УДК: 004.896

АЛГОРИТМЫ РАННЕГО ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

Шарибаев А Н, магистрант

Московский физико-технический институт

Шарибаев Р Н, магистрант

Наманганский инженерно-технологический институт

Абдулазизов Б. Т, доцент

Тохиржонова М. Р, студент

Наманганский государственный университет

Аннотация: *В работе обсуждается возникновение обучения с подкреплением, включая его ранние применения в психологии и исследованиях поведения животных. Исследуется разработка алгоритмов обучения с подкреплением, от простых методов проб и ошибок до более сложных методов глубокого обучения с подкреплением.*

Ключевқе слова: *алгоритмы, обучения с подкреплением, глубокого обучения с подкреплением*

REINFORCEMENT EARLY LEARNING ALGORITHMS

Sharibaev A N, undergraduate

Moscow Institute of Physics and Technology

Sharibaev R N, undergraduate

Namangan Institute of Engineering and Technology

Abdulazizov B. T, Associate Professor

Tokhirjonova M. R, student

Namangan State University

Abstract: This paper discusses the emergence of reinforcement learning, including its early applications in psychology and animal behavior research. The development of reinforcement learning algorithms is explored, from simple trial and error methods to more complex deep reinforcement learning methods.

Keywords: algorithms, reinforcement learning, deep reinforcement learning

Одним из главных преимуществ Q-learning и TD learning является их простота и эффективность. Эти алгоритмы могут быть применены к широкому кругу задач, не требуя сложных моделей окружающей среды, и они могут сходиться к оптимальным политикам с относительно небольшим количеством итераций обучения.

Однако оба алгоритма также имеют ограничения. Например, Q-learning может страдать от проблем конвергенции в средах с многомерными пространствами состояний или большими пространствами действий. Обучение TD может быть чувствительным к выбору скорости обучения и может плохо работать в условиях с длительными временными горизонтами или скудным вознаграждением.

За десятилетия, прошедшие с момента их разработки, Q-learning и TD-learning были расширены и модифицированы различными способами, чтобы улучшить их производительность и устранить их ограничения. Например, исследователи разработали варианты Q-learning, которые используют аппроксимацию функций, такую как нейронные сети, для представления Q-value в пространствах состояний высокой размерности. Эти алгоритмы, известные как deep Q-networks (DQNs), использовались для достижения производительности на уровне человека в различных видеоиграх и были распространены на другие области, такие как робототехника и обработка естественного языка.

Другим расширением TD-learning является алгоритм SARSA, который расшифровывается как Состояние-Действие-Вознаграждение-Состояние-Действие. SARSA - это не зависящий от модели алгоритм, который определяет оптимальную политику путем оценки ожидаемого совокупного вознаграждения за совершение действия в определенном штате и следования определенной политике с этого момента. В отличие от Q-learning, SARSA обновляет Q-value, используя фактическое действие, предпринятое агентом в следующем состоянии, а не действие, выбранное жадной политикой. Это делает SARSA более устойчивой к стохастичности и исследованию.

Одной из основных проблем алгоритмов обучения с подкреплением является балансировка между исследованием и эксплуатацией. Эксплуатация относится к процессу принятия мер, которые хороши на основе прошлого опыта. В то время как исследование предполагает принятие мер, которые могут быть неоптимальными в краткосрочной перспективе, но потенциально могут принести более высокие долгосрочные выгоды. Ключевой вопрос в обучении с подкреплением заключается в том, как сбалансировать эти две стратегии, чтобы выучить оптимальную политику, избегая при этом застревания в локальных оптимумах.

Чтобы решить эту проблему, исследователи разработали множество стратегий исследования, таких как методы “epsilon-gready”, “softmax” и “upper confidence bound” (UCB). Эти стратегии позволяют агентам более эффективно анализировать окружающую среду, а также обеспечивать принципиальный баланс между исследованием и эксплуатацией.

Другой проблемой в обучении с подкреплением является работа с частичной наблюдаемостью, когда агент имеет доступ только к ограниченному подмножеству информации о состоянии среды. Это часто имеет место в реальных сценариях, где агент может не иметь доступа к полному представлению окружающей среды или может иметь дело с зашумленными или неполными наблюдениями. Чтобы решить эту проблему, исследователи разработали частично наблюдаемые модели марковского процесса принятия решений (POMDP), которые позволяют агентам рассуждать о неопределенности и принимать решения на основе частичной информации.

Наконец, алгоритмы обучения с подкреплением могут быть чувствительны к выбору гиперпараметров, таких, как скорость обучения и коэффициент дисконтирования, что может повлиять на скорость сходимости и стабильность процесса обучения. Чтобы решить эту проблему, исследователи разработали такие методы, как поиск по сетке, случайный

поиск и байесовская оптимизация для автоматической настройки гиперпараметров.

Таким образом, обучение с подкреплением значительно эволюционировало с момента своего появления в психологии и исследованиях поведения животных и стало мощным инструментом для изучения оптимальных стратегий в широком спектре областей. Несмотря на свои успехи, обучение с подкреплением остается активной областью исследований, и многие проблемы еще предстоит решить. Однако, благодаря продолжающейся разработке новых алгоритмов, моделей и методов, обучение с подкреплением может оказать серьезное влияние на многие области науки и техники в ближайшие годы.

Литература

[1]. Schaul, T., Quan, J., Antonoglou, I., Silver, D. (2016). Prioritized experience replay. arXiv preprint arXiv:1511.05952.

[2]. Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In International conference on machine learning (pp. 1928-1937).