

ЗАДАЧА MARL ДЛЯ СВЕТОФОРА НА ПЕРЕКРЁСТКЕ

Т.И. Тисленко

9 мая 2021 г.

В работе рассматривается задача оптимизации сети светофоров с точки зрения уменьшения задержек трафика автомобилей. Объектом исследования являются временные интервалы, в течение которых машины не перемещаются. Предметом является согласование действий сети светофоров по управлению процессом выбора задержки сигнала: зеленого, красного. В качестве метода решения поставленной задачи предложен мульти-агентный подход. Каждый интеллектуальный агент решает задачу целесообразности выбора того или иного сигнала. Предложены математические модели системы, учитывающие возможность совместного выполнения задачи.

1 Основные определения

1.1 Интеллектуальным агентом называется метаобъект, наделенный долей субъектности, взаимодействующий с другими агентами и средой, выполняющий определенные функции для достижения поставленных целей.

1.2 Средой называется множество объектов, не принадлежащих агенту.

1.3 Задачами/Ресурсами называется объект, распределяемый агентами в ходе достижения их целей.

1.4 Мультиагентная система – совокупность взаимосвязанных агентов.

1.5 RL(Reinforcement Learning) — Обучение с подкреплением, где в роли учителя выступает среда. Как правило, RL используется для одного агента в среде, чтобы максимизировать его долгосрочную (накопительную, кумулятивную) награду. Модель среды – Марковский процесс принятия решения - MDP(Markov decision process), предполагается, что эта среда стационарна, т.е. состояние среды зависит только от действий агента. Самая общая модель обучения одного агента в среде — Q-обучение. Q-обучаемый агент учится оптимальному сопоставлению действия $a(action)$ состоянию окружающей среды $s(state)$ на основе кумулятивных наград $r(s, a)$ (reward).

2 Постановка задачи MARL

Рассмотрим модель обучения одного агента.

В качестве агента выступает светофор. Ресурсами такой агент не располагает.

Среда — перекресток с машинами, где на отрезках дорог за 100м до стоп-линий засекается время.

Состояние среды отражает активность фазы светофора. Обозначим их фаза0, фаза1. Пространство состояний $S = \{ s_0 = \text{«активна фаза0»}, s_1 = \text{«активна фаза1»} \}$

В момент времени t_k активна фаза светофора S_k , суммарное засеченое всех машин, проходящих через отрезок дороги называется задержкой на фазе S_k

Множество решений $A = \{ a_0 = \text{«оставить фазу»}, a_1 = \text{«сменить фазу»} \}$

Будем считать, что функция вознаграждения всецело определяется текущим состоянием, выбранной стратегией и состоянием, в которое перейдет процесс на следующем шаге:

$r(s_k, a_0)$ = задержка на фазе s_k

$r(s_k, a_1)$ = задержка на фазе $s_{1-k}, k = 1, 2$.

$p(i, k; j)$ вероятность того, что система из состояния i при выборе решения k попадает в состояние j , полностью определяется состоянием, в которое переходит процесс.

$V^*(s)$ — функция суммарных внешних доходов от оптимальной политики в состоянии s

$$V^*(s) = \max_{a(\cdot)} \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t). \quad (1)$$

Исходя из описания задачи уравнение Вальда - Беллмана [1] для управляемого марковского процесса с конечным числом действий и состояний имеет вид:

$$V^*(s) = \max_{a \in A} \left\{ \sum_{s' \in S} p(s, a; s') (r(s, a; s') + \gamma V^*(s')) \right\} = \max_{a \in \{a_0, a_1\}} \left\{ \sum_{s' \in \{s_0, s_1\}} p(s, a; s') (r(s, a; s') + \gamma V^*(s')) \right\}. \quad (2)$$

вероятности в правойой части неизвестны

Пусть Q -функция имеет следующий вид:

$$Q(s, a) = \sum_{s' \in S} p(s, a; s') (r(s, a; s') + \gamma V^*(s')). \quad (3)$$

тогда:

$$V^*(s) = \max_{a \in A} Q(s, a). \quad (4)$$

$$Q(s, a) = \sum_{s' \in S} p(s, a; s') (r(s, a; s') + \gamma \max_{a' \in A} Q(s', a')). \quad (5)$$

Итак, $Q = \{Q(s, a)\}_{s \in S, a' \in A}$, можно записать итеративно $Q_{t+1} = A(Q_t)$, где $A: \mathbb{R}_{\infty}^1 \rightarrow \mathbb{R}_{\infty}^1$ — сжимающее отображение.

$$\begin{aligned} \rho((A \circ Q_1)(s, a), (A \circ Q_2)(s, a)) &= \max_{a' \in A, s' \in S} \left| \sum_{s' \in S} p(s, a; s') (r(s, a; s') + \gamma \max_{a' \in A} Q_1(s', a')) - \right. \\ &\left. \sum_{s' \in S} p(s, a; s') (r(s, a; s') + \gamma \max_{a' \in A} Q_2(s', a')) \right| \leq \max_{a' \in A, s' \in S} |\gamma \max_{a' \in A} Q_1(s', a') - \gamma \max_{a' \in A} Q_2(s', a')| = \\ &= \gamma \rho(Q_1(s, a), Q_2(s, a)), \gamma \in (0; 1) \end{aligned} \quad (6)$$

Последовательность $\{Q_t\}$ представляет собой приближенные решения $AQ = Q$ эффективный способ оценки точности которого:

$$\rho(Q_{t_n}(s, a), Q_{t_0}(s, a)) = \rho((A^n \circ Q_{t_0})(s, a), Q_{t_0}(s, a)) \leq \frac{\gamma^n \rho(Q_{t_1}(s, a), Q_{t_0}(s, a))}{1 - \gamma} = \frac{\gamma^n \rho((A \circ Q_{t_0})(s, a), Q_{t_0}(s, a))}{1 - \gamma} \quad (7)$$

Идея Q -обучения заключается в оценке невычислимой правой части:

$$Q_{t+1}(s, a) = Q_t(s, a) + \alpha_t(s, a) \left(r(s, a) + \gamma \max_{a' \in A} Q_t(s', a') - Q_t(s, a) \right) \quad (8)$$

где s' — положение процесса на шаге $t+1$, если на шаге t процесс был в состоянии s и было выбрано действие a . Если на шаге t процесс находился в состоянии s и было выбрано действие a , то $0 \leq \alpha_t(s, a) \leq 1$, иначе $\alpha_t(s, a) = 0$.

$$V_t^*(s) = \max_{a \in A} Q_t(s, a). \quad (9)$$

Оказывается, что если используемая стратегия $a(s)$ приводит к тому, что с вероятностью 1 каждая пара (s, a) будет бесконечное число раз встречаться на бесконечном горизонте наблюдения, то из отмеченного выше условия сжимаемости при

$$\sum_{t=0}^{\infty} \alpha_t(s, a) = \infty, \sum_{t=0}^{\infty} \alpha_t(s, a)^2 \leq \infty \quad (10)$$

удет следовать сходимость (с вероятностью 1) процесса 8

$$V^*(s) = \max_{a \in A} \lim_{t \rightarrow +\infty} Q_t(s, a). \quad (11)$$

$$a_t(s) = \arg \max_{a' \in A} Q_t(s, a') \quad (12)$$

3 Цель работы

Целью работы являлось ознакомление с подходами, позволяющими оптимизировать процесс выбора сигнала светофора, с учетом текущей загрузки транспорта, с точки зрения минимизации задержки. В работе получены следующие результаты:

1. Математическая модель процесса выбора фазы светофора, отличающаяся учетом текущего расположения светофоров и их загрузки и позволяющая сформулировать оптимизационные задачи, целью которых является минимизация задержки трафика автомобилей.
2. Структура мультиагентной системы, включающая в себя единственного агента – светофор, обеспечивающая наиболее эффективное распараллеливание всей задачи на подзадачи, которые будут решены агентом.

Список литературы

- [1] Лекции по случайным процессам : учебное пособие / А. В. Гасников, Э. А. Горбунов, С. А. Гуз и др. ; под ред. А. В. Гасникова. – «Москва» : МФТИ, 2019. – 285 с. ISBN 978-5-7417-0710-4
- [2] Марковские процессы принятия решений. / Майн Х., Осаки С. Главная редакция физико-математической литературы издательства «Наука», 1977. - 176 с. УДК 519.283