УДК 65.015.013

ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ В ЗАДАЧЕ ДИНАМИЧЕСКОГО ПЛАНИРОВАНИЯ РАБОТ

Воронин В.В. (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО)
Научный руководитель – кандидат технических наук, доцент, Гулева В.Ю.
(федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО)

В работе рассматриваются алгоритмы обучения с подкреплением для решения задачи динамического планирования работ

Введение. Задача планирования работ, так называемая job shop scheduling problem, является NP-трудной оптимизационной задачей. Классическая постановка (JSSP) почти всегда далека от реальных производственных процессов и большинство усилий направлено на решение динамической версии данной задачи (DJSSP), которая формулируется следующим образом: имеется N работ $J = \{j_1, j_2, \dots j_N\}$, которые должны быть распределены для выполнения на M машинах $M = \{m_1, m_2 \dots m_M\}$ таким образом, чтобы минимизировать общую продолжительность выполнения всех работ. Длительность t_i выполнения работы j_i задается распределением $p(t_i|j_i)$. Каждая машина может выполнять одномоментно только одну работу. Следующая работа не может быть начата, пока не закончена предыдущая.

Мы расширяем постановку, вводя следующие требования: выполнение работы j_i с вероятностью $p(k|j_i)$ приводит к добавлению k новых работ. Работы связаны между собой, т.е. после выполнения работы j_i можно выбрать не любую из оставшихся, а лишь работы из подмножества $\{j_a, \dots j_n\}, n \leq N$. Связь работ определяется матрицей A.

Основная часть. Обучение с подкреплением определяется как взаимодействие агента и среды. Агент определяется через стратегию выбора действий π, а окружающая среда – через распределение реакций среды $p(s', r|s, a)$, где s' и s – следующие и текущее состояния среды, a действие агента, r – моментальная награда. В нашем случае, так как задача эпизодическая, то моментальная награда равна 0, а агент получает выигрыш лишь в случае успешного завершения последовательности всех N работ. Последовательность называется успешной, если соблюдены связи между работами согласно матрице A. Действие a доступное для выполнения из состояния s мы определим как работу j_i, доступную для выполнения после завершения предыдущей. Наконец, состояние s мы определим как множество еще не завершенных работ.

Обзор существующих методов Direct RL, которые мы будем использовать из-за отсутствия готового распределения реакций среды $p(t_i|j_i)$ показывает, что в отличии от Q-learning/SARSA, policy-based алгоритмы используются редко. Предлагается, разбить граф, задаваемый матрицей A на подграфы, с помощью policy gradient algorithm обучить каждый из них и обобщить полученную таким образом стратегию на весь граф, задаваемый матрицей A.

Выводы. Модель может быть использована при планировании производственных процессов в различных отраслях народного хозяйства, однако, изначально работа была инициирована запросом одной из действующих авиакомпаний России. Она заинтересована в уменьшении затрат на простой самолета, возникшего из-за ошибок планирования процесса технического обслуживания. По грубым оценкам, один день простоя на земле самолета типа Airbus 320F обходится ей в $25000. Имеется договоренность об использовании исторических данных авиакомпании для испытаний разрабатываемых моделей. При хороших результатах на исторических данных будет поставлен вопрос о внедрении.

Воронин В.В. (автор) _____

Гулева В.Ю. (научный руководитель) _____