но с одинаковым действием, получаем набор обучающих выборок, как представлено в табл. 1 ниже.

Таблица 1

	a_1	a_2	...	a_k	Q	f_1	...	f_n
s_1	1	0	...	0	$Q(s_1, a_1)$	$\delta_{1,1}$...	$\delta_{1,n}$
...
s_m	1	0	...	0	$Q(s_m, a_1)$	$\delta_{m,1}$...	$\delta_{m,n}$
s_1	0	1	...	0	$Q(s_1, a_2)$	$\delta_{1,1}$...	$\delta_{1,n}$
...
s_m	0	1	$Q(s_m, a_k)$	$\delta_{m,1}$...	$\delta_{m,n}$

Для фиксированного действия (например, a_1) обучающими примерами будут те строчки, где $Q(s_j, a_1) \geq 0$, а контр-примерами – дополнительные (где $Q(s_j, a_1) < 0$). На битовых строчках $\delta_{i,1} \dots \delta_{i,n}$, и $\delta_{j,1} \dots \delta_{j,n}$, соответствующих обучающим примерам, есть бинарная операция сходства – побитовое умножение.

При адекватном кодировании состояний битовыми строками $\delta_{i,1} \dots \delta_{i,n}$ можно надеяться, что общие признаки будут определять обобщение конкретных ситуаций (обучающих примеров), в которых применение выбранного действия приводит к большой награде. Контр-примеры необходимы для уменьшения числа сходств – устранения переобучения. Поэтому присоединение ВКФ-метода, вероятностно порождающего такие сходства, может превратить МКПД в хорошую альтернативу нейросетевым методам.

Заключение

В настоящей работе представлено формальное описание теоретико-решеточного подхода к обучению с подкреплением.

Возможность применения вероятностно-комбинаторного формального метода к задаче обобщения результатов оценивания методом Монте-Карло с поиском по дереву состояний позволяет надеяться на возрождение использования классического метода МКПД для ОСП.

Для апробации предложенного подхода автор запрограммировал на языке C++23 с использованием библиотеки oneTBB 2022.1 прототип системы ОСП для игр двух лиц с полной информацией Noughts&Crosses и нескольких вариантов игры Nim. На первом этапе существенную помощь ему оказала его аспирантка в ФИЦ ИУ РАН Л.А. Якимова. В дальнейшем планируется сравнение с известными нейросетевыми алгоритмами PPO [Schulman et al., 2017] от OpenAI и Deep Graph Network RL [Munikoti et al., 2022] при сотрудничестве с магистрантом МФТИ А.С. Мисником.

Благодарности. Идея написать настоящую работу возникла у автора в результате обсуждения со студентом МФТИ А.С. Мисником результатов совместной работы автора со своей бывшей аспиранткой ФИЦ ИУ РАН Л.А. Якимовой [Виноградов и др., 2024]. Им обоим автор выражает свою благодарность за сотрудничество и интерес к его работе. Автор считает своим приятным долгом поблагодарить своих коллег по ВЦ им. А.А. Дородницына ФИЦ ИУ РАН за поддержку и конструктивные дискуссии.

Список литературы

- [Виноградов, 2021] Виноградов Д.В. Проекция полурешеток: язык теории категорий // Научно-техническая информация. Серия 2. – 2021. – № 6. – С. 27-31.
- [Виноградов, 2022] Виноградов Д.В. Алгебраическое машинное обучение: упор на эффективность // Автоматика и телемеханика. – 2022. – № 6. – С. 5-23.

- [Виноградов и др., 2024] Виноградов Д.В., Якимова Л.А. Вероятностный подход к «доброму старомодному» искусственному интеллекту // Научно-техническая информация. Серия 2. – 2024. – № 3. – С. 21-26.
- [Саттон и др., 2020] Саттон Р.С., Барто Э.Дж. Обучение с подкреплением. – М.: ДМК Пресс, 2020. – 552 с.
- [Ganter et al., 1999] Ganter B., Wille R. Formal Concept Analysis. – Berlin: Springer, 1999.
- [Munikoti et al., 2022] Munikoti S., Agarwal D., Das L., Halappanavar M., Natarajan B. Challenges and opportunities in deep reinforcement learning with graph neural networks: A comprehensive review of algorithms and applications // arXiv preprint. – 2022. – doi: 10.48550/arXiv.2206.07922.
- [Schulman et al., 2017] Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal policy optimization algorithms // arXiv preprint. – 2017. – doi: 10.48550/arXiv.1707.06347.
- [Świechowski et al., 2023] Świechowski M., Godlewski K., Sawicki B., Mańdziuk J. Monte Carlo Tree Search: a review of recent modifications and applications // Artificial Intelligence Review. – 2023. – Vol. 56, – P. 2497-2562.
- [Valiant, 1984] Valiant L.G. An theory of the learnable // Communications of the ACM, – 1984. – Vol. 27, No. 11. – P. 1134-1142.