ЗАДАЧА MARL ДЛЯ СВЕТОФОРА НА ПЕРЕКРЁСТКЕ

Тисленко Тимофей Иванович

ФГАОУ ВО «СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт математики и фундаментальной информатики

Научный руководитель — к.ф.-м.н., доцент Д.В. Семенова

Томск, МПОИТЭС 2021

Актуальность

Цели и задачи

Цель работы

Разработка и исследование математической модели мультиагентной системы для задачи оптимизации движения на перекреске. . .

Задачи

1. Сделать обзор литературы по соответствующей тематике.
2. Описать математическую модель.
3. Описать алгоритм.
4. Продемонстрировать результаты работы алгоритма.

Основные определения

Определения:

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

Средой называется множество объектов, не принадлежащих агенту.

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

Мультиагентная система – совокупность взаимосвязанных агентов.

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

Постановка задачи

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 = $\{ s_0 =$ «фаза0» , $s_1 =$ «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 = «оставить фазу» , a_1 = «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

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

Уравнение Вальда - Беллмана [**?**] для управляемого марковского процесса с конечным числом действий и состояний имеет вид:

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

Т.е.

$V^*(\text{«фаза0»}) = (p(\text{«фаза0»}, \text{«оставить фазу»}; \text{«фаза1»})(r(\text{«фаза0»}, \text{«оставить фа}$
$\gamma V^*(\text{«фаза1»}))+p(\text{«фаза0»}, \text{«оставить фазу»}; \text{«фаз01»})(r(\text{«фаза0»}, \text{«оставить фа}$
$\gamma V^*(\text{«фаза0»})) > (p(\text{«фаза0»}, \text{«сменить фазу»}; \text{«фаза1»})(r(\text{«фаза0»}, \text{«сменить ф}$
$\gamma V^*(\text{«фаза1»}))+p(\text{«фаза0»}, \text{«сменить фазу»}; \text{«фаз01»})(r(\text{«фаза0»}, \text{«сменить фаз}$
$\gamma V^*(\text{«фаза0»}))))?$
$p(\text{«фаза0»}, \text{«оставить фазу»}; \text{«фаза1»})(r(\text{«фаза0»}, \text{«оставить фазу»}; \text{«фаза1»})+$

Пусть Q-функция имеет следующий вид:

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

Идея Q-обучения заключается в оценке невычислимой правой части:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t(s,a)\left(r(s,a) + \gamma \max_{a' \in A} Q_t(s',a') - Q_t(s,a)\right) \quad (4)$$

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 $\alpha_t(s,a) = 0$.
тогда:

$$V^*(s) = \max_{a \in A} Q(s,a) = \max_{a \in A} Q_t(s,a). \quad (5)$$

Сходимость

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

$$\rho((A\circ Q_1)(s,a),(A\circ Q_2)(s,a)) = \max_{a'\in A, s'\in S} |\sum_{s'\in S} p(s,a;s')(r(s,a;s') + \gamma \max_{a'\in A}$$

$$\sum_{s'\in S} p(s,a;s')(r(s,a;s') + \gamma \max_{a'\in A} Q_2(s',a'))| \le \max_{a'\in A, s'\in S} |\gamma \max_{a'\in A} Q_1(s',a')) - \gamma \max_{a'\in A} Q_2(s',$$

$$= \gamma\rho(Q_1(s,a), Q_2(s,a)), \gamma \in (0;1) \tag{6}$$

Оказывается, что если используемая стратегия $a(\cdot)$ 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

∞ ∞

Основные результаты работы

Целью работы являлось ознакомление с подходами, позволяющими оптимизиров 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 распараллелив всей задачи на подзадачи, которые будут решены агентом.

- 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
- Марковские процессы принятия решений. / Майн Х., Осаки С. Главная редакция физико-математической литературы издательства «Наука», 1977. - 176 с. УДК 519.283

СПАСИБО ЗА ВНИМАНИЕ!!!