Научная статья
УДК 33
https://doi.org/10.24412/2073-0454-2024-5-244-248
EDN: https://elibrary.ru/tgvxiz
NIION: 2003-0059-5/24-173
MOSURED: 77/27-003-2024-05-372

Методы машинного обучения с подкреплением: теоретические основы и практическое применение

Каринэ Суреновна Хачатурян[1], Сурен Арутюнович Хачатурян[2]
[1] Финансовый университет при Правительстве Российской Федерации, Москва, Россия, kara111315hks@yandex.ru
[2] ФГБНУ «Аналитический центр», Москва, Россия, sure1311@gmail.com

Аннотация. Анализируются концепции и методы машинного обучения с подкреплением (RL), подробно раскрывая ключевые элементы этой технологии: агента, среду, стратегию, сигнал вознаграждения, функцию ценности и модель среды. Приводится обзор процесса формирования стратегий RL, который основан на использовании наград для принятия оптимальных решений. Исследуется важность марковских процессов и функции ценности для создания долгосрочных стратегий. Делается акцент на различии между модель-ориентированными и безмодельными подходами RL, выявляются их преимущества для адаптивного поведения и успешного выполнения задач.

Ключевые слова: обучение с подкреплением, агент, стратегия, функция ценности, марковский процесс, модель среды, оптимизация поведения

Благодарности: работа выполнена в рамках государственного задания Минобрнауки России по теме «Развитие методологии производства продукции двойного назначения высокотехнологичными компаниями России с использованием элементов искусственного интеллекта в условиях цифровизации экономики и санкционного давления» № 123011600034-3.

Для цитирования: Хачатурян К. С., Хачатурян С. А. Методы машинного обучения с подкреплением: теоретические основы и практическое применение // Вестник Московского университета МВД России. 2024. № 5. С. 244–248. https://doi.org/10.24412/2073-0454-2024-5-244-248. EDN: TGVXIZ.

Original article

Machine learning methods with reinforcement: theoretical foundations and practical applications

Karine S. Khachaturyan[1], Suren A. Khachaturyan[2]
[1] Financial University under the Government of the Russian Federation, Moscow, Russia, kara111315hks@yandex.ru
[2] Federal Center of Analyzis, Moscow, Russia, sure1311@gmail.com

Abstract. The concepts and methods of reinforcement learning (RL) machine learning are analyzed, detailing the key elements of this technology: agent, environment, strategy, reward signal, value function, and environment model. An overview is given of the RL strategy formation process, which is based on the use of rewards to make optimal decisions. The importance of Markov processes and the value function in creating long-run strategies is explored. Emphasis is placed on the distinction between model-based and model-free RL approaches, analyzing their benefits for adaptive behavior and successful task performance.

Keywords: reinforcement learning, agent, strategy, value function, Markov process, environment model, behavior optimization

Acknowledgements: the work was carried out within the framework of the state assignment of the Ministry of Education and Science of the Russian Federation on the topic «Development of methodology for the production of dual-use products by high-tech companies of Russia using elements of artificial intelligence in the conditions of digitalization of the economy and sanctions pressure» № 123011600034-3.

For citation: Khachaturyan K. S., Khachaturyan S. A. Machine learning methods with reinforcement: theoretical foundations and practical applications. Bulletin of the Moscow University of the Ministry of Internal Affairs of Russia. 2024;(5):244–248. (In Russ.). https://doi.org/10.24412/2073-0454-2024-5-244-248. EDN: TGVXIZ.

С конца 1960-х гг. мнение многих специалистов в области искусственного интеллекта заключалось в том, что интеллект формируется за счет различных специализированных техник, процедур и эвристик, не опираясь на универсальные принципы. Приемы, которые базируются на таких общих методиках, как, к примеру, поиск или обучение, считались «слабыми методами», а подходы, которые базировались на конкретных знаниях, получили название «сильных методов». Данная точка зрения остается распространенной и в настоящее время, но в данный момент она не занимает господствующих позиций. По нашему мнению, такое убеждение возникло преждевременно из-за недостаточных усилий, направленных на изучение общих принципов, а это не позволяет утверждать об их полном отсутствии.

В структуре обучения с подкреплением ключевыми элементами являются агент, его среда и четыре

структурных компонента: стратегический подход, индикаторы вознаграждений, механизмы оценки и, при необходимости, моделирование среды.

Стратегия устанавливает линию поведения агента, функционирует как путеводитель, который прокладывает курс действий от текущего положения в среде к оптимальному выбору реакций. Рассматриваемый нами процесс напоминает механизм формирования условных рефлексов в психологии, превращает раздражители в целенаправленные ответы. В зависимости от сложности задачи, стратегия может принимать форму простого соответствия или эволюционировать в комплексный вычислительный алгоритм, который нацелен на поиск оптимальных решений. Она лежит в основе адаптации агента, обучающегося через подкрепление, и может быть либо полностью детерминированной, либо стохастической, когда вероятности действий меняются в зависимости от обстоятельств [1].

Сигнал вознаграждения выступает как основополагающий элемент в процессе обучения с подкреплением; определяет цели для агента. На каждом этапе его деятельности окружающая среда предоставляет ему числовое значение, которое отражает эффективность его действий в данный момент. Данное значение служит оценкой того, насколько успешно или неуспешно агент справился с задачей. Основная задача агента заключается в том, чтобы максимизировать сумму получаемых вознаграждений в течение всего времени его функционирования. Сигнал вознаграждения, таким образом, является критерием для определения того, какие действия считаются положительными, а какие отрицательными в контексте поведения агента.

В биологических системах ощущения удовольствия или боли выполняют функции вознаграждения, поскольку они непосредственно воздействуют на поведение организма. Вознаграждение служит ключевым показателем для агента, указывает на значимые аспекты стоящей перед ним задачи и становится центральным элементом при адаптации его стратегии. Если конкретное действие не приводит к ожидаемому вознаграждению, необходима корректировка стратегии, чтобы в аналогичных будущих сценариях выбор пал на более эффективное действие. Сигнал вознаграждения можно описать как стохастическую функцию, зависящую от текущего состояния окружающей среды и выполненных действий [2].

В дискуссии о вознаграждении следует учитывать несколько основных моментов:

♦ вознаграждение не предоставляет прямых инструкций о методах решения задачи или о том, какие действия предпринять;

♦ вознаграждение часто откладывается во времени, а это существенно отличает его от непосредственной обратной связи в обучении с учителем;

♦ вознаграждение выступает в роли сигнала для обучения, указывает на успешность или неудачу предпринятых действий; в отличие от самостоятельного обучения, в нем такой обратной связи не предусмотрено.

Функция ценности служит важнейшим инструментом для оценки перспективности различных состояний в долгосрочной перспективе, выступает как прогнозирующий инструмент, который антисипирует общую сумму вознаграждений, на которые агент может рассчитывать, начиная действовать из данного момента. Вышеописанное отличается от немедленного вознаграждения, которое отражает текущую привлекательность ситуации, в то время как ценность предсказывает долгосрочную пользу; сюда же относятся потенциальные будущие сценарии и соответствующие вознаграждения, которые могут последовать. Например, исходное положение может обеспечивать стабильное, но скромное вознаграждение, однако также может открывать двери к более значимым вознаграждениям впоследствии [8].

Вознаграждение, таким образом, можно ассоциировать с мгновенными ощущениями удовольствия или разочарования, зависящими от его объема. Ценность, в свою очередь, можно сравнить с более глубоким, стратегическим размышлением о том, насколько удовлетворительным или недостаточным является текущее состояние. В этом контексте вознаграждение представляет собой первичный стимул, а ценность выступает в качестве вторичного аспекта, поскольку она представляет собой ожидание будущих вознаграждений и обеспечивает более сложный анализ возможных действий.

Без наличия вознаграждения понятие ценности теряло бы свою актуальность, поскольку основная задача оценки ценности заключается в максимизации получаемого вознаграждения. В контексте принятия решений и анализа их последствий именно ценность выходит на передний план. Решения формируются на основе оценок ценности, и предпочтение отдается тем действиям, которые ведут к наиболее ценным состояниям, а не обязательно к мгновенно большим вознаграждениям. В первую очередь это связано с тем, что такие действия обещают наибольшую выгоду в долгосрочной перспективе.

Вознаграждение непосредственно поступает от внешней среды, а ценность требует активного и непрерывного анализа на основе собранной информации о прошлых взаимодействиях. Ключ к большинству алгоритмов обучения с подкреплением лежит в умении точно оценивать ценность, опираясь на предыдущий опыт.

Модель окружающей среды, занимающая четвертое место в иерархии компонентов системы обучения с подкреплением и изображенная на рис. 1, выполняет функции симулятора. Она позволяет делать предположения о будущем поведении среды в ответ на действия агента, становится незаменимым инструментом для планирования и прогнозирования в динамичных условиях [3; 7].

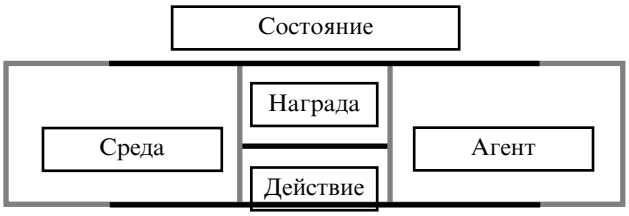

Рис. 1. Взаимодействие агента со средой [4]

Опираясь на текущие данные о состоянии и выбранных действиях, модель предсказывает будущее состояние и ожидаемые вознаграждения. Рассматриваемые нами прогностические модели служат фундаментом для стратегического планирования, которое включает разработку последовательности действий на основе тщательного анализа возможных будущих событий, еще до их наступления. Такие подходы в обучении с подкреплением, акцентирующие внимание на моделировании и стратегическом планировании, известны как модель-ориентированные. Они формируют явный контраст с более примитивными безмодельными методами, где учебный процесс происходит непосредственно через пробы и ошибки. Современные методики обучения с подкреплением представляют широкий спектр от простых методов испытаний до продвинутых систем стратегического планирования.

Обучение с подкреплением интенсивно использует понятие «состояние», которое занимает центральное место в его структуре. Состояние предоставляет критическую информацию, необходимую для разработки стратегий, вычисления функции ценности, и действует как основной вход и выход для моделирования окружающей среды. Данный параметр можно представить как информационный сигнал, который дает агенту представление о текущих условиях окружающей его среды. Обычно предполагается, что такой сигнал состояния генерируется встроенной системой предобработки данных, официально интегрированной в окружающую среду агента.

В рамках обучения с подкреплением можно разработать алгоритм, имитирующий процесс обучения через метод проб и ошибок. В отличие от традиционных подходов, в которых используются заранее подготовленные учебные наборы данных, этот алгоритм активно взаимодействует с окружающей средой. Величина вознаграждения, выдаваемая после каждого действия, служит показателем эффективности алгоритма в достижении целей, действует как мера его успеха в каждом конкретном случае взаимодействия.

В рамках обучения с подкреплением, задачи структурируются через марковский процесс принятия решений (MDP), включающий в себя компоненты S, A, P и r:

♦ **S** — пространство состояний, описывает все возможные кондиции, в которых может находиться система. В контексте управления запасами, это может быть, например, количество единиц товара на складе;

♦ **A** — пространство действий, включает в себя все доступные альтернативы действий, которые агент может предпринять в каждый данный момент. К примеру, это может быть, скажем, решение о пополнении запасов или их сохранении без изменений;

♦ **P** — функция переходов, определяет, как изменится среда после того, как агент примет действие a в состоянии s, обычно моделируется как вероятностное распределение, показывающее, с какой вероятностью осуществится переход в новое состояние s′;

♦ **r** — функция награды, присваивает числовое значение вознаграждению за выполнение действия; в состоянии s, действуя как ключевой мотивационный и обучающий сигнал для агента, стимулирует его к оптимизации действий для достижения максимальной выгоды.

Марковские процессы принятия решений представляют собой элегантное воплощение трех фундаментальных аспектов: восприятия, действия и целеполагания, представленных в удивительно простой, но не банальной форме. Методы, которые эффективно справляются с этими задачами, принято классифицировать как техники обучения с подкреплением.

В этих системах агент следует определенной стратегии, которая регулирует его действия на основе текущего состояния окружающей среды, известной как «policy» или «политика», и представляется как функция распределения $\pi(a|s)$. Именно эта стратегия является центральным объектом разработки.

Динамика управляемой марковской цепи, которая используется для моделирования системы, поведение которой можно описать последовательностью состояний с учетом выбранных действий. В таком подходе на каждом шаге времени, обозначенном *t*, мы, принимая некоторое управленческое решение a_t, определяем вероятностное распределение, согласно которому система переходит в следующее состояние $s_t + _1$.

Ключевым предположением, лежащим в основе данного подхода, является соблюдение **марковского свойства** [3; 6]. Оно утверждает, что вероятность перехода системы из текущего состояния s_t в следующее состояние $s_t + _1$ зависит исключительно от текущего состояния s_t и текущего действия a_t, без учета предыдущих состояний и действий, которые привели к s_t. Таким образом, вся необходимая информация для принятия решения на следующем шаге содержится в текущем состоянии системы.

Это предположение формально выражается следующим образом:

$$p(s_t + _1|s_t,a_t,s_t - _1,a_t - _1,\ldots,s0,a0) = p(s_t + _1|s_t,a_t). \quad (1)$$

Данная формула выражает **марковское условие** (или **условие марковской независимости**), утверждающее, что переходы в системе подчиняются принципу независимости от прошлого при условии текущего состояния и действия. Это свойство упрощает задачи прогнозирования и управления в системе, так как для оптимального управления требуется учитывать лишь текущее состояние и текущее действие, что значительно сокращает объем необходимой информации и вычислительных ресурсов.

В такой системе окружающая среда функционирует как управляемая марковская цепь, в которой каждый выбор действия формирует вероятностное распределение, предсказывающее возможные будущие состояния. Это означает, что последовательность переходов между состояниями определяется исключительно текущим контекстом и не зависит от предшествующих действий или состояний. Также предполагается, что функция переходов не меняется со временем; это придает системе стационарный характер.

При обучении с подкреплением центральным элементом является задача оптимизации определенных функционалов. В отличие от традиционных подходов, в машинном обучении, в которых выбор функции потерь может быть скорректирован инженерным путем, здесь функция вознаграждения предоставляется как данность и служит мерилом для оценки желаемых результатов действий агента, указывает направление для максимизации достижений.

Современные подходы к решению таких задач часто используют динамическое программирование, но для разработки более продвинутых решений важно глубоко понимать структуру задачи, предложенную в рамках MDP. Вышеописанное включает определение текущего состояния s и выбора оптимального действия *a*, что приводит к награде r = *r(s,a)*, и последующее переходное состояние s', подготавливая сценарий для аналогичной задачи выбора в новой ситуации.

Большинство рассматриваемых методов обучения с подкреплением построено на оценивании функций ценности, но это необязательно. Функция ценности состояния *V(s)* определяет ожидаемую суммарную дисконтированную стоимость за бесконечный период времени, начиная с состояния *s* и далее следуя политике π; т. е., *V(s)* показывает «ценность» нахождения в состоянии *s* с точки зрения минимизации затрат.

Эта функция может быть выражена рекурсивно через функцию ценности следующих состояний *s'*, в которые может перейти система. Основное уравнение для функции ценности выглядит так:

$$V(s) = \frac{min}{a \in A} [\![c(s,a) + \gamma \frac{\sum}{s \in S} P(s'|s,a) V(s')]\!], \quad (2)$$

где:

- c(s,a) — ожидаемые затраты на действие *a* из состояния *s*;
- γ — коэффициент дисконтирования, который уменьшает вес будущих затрат (обычно $0 \leq \gamma < 1$);
- $\sum_{s' \in S} P(s'|s,a) V(s')$ — ожидаемая ценность следующего состояния *s'*, в которое система может перейти из состояния *s* при действии *a*.

Это уравнение называется уравнением Беллмана и используется для нахождения минимальной стоимости при переходах между состояниями. Если заменить сумму интегралом, то можно получить непрерывный аналог данного уравнения для задач с непрерывными пространствами состояний и действий.

Для нахождения оптимального действия в каждом состоянии используется функция ценности действия *Q(s,a)*, которая показывает, какие затраты будут при выборе действия *a* в состоянии *s*:

$$Q(s,a) = c(s,a) + \gamma \frac{\sum}{s \in S} P(s'|s,a) V(s'). \quad (3)$$

Эта функция помогает агенту оценивать выгоду каждого возможного действия. Функция *V(s)* показывает минимальные затраты в состоянии s, тогда как *Q(s,a)* описывает затраты конкретного действия *a* в этом состоянии. Оптимальное действие выбирается так, чтобы минимизировать значение *Q(s,a)*.

Таким образом выявлена важность обучения с подкреплением как подхода к формированию адаптивного и оптимального поведения агента. Основываясь на

марковских процессах, RL позволяет моделировать ситуации, требующие долгосрочных стратегий, где агент постепенно учится на основании вознаграждений и накопленных данных о среде. Выделение функций стратегии и ценности, наряду с оптимизацией вознаграждений, обеспечивает агенту возможность принимать обоснованные решения, фокусируясь на выгоде в долгосрочной перспективе. Перспективы RL включают интеграцию более сложных алгоритмов, которые могут улучшить точность моделирования, делая систему способной решать задачи высокой сложности.

Список источников

1. Жерон О. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow. М., 2017.

2. A Neuroevolution Approach to General Atari Game Playing // URL: https://www.cs.utexas.edu/mhauskn/papers/atari.pdf

3. Выборнова О. Н. Применение машинного обучения с подкреплением для автоматизированного поиска уязвимостей информационных систем // MMTT. 2020. Т. 4. С. 110–113.

4. Саттон Р. С., Барто Э. Дж. Обучение с подкреплением. Введение. 2-е изд. М., 2020.

5. Bellman R. E. A Markov decision process // Journal of Mathematical Mechanics. 1957. P. 679–684.

6. Хачатурян К. С., Хачатурян С. А. Технологии искусственного интеллекта для эффективного управления запасами // Экономика и управление: проблемы, решения. 2024. Т. 3. № 6(147). С. 283–287.

7. Хачатурян К. С. Влияние процессов цифровой трансформации на глобальную экономику // Вестник Московского университета им. С.Ю. Витте. 2024. № 2(49). С. 36–41.

8. Хачатурян А. А., Абдулкадыров А. С. Современные проблемы совершенствования организации и управления высокотехнологическими кластерами // Вестник Московского университета МВД России. 2024. № 1. С. 230–236.

References

1. Geron O. Simple mechanical learning using Scikit-Learn and TensorFlow. M., 2017.

2. A Neuroevolution Approach to General Atari Game Playing // URL: https://www.cs.utexas.edu/mhauskn/papers/atari.pdf

3. Vybornova O. N. The use of machine learning with reinforcement for automated vulnerability search of information systems // MMTT. 2020. Vol. 4. P. 110–113.

4. Sutton R. S., Barto E. J. Reinforcement learning: An introduction. 2nd ed. M., 2020.

5. Bellman R. E., Markov decision-making process // Journal of Mathematical Mechanics. 1957. P. 679–684.

6. Khachaturyan K. S., Khachaturyan S. A. Artificial Intelligence Technologies for Effective Inventory Management // Economics and Management: Problems, Solutions. 2024. Vol. 3. No. 6(147). P. 283–287.

7. Khachaturyan K. S. The Impact of Digital Transformation Processes on the Global Economy // Bulletin of Witte University of Moscow. 2024. No. 2(49). P. 36–41.

8. Khachaturyan A. A., Abdulkadyrov A. S. Modern Problems of Improving the Organization and Management of High-Tech Clusters // Bulletin of Moscow University of the Ministry of Internal Affairs of Russia. 2024. No. 1. P. 230–236.

Информация об авторах

К. С. Хачатурян — профессор кафедры экономической теории Финансового университета при Правительстве Российской Федерации, доктор экономических наук, профессор;

С. А. Хачатурян — старший научный сотрудник ФГБНУ «Аналитический центр», кандидат экономических наук.

Information about the authors

K. S. Khachaturyan — Professor of the Department of Economic Theory of the Financial University under the Government of the Russian Federation, Doctor of Economic Sciences, Professor;

S. A. Khachaturyan — Senior Researcher of the Federal Center of Analyzis, Candidate of Economic sciences.

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Статья поступила в редакцию 12.09.2024; одобрена после рецензирования 30.09.2024; принята к публикации 07.10.2024.

The article was submitted 12.09.2024; approved after reviewing 30.09.2024; accepted for publication 07.10.2024.