https://doi.org/10.26160/2474-5901-2023-35-113-117

КОНЦЕПЦИЯ АУГМЕНТАЦИИ АЛГОРИТМОВ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ ПРИ ПОМОЩИ КЛЕТОЧНЫХ МОДЕЛЕЙ ТРАНСПОРТНОЙ СЕТИ В ЗАДАЧЕ СВЕТОФОРНОГО РЕГУЛИРОВАНИЯ

Матросов С.В.
Московский государственный университет им. М.В. Ломоносова, Москва, Россия

Ключевые слова: транспортная сеть, светофорное регулирование, обучение с подкреплением, клеточная транспортная модель.
Аннотация. В данной работе рассматривается вопрос применения клеточных моделей транспортной сети совместно с методами обучения с подкреплением в задаче светофорного регулирования. Предлагается схема управления, позволяющая расширять пространство состояний при помощи результатов моделирования. Также предлагается архитектура слоя нейросети, отвечающего за извлечение вектора признаков из прогноза клеточной модели.

THE CONCEPT OF AUGMENTATION OF REINFORCEMENT LEARNING ALGORITHMS USING CELLULAR MODELS OF A TRAFFIC NETWORK IN A TRAFFIC SIGNAL CONTROL PROBLEM

Matrosov S.V.
Lomonosov Moscow State University, Moscow, Russia

Keywords: traffic network, traffic signal control, reinforcement learning, cellular traffic model.
Abstract. In this paper we consider the application of cellular models of transport network together with reinforcement learning methods in a traffic signal control problem. We propose a control scheme that allows to expand the state space using simulation output. Also we propose the architecture of neural network layer, which is responsible for extracting the feature vector from the prediction of the cellular model.

Алгоритмы адаптивного светофорного регулирования представляют большой практический интерес, т.к. позволяют повысить пропускную способность транспортной сети [1].

В последние годы набирают популярность методы светофорного регулирования, основанные на обучении с подкреплением [2]. В тестовых сценариях они показывают себя лучше, чем классические алгоритмы, однако их практическое применение сопряжено с рядом сложностей.

Большинство современных алгоритмов светофорного регулирования, опирающихся на методы обучения с подкреплением, используют для описания состояния системы и расчета функции награды только показания детекторов. Это может приводить к значительному падению качества управления на слабо оснащенных детекторами участках транспортной сети. В литературе можно найти множество моделей трафика, способных достаточно точно предсказывать поведение транспортных потоков [3]. Аугментация состояния системы при помощи прогнозов таких моделей может улучшить качество работы алгоритмов управления.

Обычно, для обучения агента применяется компьютерная модель управляемой системы. В большинстве случаев, построение модели транспортной сети оказывается очень трудоемкой задачей. При этом полученная модель используется только при обучении агента, но не задействована явно в процессе управления.

В настоящей работе предлагается схема совместного применения нейросетей и клеточных моделей транспортной сети в задаче светофорного регулирования. Основой метода является аугментация показаний детекторов при помощи предсказаний модели. В работе предлагается архитектура слоя нейросети, отвечающего за извлечение признаков из прогнозов клеточной модели. Предложенный метод совместим со всеми стандартными model-free алгоритмами обучения с подкреплением.

Общий подход обучения с подкреплением. В рамках подхода обучения с подкреплением агент учится максимизировать свою выгоду при взаимодействии со средой. После каждого взаимодействия он получает вознаграждение. В ходе обучения агент методом проб и ошибок должен найти оптимальную стратегию поведения.

Среда обычно описывается как марковский процесс принятия решений (S,A,P,r,γ). Множества S и A определяют возможные состояния системы и доступные агенту действия. Модель среды $P(s_{t+1}|s_t,a_t)$ задает вероятность перехода в состояние s_{t+1} из состояния s_t при выполнении действия a_t. Функция $r(s_t,a_t,s_{t+1})$ определяет награду, получаемую агентом при переходе в состояние s_{t+1} из состояния s_t при выполнении действия a_t, а γ задает коэффициент дисконтирования.

Стратегия агента должна задавать действие a_t, которое он предпримет, находясь в состоянии s_t. В наиболее общем случае стратегия задается как условное распределение $\pi(a_t|s_t)$. Задачей алгоритма обучения является поиск стратегии π^*, приносящей максимальную среднюю суммарную награду:

$$\pi^* = \arg\max_{\pi}\left(E_P\left[\sum_{i=0}^{\infty} \gamma^i r_{t+i} \right] \right).$$

Для того чтобы сформулировать задачу светофорного регулирования как задачу обучения с подкреплением необходимо описать используемые множества состояний, действий и функцию награды.

Модель среды в задаче светофорного регулирования. Введем ряд понятий, необходимых для постановки задачи. *Светофорным контроллером* называется специальное устройство, осуществляющее управление светофорами на перекрестке. Контроллер переключается между *фазами*, задающими разрешенные направления движения. *Детекторы транспорта* измеряют параметры транспортных потоков в транспортной сети.

Управление светофорными контроллерами осуществляется эпизодически – для каждого эпизода t длительностью ΔT выбирается управление $a_t \in A$, которое определяет какие фазы будут включены на контроллере в этого время. Можно выделить три основных стратегии формирования множества действий A [2].

1. *Прямое переключение фаз* – в момент времени t для каждого контроллера выбирается фаза, которая будет включена в следующем эпизоде.

2. *Продление фаз* – похожа на предыдущий подход, но последовательность фаз для каждого контроллера фиксируется, и агент может выбирать между продлением текущей фазы или включением следующей

3. *Построение программ* – для каждого контроллера задается программа, фиксирующая последовательность и длительность фаз на следующий эпизод.

Функция награды должна оценивать, как управление влияет на пропускную способность транспортной сети. Обычно ее выбирают в зависимости от того, какие параметры могут измерять детекторы транспорта. Часто используются следующие функции награды: длина очереди, давление на перекрестке, средняя скорость транспортного потока, средняя интенсивность транспортного потока и их комбинации.

Отметим, что все перечисленные функции оценивают эффективность управления только в окрестности детекторов. Из-за этого агент может неправильно оценивать влияние своих действий в процессе обучения, что приводит к падению итогового качества управления.

Состояние системы на основе прогнозов клеточной модели транспортной сети. Можно выделить несколько распространенных подходов для описания множества состояний S [2].

1. *Табличное представление* – состояние системы описывается при помощи таблицы признаков. Как правило, признаки либо совпадают с показаниями детекторов, либо вычисляются на их основе. Примером таких признаков можно считать длину очереди или время ожидания перед светофором, среднюю скорость на полосе, суммарную длину очереди на перекрестке и т.п.

2. *Представление в виде снимков* – к табличным данным добавляются снимки/видео с камер видеонаблюдения или схематичные представления состояния перекрестков, такие как DTSE (discrete traffic state encoding)

Можно предположить, что более богатое пространство признаков позволяет агенту более эффективно принимать решения. При этом указанные выше методы используют для описания состояния системы только данные детекторов, которые, как правило, покрывают транспортную сеть неравномерно. Из-за этого возникает ситуация, когда агент вынужден принимать решения, опираясь на неполную информацию о состоянии системы.

Для решения данной проблемы предлагается использовать клеточные модели [3], при помощи которых можно достраивать состояние транспортной сети. Клеточные модели разделяют транспортную сеть на сегменты (клетки). Каждой клетке соответствует набор параметров, характеризующий транспортный поток в области. Примером таких характеристик может быть интенсивность, скорость или плотность потока. Динамика системы задается при помощи набора правил, по которым меняются параметры клеток.

Транспортную сеть будем представлять как направленный граф, в котором ребрам соответствуют дороги, а вершинам перекрестки. Перекрестки могут быть управляемыми, т.е. оснащенными светофорными контроллерами, и неуправляемыми. Состояние дороги l в момент времени t задается матрицей $U_{l,t} = [u_{l,t}^1, \ldots, u_{l,t}^N]$, где $u_{l,t}^i$ это вектор параметров i-й клетки. Состояние среды $s_t \in S$ задается как состояние клеточной модели транспортной сети в совокупности с показаниями детекторов.

Отметим, что теперь функцию награды можно строить не только на основании показаний детекторов, но и с использованием прогноза клеточной модели. Это позволяет оценивать состояние всей транспортной сети, а не отдельных ее участков.

Следует отметить, что обширный класс макроскопических моделей [4] также можно связать с клеточными моделями. Макроскопические модели описывают состояние трафика как распределение плотности $\rho(x,t)$, интенсивности $Q(x,t)$ и средней скорости потока $v(x,t)$ в каждой точке x транспортной сети в настоящий момент времени t, а динамика этих параметров задается системой дифференциальных уравнений в частных производных. Для расчета этих моделей часто используются разностные схемы, которые в совокупности с алгоритмом обновления параметров на сетке также можно рассматривать как клеточные модели.

Слой извлечения признаков. Чтобы эффективно использовать прогнозы клеточной модели транспортной сети**,** предлагается добавить в нейросеть специализированный слой, отвечающий за формирование вектора признаков. Этот слой будет состоять из трех блоков: свертки для ребер графа, детекторный блок, и блок, отвечающий за соединение дорог в перекрестки.

Свертки по ребрам графа будут извлекать признаки транспортного потока, движущегося по дороге. Архитектура блока – одномерная сверточная нейросеть. Число каналов в первом сверточном слое совпадает с размерностью вектора $u_{l,t}^i$. Предполагается, что один и тот же блок будет использоваться для каждой дороги в графе транспортной сети. Это позволит уменьшить количество обучаемых весов, что должно повысить скорость обучения.

Детекторный блок отвечает за признаки, связывающие показания детекторов и предсказания клеточной модели в области детекции. Архитектура блока – полносвязная нейросеть. На вход блока подается конкатенация показаний детектора и параметров клетки $u_{l,t}^i$. Для детекторов одного типа используется один и тот же блок.

Блок, отвечающий за признаки на перекрестках, формирует признаки, объединяя результаты применения сверток по всем ребрам, смежным с данным перекрестком. Архитектура – полносвязная нейросеть. На вход подается конкатенация результатов свертки смежных ребер графа. Если перекресток управляемый, то на вход также подается текущая включенная на контроллере фаза. Для каждого перекрестка создается отдельный экземпляр блока со своим набором весов.

Состояния транспортной сети $s_t \in S$ преобразуется в вектор признаков по следующему алгоритму – сверточным блоком обрабатываются все ребра графа, одновременно детекторный блок обрабатывает показания детекторов, после вычисления сверток считаются перекрестки, результаты этих операций объединяются в один вектор, который передается в последующие слои нейросети.

Обучения и эксплуатация нейросетевого агента. Процесс обучения можно разделить на два этапа. Сначала строится граф транспортной сети, и калибруются параметры клеточной модели. Потом для этой модели строится нейросеть и запускается стандартный процесс обучения.

Каждая итерация цикла управления происходит в два этапа – сначала оценивается состояние транспортной сети при помощи клеточной модели, а потом нейросеть использует результаты моделирования для выбора оптимальных фаз для светофорных контроллеров.

Заключение. В настоящей работе была предложена техника аугментации показаний детекторов при помощи прогнозов клеточной модели транспортной сети, для решения задачи светофорного регулирования методами обучения с подкреплением. Этот подход позволяет ослабить требования к плотности покрытия управляемой зоны детекторами. Большой интерес представляет апробация предложенного метода в компьютерном эксперименте и сравнение его с другими нейросетевыми алгоритмами светофорного регулирования на общем наборе тестовых сценариев [5].

Список литературы / References

1. Khattak Z.H., Magalotti M.J., Fontaine M.D. Operational performance evaluation of adaptive traffic control systems: A Bayesian modeling approach using real-world GPS and private sector PROBE data // Journal of Intelligent Transportation Systems. 2020, vol. 24, no. 2, pp. 156-170. doi.org/10.1080/15472450.2019.1614445
2. Haydari A., Yılmaz Y. Deep Reinforcement Learning for Intelligent Transportation Systems: A Survey // IEEE Transactions on Intelligent Transportation Systems. 2022, vol. 23, no. 1, pp. 11-32. doi.org/10.1109/TITS.2020.3008612.
3. Femke van Wageningen-Kessels, Hans van Lint, Kees Vuik, Serge Hoogendoorn. Genealogy of traffic flow models // Journal on Transportation and Logistics. 2015, vol. 4, no. 4, pp. 445-473. doi.org/10.1007/s13676-014-0045-5.
4. Mohan R., Ramadurai G. State-of-the art of macroscopic traffic flow modelling // International Journal of Advances in Engineering Sciences and Applied Mathematics. 2013, vol. 5, pp. 158-176. http://doi.org/10.1007/s12572-013-0087-1.
5. Ault J., Guni S. Reinforcement Learning Benchmarks for Traffic Signal Control // Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track. – 2021.

Матросов Сергей Владимирович – аспирант	**Matrosov Sergey Vladimirovich** – graduate student
matrosik14@gmail.com	

Received 05.05.2023