УДК 519.83

ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ В ЭВОЛЮЦИОННЫХ ИГРАХ

А.Ю. Павлов
Московский физико-технический институт
Россия, 141701, Долгопрудный, Институтский пер., 9
E-mail: admpavlov@yandex.ru

Ключевые слова: обучение с подкреплением, эволюционные игры, игры на графах, кооперирование, дилемма заключенных.

Аннотация: В докладе[1] изучаются характеристики динамики системы, состоящей из N игроков, находящихся в узлах графа, результат взаимодействия между которыми в рамках одного периода описывается классической дилеммой заключенных. Поскольку рассматривается динамическая версия игры, агенты максимизируют не текущий одноперидный выигрыш, а суммарную приведенную ожидаемую полезность за весь период игры и, таким образом, действуют дальновидно. Изучается возможность использования алгоритмов обучения с подкреплением для реализации такой дальновидной стратегии игроков. Интерес представляет изучение свойств алгоритмов, позволяющих получить кооперирование двух агентов. Одним из таких свойств является необходимость разделения динамики на две стадии: обучение и непосредственно игра. Кроме того, изучается зависимость свойств динамики от степени дальновидности агентов.

1. Введение

Интерес к изучению характеристик эволюции игр, обучение игроков в которых описывается алгоритмами обучения с подкреплением, возникает в контексте разработки мультиагентных адаптивных систем [1,2], поскольку теоретико-игровое описание – это, вероятно, единственное возможное описание взаимодействия агентов в таких системах. Ввиду этого, классической отправной точкой для изучения свойств систем агентов, обучающихся посредством алгоритмов обучения с подкреплением, является изучение различных модификаций дилеммы заключенных, как одного из наиболее хорошо изученных примеров игр.

В докладе изучаются свойства динамики системы, состоящей из N игроков, находящихся в узлах графа, результат парного взаимодействия между которыми в рамках одного периода описывается классической дилеммой заключенных. Два агента (заключенные) делают выбор из двух вариантов: обвинить оппонента (D) или молчать (C). Матрица выплат в данной игре представлена в Таблице 1. Как известно, в одноперидной игре стратегия D является доминирующей.

Таблица 1. Матрица выплат в одноперидной дилемме заключенных.

	C	**D**
C	3;3	-1;5
D	5;-1	1;1

[1] Доклад основан на материалах работы Leonidov A., Pavlov A., Serebryannikova E. Reinforcement learning in evolutionary games on graphs. *Work in progress*.

Пусть агенты играют T раундов, причем T агентам не известно. Цель агентов в данном случае – максимизация суммарной приведенной полезности следующего вида

$$W(S_i^1, S_{-i}^1) = \sum_{\tau=1}^{T} \gamma^{\tau-1} r(s_i^\tau, s_{-i}^\tau), \quad (1)$$

где $r(s_i^\tau, s_{-i}^\tau)$ – выплата игрока i в момент времени τ, если он выбрал стратегию s_i^τ, s_{-i}^τ – стратегии других игроков в момент времени τ, γ – норма дисконтирования, $S_i^t = (s_i^t, s_i^{t+1}, \ldots, s_i^T)$, $S_{-i}^t = (s_{-i}^t, s_{-i}^{t+1}, \ldots, s_{-i}^T)$.

В такой постановке стратегия D не обязательно является доминирующей: играя постоянно D, можно ожидать, что оппонент ответит тем же, и средний доход за раунд будет близок к единице, тогда как при кооперативном поведении обоих игроков выигрыш мог бы быть больше.

Вообще говоря, для выбора стратегии, максимизирующей полезность (1), игрок должен воспользоваться принципом Беллмана, то есть найти функцию Беллмана $V(x,t)$, определяемую следующим соотношением

$$V(\Phi, t) = \max_{S_i^t} \sum_{\tau=t}^{T} \gamma^{\tau-1} r(s_i^\tau, s_{-i}^\tau), \quad (2)$$

так что

$$V(\Phi_1, 1) = \max_{S_i^1} W(S_i^1, S_{-i}^1), \quad (3)$$

где Φ – это вектор, характеризующий состояние системы.

Обучение с подкреплением – это один из подходов к получению некоторой аппроксимации функции Беллмана. В данных алгоритмах итеративно ищется функция ценности $Q(x,a)$ каждого действия (a) в зависимости от состояния игры (x). Состояниями при этом считаются n последних результатов игры, где n параметр алгоритма. Действие затем выбирается так, чтобы максимизировать значение функции *Q(x,a)* при заданном x.

Важным параметром обучения является параметр γ, характеризующий ценность будущих состояний при оценке текущего действия. Нулевое значение этого параметра отвечает случаю игры близоруких агентов, при котором стратегия D является доминантной. Чем ближе значение γ к единице, тем большее влияние оказывают будущие выплаты на действие, совершаемое в текущий момент времени.

Ключевые показатели, характеризующий эволюцию данной системы - частота результата игры (D, D) (DR, defect rate) и частота результата игры (C, C) (CR, cooperate rate).

2. Результаты

Исследование эволюции системы производилось в игре двух агентов, обучение которых производилось при помощи алгоритма Q-learning. В дальнейшем будет проведено исследование случая игры N агентов на графе.

Выше отмечалось, что в каждом периоде действие выбирается так, чтобы достигался максимум функции $Q(x,a)$ при заданном состоянии x. Однако в задачах обучения возникает так называемый exploration-exploitation tradeoff (см., например, [2]), заключающийся в том, что эффективный процесс обучения возможен только если агент с ненулевой вероятностью исследует всевозможные комбинации состояние-действие, но, с другой стороны, при этом он с необходимостью будет совершать неоптимальные шаги. Поэтому при выборе оптимального действия используется процедура $\mathrm{softmax}_a Q(x,a)$

при фиксированном состоянии x. Однако это приводит к доминации стратегии D вне зависимости от величины γ (рис. 1).

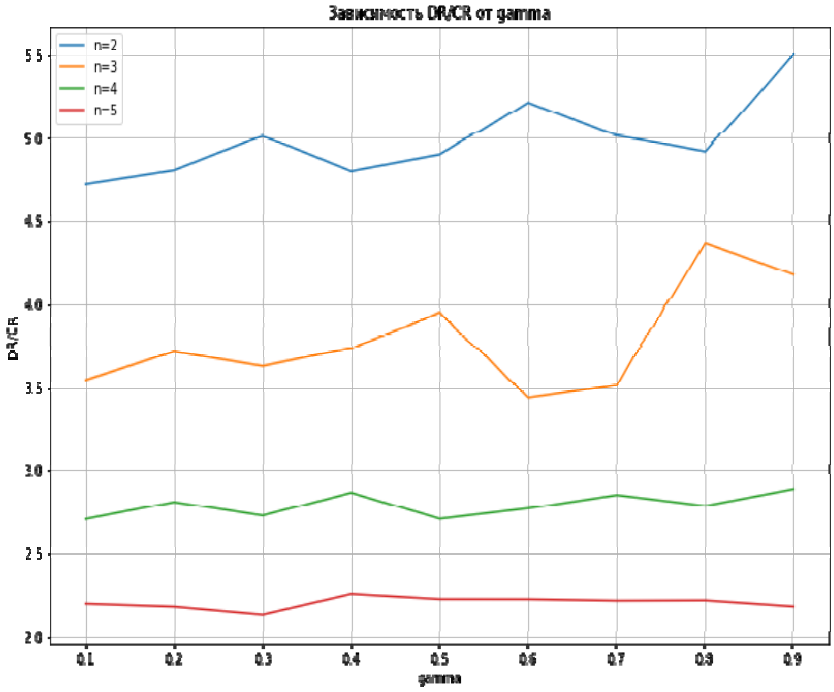

Рис. 1. Зависимость DR/CR от gamma и n. Даже при больших gamma DR/CR > 1.

Избежать этого можно путем разделения алгоритма на две стадии: обучение – выбор действия происходит по функции softmax; непосредственная игра – выбор действия происходит по функции argmax, т.е. выбирается наилучшее действие. Из рис. 2 видно, что при этом DR, как и ожидалось, уменьшается с ростом предусмотрительности агентов (описываемой параметром γ). Таким образом, при достаточной степени предусмотрительности агенты начинают кооперироваться чаще.

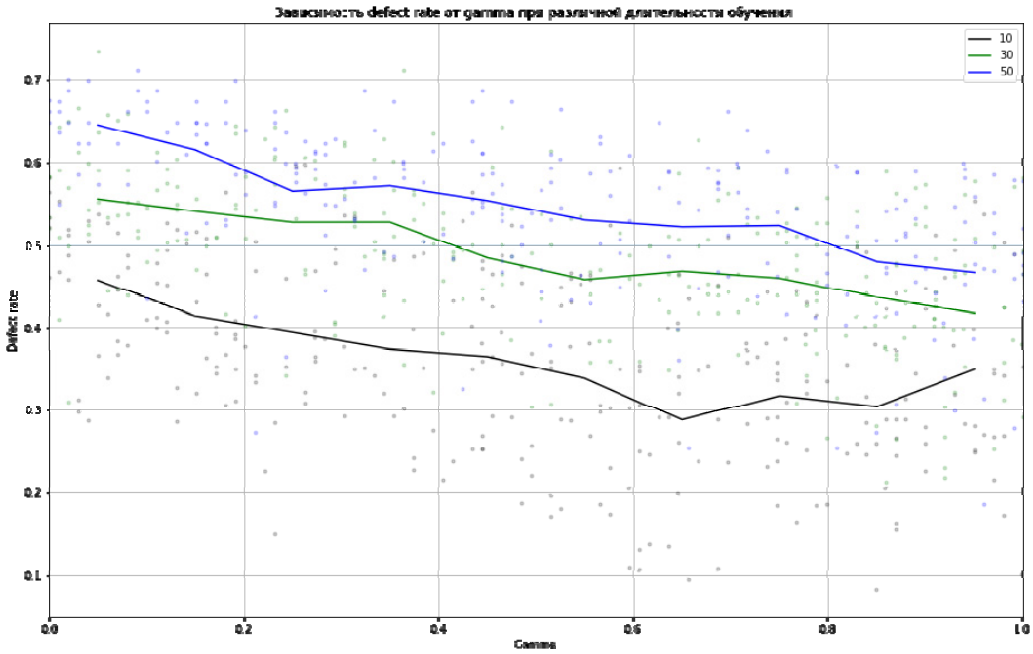

Рис 2. Зависимость DR от gamma и длительности обучения. При больших gamma низкий DR.

3. Заключение

Сформулирована задача исследования эволюции игры N агентов на графе, выбор стратегий которых определяется на основе результатов применения алгоритмов обучения с подкреплением. В дальнейшем предполагается проведение численных экспериментов для исследования влияния топологии сети взаимодействия агентов на эволюцию системы, а также более глубокое изучение влияния степени предусмотрительности агентов на результаты игры.

Список литературы

1. Tuyls K., Nowé A. Evolutionary game theory and multi-agent reinforcement learning // The Knowledge Engineering Review. 2005. Т. 20. №. 1. С. 63-90.
2. Buşoniu L., Babŭska R., De Schutter B. Multi-agent reinforcement learning: An overview // Innovations in Multi-Agent Systems and Applications – 1 (D. Srinivasan and L.C. Jain, eds.). Mol. 310 of Studies in Computational Intelligence, Berlin, Germany: Springer, 2010. Chapter 7. P. 183-221.
3. Watkins C.J.C.H., Dayan P. Q-learning // Machine learning. 1992. Vol. 8, No. 3-4. P. 279-292.