УДК 004.8

ОБЗОР НАДЕЖНЫХ МЕТОДОВ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ ДЛЯ НЕПРЕРЫВНОГО КОНТРОЛЯ

Барбахан И. (Университет ИТМО)
Научный руководитель – кандидат физико-математических наук, доцент Фильченков А.А.
(Университет ИТМО)

Эта работа представляет собой исследование, посвященное рассмотрению того, как современные методы исследования подходят к проблеме непрерывного контроля в обучении с подкреплением, и мы в основном сосредоточились на обучении с подкреплением в автономном режиме и концепции надежности в поисках темы в этом направлении.

Введение.

Задачи обучения с подкреплением могут иметь либо дискретное, либо непрерывное пространство действий, что сильно влияет на используемый алгоритм.

Алгоритмы глубокого обучения с подкреплением уже применялись как к дискретным, так и к непрерывным пространствам действий с разной эффективностью. Были проведены исследования для сравнения производительности хорошо зарекомендовавших себя алгоритмов глубокого обучения с подкреплением без моделей, таких как DQN и другие.

Результаты таких экспериментов продемонстрировали необходимость большего поиска в области непрерывного контроля, поскольку стандартные подходы обучения с подкреплением должны были достигать лучших результатов в дискретном пространстве действий, чем в пространстве непрерывного действия.

Основная часть.

Автономное обучение с подкреплением (RL) — это вновь возникающая область исследований, целью которой является изучение поведения с использованием только зарегистрированных данных, таких как данные предыдущих экспериментов или демонстраций человека, без дальнейшего взаимодействия с окружающей средой. У него есть потенциал для достижения огромного прогресса в нескольких реальных проблемах принятия решений, где активный сбор данных является дорогостоящим, таким как робототехника, открытие лекарств, создание диалогов, системы рекомендаций и другие, или небезопасным/опасным, таким как здравоохранение, автономное вождение, или образование.

Такая парадигма обещает решить ключевую проблему переноса алгоритмов обучения с подкреплением из ограниченных лабораторных условий в реальный мир, и это включает в себя работу с непрерывными управляющими действиями как общей характеристикой реальных проблем.

Автономное обучение с подкреплением обещает преодолеть разрыв между алгоритмами обучения с подкреплением и реальными приложениями. Благодаря использованию больших предварительно собранных наборов данных это может смягчить технические проблемы, связанные со сбором данных в режиме онлайн, поскольку взаимодействие с реальными средами в реальных приложениях часто дорого или опасно. Эти обещания вызвали всплеск интереса к исследованиям этой проблемной области, требующей заметных улучшений.

Чтение литературы об автономном обучении с подкреплением привело к связи между достижением хорошего решения задачи непрерывного контроля и обеспечением того, чтобы решение было надежным и способным к обобщению.

Алгоритм является надежным, если он хорошо работает даже при наличии небольших ошибок во входных данных. Это определение поднимает некоторые вопросы, например, что значит работать хорошо? А что считать мелкой ошибкой? И как вычислить надежные решения?

Надежность не важна для некоторых задач управления, таких как перевернутый маятник, а также для компьютерных игр, таких как Atari или Minecraft, или настольных игр, таких как

шахматы и го, и это в основном потому, что эти приложения имеют детерминированную динамику, и у них есть быстрые и точные симуляторы. где есть много доступных данных и легко протестировать политику, и самое главное, что в таких приложениях невыучить хорошую политику дешево.

Надежность считается важной в реальных приложениях, где мы учимся на зарегистрированных данных (пакетное RL), где нет симулятора, никогда не бывает достаточно данных, и тот факт, что нет перекрестной проверки, что затрудняет тестирование политики, и в случае плохой политики будет высокая цена неудачи.

В реальных приложениях, таких как сельское хозяйство, когда нам нужно планировать пестициды, одна итерация составляет один год, а плохая политика означает отсутствие урожая. Другим примером является оптимизация обслуживания инфраструктуры или здравоохранения для улучшения управления инсулином при диабете, решения таких проблем, как автономные транспортные средства и робототехника, и многих других.

Таким образом, цель устойчивости состоит в том, чтобы получить наилучшую политику в отношении входных данных со всеми возможными небольшими ошибками, и это называется состязательной устойчивостью для обучения с подкреплением, и ошибка может быть либо в вознаграждениях, либо в переходах.

Когда ошибка находится в функции вознаграждения, цель состоит в том, чтобы максимизировать политику, сводя к минимуму небольшую ошибку в вознаграждении, подходы к решению этой проблемы, как правило, превращают эту проблему в регуляризацию с двойственностью линейного программирования. Однако в целом не существует аналогичных алгоритмов итерации с известным значением или итерации политики.

Когда ошибка находится в функции перехода, цель состоит в том, чтобы максимизировать политику, сводя к минимуму небольшую ошибку перехода. В общем, это NP-трудно решить, нет известных линейных формулировок, однако его можно решить с помощью итераций значений и итераций политики.

Наиболее многообещающим подходом к достижению устойчивости к ошибке функции перехода является использование множества неоднозначности, также известного как множество неопределенности.

Решить это можно по-разному, используя два подхода множества неоднозначности: S-прямоугольный и SA-прямоугольный, основное различие между ними в том, что в S-прямоугольном характере шум выбирает, не зная, какое действие мы выбираем, потому что мы определяем ограничения для каждого состояния, но тогда это происходит через несколько действий из состояния, в то время как в SA-прямоугольнике мы принимаем более сильную природу, видим состояние, в котором мы находимся, и можем реагировать на любое действие, которое мы предпринимаем, другим видом шума.

Другими словами, в S-прямоугольном мы видим только состояние, а не действие, он должен зафиксировать шум, прежде чем узнать, какие действия мы предпринимаем, SA-прямоугольный видит состояние и действие и на основе этого выбирает другое действие. вид шума.

в обоих случаях это полиномиальное время разрешимо с оптимальностью Беллмана.

Рассматривая это как надежный марковский процесс принятия решений, подобный игре, которую мы играем против природы, где у нас есть состояние и мы выбираем действие, природа выбирает мою вероятность перехода, чтобы определить, в какое состояние мы переходим и какую награду мы получаем, и принимая принимая во внимание, что природа может быть либо статичной, когда природа устанавливает ошибку в состояние, она должна представлять одну и ту же ошибку каждый раз, когда мы посещаем состояние, либо динамичной, когда природа зависит от истории и может меняться при каждом посещении.

Выводы.

Это исследование поднимает пару вопросов о выборе уровня надежности:

1- Каков правильный размер уровня шума набора неоднозначности?

2- Должен ли этот уровень шума быть одинаковым для каждого состояния и действия?

3- Зачем использовать норму L1? Как насчет L∞, KL-дивергенции и прочего?

4- Какую прямоугольность использовать (если есть)?

Ответы на все эти вопросы зависят от того, почему возникают ошибки. Чтобы понять это, также исследуйте, как обучение с подкреплением на основе модели с использованием подхода, подобного Dyna, путем сбора данных о переходе, затем использование машинного обучения для построения модели перехода, решение модели MDP для получаем политику, а затем развертываем политику, однако модель все еще может быть ошибочной по многим причинам, таким как упрощение модели, ограниченные данные, нестационарная среда, зашумленные наблюдения и другие, и каждый из этих источников ошибок требует разной обработки.

Таким образом, модель становится надежным обучением с подкреплением на основе модели путем сбора данных перехода, использования ML для уверенного построения модели перехода, решения надежной модели MDP для получения политики, а затем мы уверенно развертываем политику. Однако эти подходы к каждому типу ошибок могут иметь некоторые плюсы и минусы в обобщении, возможности использовать априорные значения и требуют больших вычислительных ресурсов.

Одним из наиболее многообещающих подходов является основанный на моделях подход к надежному табличному RL с эффективной выборкой вне политики путем изучения моделей и достоверности, и он открывает множество направлений исследований в отношении масштабируемости, ослабления прямоугольности и понимания реального воздействия и ограничений методов.

Барбахан И. (автор) Подпись

Фильченков А.А. (научный руководитель) Подпись