УДК 004.021 DOI: 10.18698/2541-8009-2021-04-694

МЕТОД ГЛУБОКОГО МЕТАМУЛЬТИАГЕНТНОГО МАШИННОГО ОБУЧЕНИЯ, ОСНОВАННЫЙ НА ПРИНЦИПЕ МАКСИМУМА

К.Д. Смирнова
iron.karina@gmail.com
SPIN-код: 7120-0782

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Аннотация

Рассмотрена концепция машинного обучения и сформулирована целевая задача как поиск максимально универсальных методов обучения. Дано описание используемой в эксперименте дискретной среды и действующих в ней агентов. Описан метод Q-обучения, на котором основан разработанный алгоритм. Приведены результаты эксперимента с концептуальным алгоритмом машинного обучения, основанным на принципе максимума, где в рамках метаобучения алгоритм использовал данные предобученных структур, и исследованы его результаты. Проанализирована работа алгоритма с различными наборами параметров. Исследованы закономерности влияния ключевых параметров обучения на результат и рассмотрены перспективы их использования.

Ключевые слова

Машинное обучение, обучение с подкреплением, метаобучение, высокоуровневая система, Q-обучение, мультиагентное обучение, коэффициент дисконтирования, скорость обучения

Поступила в редакцию 08.04.2021

Введение. Машинное обучение представляет собой класс методов реализации искусственного интеллекта, которые базируются на идее о том, что аналитические системы могут учиться выявлять закономерности и принимать решения с минимальным влиянием человеческого фактора. Машинное обучение в настоящее время насчитывает огромное количество методов, часто имеющих определенную специфику для решения конкретных типов задач, и стремление к универсальному и сильному искусственному интеллекту наиболее преследуют методы машинного метаобучения.

Осмысливая главную концепцию машинного обучения не как непосредственное прямое решение задачи, а как способность к обучению, позволяющую решать широкий спектр задач, и ориентируясь на парадигмы искусственного интеллекта, можно сформулировать целевую задачу как поиск максимально универсальных и всеобъемлющих методов обучения [1]. Решая эту задачу, можно сказать, что предполагаемая модель должна быть обучена на огромном множестве задач, характеризуемых метапараметрами, т. е. метаобучена [2]. Однако, отталкиваясь от уже очевидной в машинном обучении возможности некоторой модели обучения покрыть некоторое множество задач, можно сделать вывод о том, что следует обучать глобальную модель на продуктах обучения различных менее универсальных моделей [3]. Так, помимо определения метаобучения как

«обучения тому, как нужно учиться» или «обучения с выделением характерных метапараметров» будет разумно рассматривать метаобучение в более практическом аспекте: метаобучение как обучение высокоуровневых структур.

Разработанный алгоритм предназначен для стартового исследования поведения модели, при построении которой используются результаты предобученных структур. В нем применяется мультиагентный подход, позволяющий избежать дополнительных ограничений по числу агентов [4].

Описание среды. Обучение и тестирование происходит в среде Starcraft II [5, 6]. Существуют две симметричные карты, на которых друг напротив друга располагаются две пары агентов (действующих сущностей): с одной стороны — пара агентов, управляемая разработанным алгоритмом, в рамках которого они исследуют среду и стреляют в нападающих (красные), с другой — пара агентов-ботов, управляемая искусственным интеллектом игры и реализующая только нападение на наших красных агентов (синие). На одной карте нападающие агенты находятся слева, защищающиеся — справа, на второй карте — наоборот. Для агентов карта представляет собой табулярный мир, описываемый в прямоугольной системе координат, что позволяет, применяя метод Q-обучения, использовать *Q*-таблицу в качестве продукта предобучения. Используемые карты представлены на рис. 1.

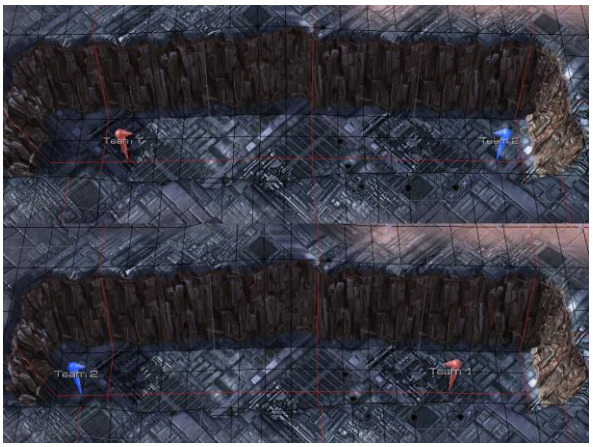

Рис. 1. Карты, используемые в эксперименте

Описание алгоритма обучения. Алгоритм обучения выстроен на основе метода *Q*-обучения и, соответственно, предусматривает заполнение *Q*-таблицы значениями награды. *Q*-таблица представляет собой матрицу, содержащую информацию о наградах агента для каждой пары состояние — действие [7].

Процесс обучения может регулироваться тремя параметрами: скоростью обучения α, коэффициентом дисконтирования γ и динамическим коэффициентом ε (подобным коэффициенту ε в эпсилон-жадной стратегии — регулирующий случайность выбора действия на первых эпизодах обучения и обеспечивающий неслучайный выбор действия на поздних эпизодах обучения) [8]. Также одним из параметров является количество эпизодов обучения. Чтобы проследить динамику поведения в зависимости от этих параметров, их задают вруч-

ную. В рамках алгоритма вычисляются координаты агента, т. е. его состояние, определяются доступные действия, и устанавливается награда для пары состояние — действие (т. е. обучается) по основному принципу *Q*-обучения, описываемому вариацией уравнения Беллмана [9]:

$$Q(s, a) = Q(s, a) + \alpha[r + \max_{a'} Q(s', a') - Q(s, a)],$$

где $Q(s, a)$ — награда для пары состояние — действие; $\max_{a'} Q(s', a')$ — максимальная ожидаемая награда впоследствии.

Полученные значения записывают в предварительно созданную пустую *Q*-таблицу.

Описание эксперимента. В основе идеи лежит образование *Q*-таблицы сложением по методу максимума *Q*-таблиц, полученных в результате обучения на симметричных картах, и исследование поведения агентов с ее использованием. Сначала агенты обучаются на двух реверсивных картах, в процессе чего для каждого из них заполняется собственная *Q*-таблица. Это и есть тот результат предобучения, который будет использоваться далее. Затем эти *Q*-таблицы тестируются и фиксируются результаты обучения: главными показателями качества обучения является средняя награда и средний винрейт (процент успешной защиты красных агентов, т. е. процент побед). *Q*-таблицы, полученные в результате нападения справа и нападения слева, суммируются по методу максимума — по своей структуре таблицы абсолютно идентичны, и в ячейку итоговой таблицы записывается большее значение, полученное путем сравнения значений из соответствующих ячеек исходных *Q*-таблиц. Полученную *Q*-таблицу используют для тестирования на обеих имеющихся картах, фиксируют результаты, анализируют и сравнивают результаты использования исходных *Q*-таблиц и *Q*-таблицы, сформированной с применением предобучения.

В рамках эксперимента проводили исследования для различного числа эпизодов обучения: 100, 200, 300. Для каждого количества эпизодов обучения исследовали наборы коэффициентов, представленные в таблице. Число тестовых эпизодов на каждом этапе эксперимента было равно 10.

Рассматриваемые наборы коэффициентов

α	γ	ε
0,3	0,3	0,1
0,3	0,3	0,7
0,3	0,9	0,1
0,3	0,9	0,7
0,9	0,3	0,1
0,9	0,3	0,7
0,9	0,9	0,1
0,9	0,9	0,7

В рамках каждой серии экспериментов подсчитаны средние показатели по обучению и по тестированию, построены графики зависимости винрейта от каждого из коэффициентов при остальных фиксированных.

Результаты эксперимента. При определении эффективности использования данных предобученных структур главной задачей являлось сравнение среднего винрейта и средней награды при первичном тестировании и при тестировании с использованием *Q*-таблицы, полученной из таблиц первичного тестирования. Поведение винрейта в зависимости от коэффициентов служит второстепенной задачей, позволяющей взглянуть на процесс подробнее.

Для серии экспериментов с числом эпизодов 100 были подсчитаны и выявлены следующие критерии и закономерности:

– средний винрейт при тестировании с первичными *Q*-таблицами — 0,04375;

– средний винрейт при тестировании с *Q*-таблицей, полученной по методу максимума, — 0,05;

– средняя награда при тестировании с первичными *Q*-таблицами — 4,00605;

– средняя награда при тестировании с *Q*-таблицей, полученной по методу максимума, — 3,41008.

По результатам экспериментов можно сделать следующие выводы.

1. Средний винрейт повысился на 14,3 %, а средняя награда уменьшилась на 14,8 % относительно результатов тестирования первичных *Q*-таблиц.

Одна из троек графиков зависимости винрейтов от одного коэффициента при двух фиксированных построена с помощью пакета Microsoft Excel и представлена на рис. 2.

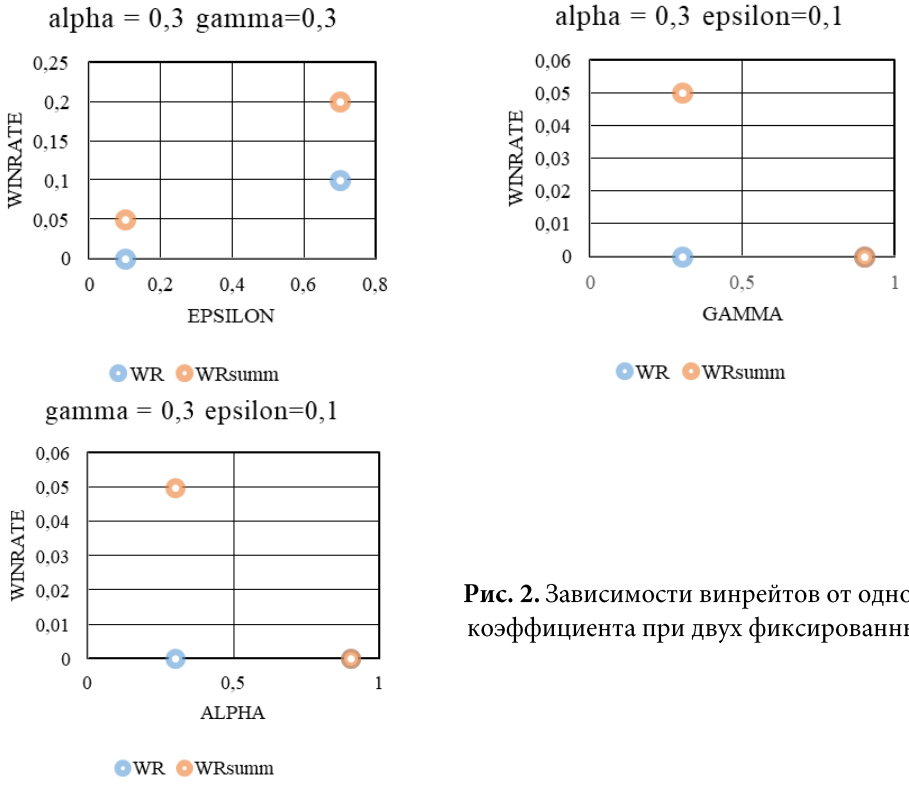

Рис. 2. Зависимости винрейтов от одного коэффициента при двух фиксированных

2. Прослеживается сравнительно (со средним) высокий винрейт при малом значении α.

3. В среднем увеличение ε положительно влияет на винрейт. Это можно объяснить тем, что при увеличении ε растет доля неслучайного выбора действия, а значит, чаще происходит выбор действия в соответствии с большей наградой в *Q*-таблице.

4. В среднем рост γ негативно влияет на винрейт. Увеличение γ обеспечивает бóльшую ориентацию на значение будущей выгоды.

В соответствии с результатами эксперимента тройка коэффициентов $\alpha = 0{,}3$, $\gamma = 0{,}3$, $\varepsilon = 0{,}7$ определена как самая оптимальная, поскольку при данных значениях самый высокий винрейт наблюдается и при первичном обучении, и при обучении с использованием данных предобучения.

Для серии экспериментов с количеством эпизодов 200 были подсчитаны и выявлены следующие критерии и закономерности:

– средний винрейт при тестировании с первичными *Q*-таблицами — 0;

– средний винрейт при тестировании с *Q*-таблицей, полученной по методу максимума, — 0,01875;

– средняя награда при тестировании с первичными *Q*-таблицами — 2,688306;

– средняя награда при тестировании с *Q*-таблицей, полученной по методу максимума, — 3,497581.

По результатам экспериментов можно сделать следующие выводы.

1. Глобально средний винрейт повысился до 1,9 %. А средняя награда увеличилась на 30,1 % относительно тестирования первичных *Q*-таблиц.

Одна из троек графиков зависимости винрейтов от одного коэффициента при двух фиксированных построена с помощью пакета Microsoft Excel представлена на рис. 3

2. Увеличение винрейта относительно первичного значения было зафиксировано при тройках коэффициентов $\alpha = 0{,}3$, $\gamma = 0{,}3$, $\varepsilon = 0{,}1$ и $\alpha = 0{,}9$, $\gamma = 0{,}9$, $\varepsilon = 0{,}1$, т. е. при одновременном увеличении коэффициентов α и γ. Винрейт при второй тройке выше.

3. В среднем увеличение ε отрицательно влияет на винрейт. Это может быть объяснимо недостаточностью в заполненности *Q*-таблицы на первых эпизодах, ведь чем больше случайных попыток — тем лучше откалибрована *Q*-таблица в начале эксперимента.

4. В среднем рост γ позитивно влияет на винрейт. Увеличение γ обеспечивает бóльшую ориентацию на значение будущей выгоды.

Результаты эксперимента с числом эпизодов 200 показали буквально нулевой винрейт при первичном обучении, однако использование данных предобучения позволило добиться винрейта, отличного от нуля. В соответствии с результатами эксперимента тройка коэффициентов $\alpha = 0{,}9$, $\gamma = 0{,}9$, $\varepsilon = 0{,}1$ определена как самая оптимальная, поскольку при данных значениях самый высокий винрейт наблюдается при обучении с использованием данных из предобучения.

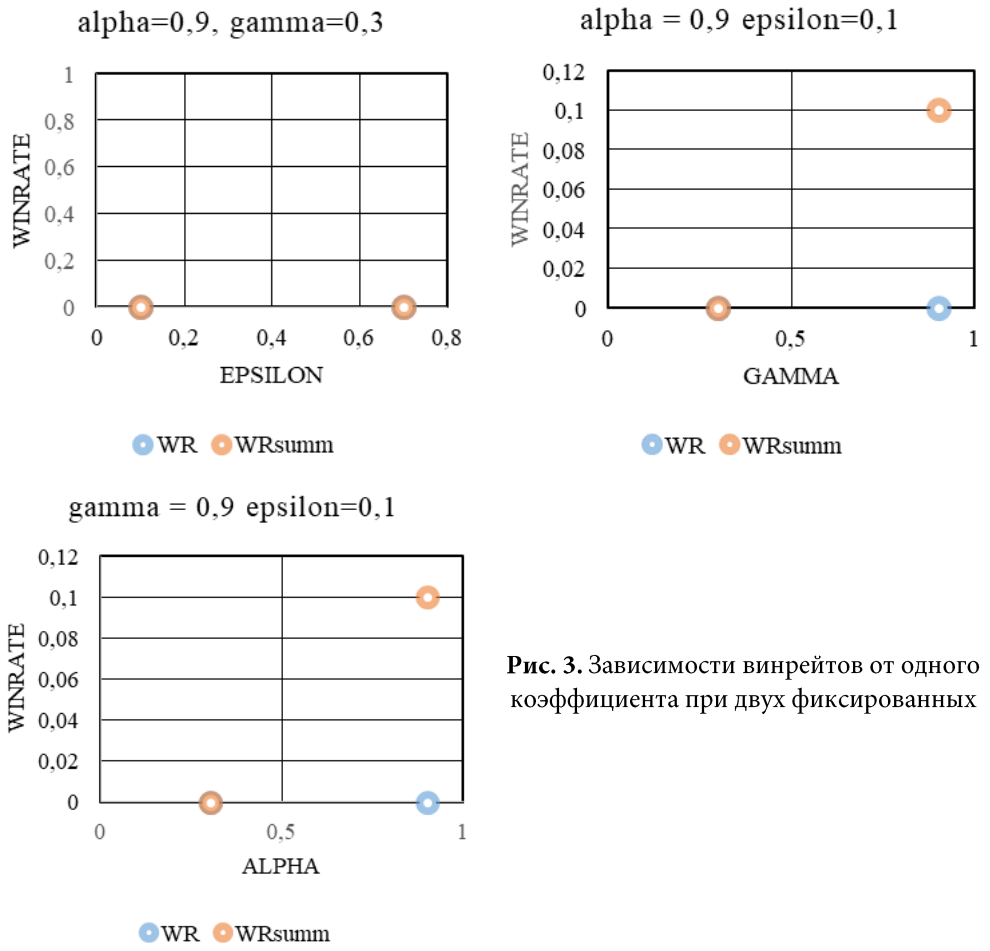

Рис. 3. Зависимости винрейтов от одного коэффициента при двух фиксированных

Для серии экспериментов с числом эпизодов 300 были подсчитаны и выявлены следующие критерии и закономерности:

– средний винрейт при тестировании с первичными *Q*-таблицами — 0,01875;

– средний винрейт при тестировании с *Q*-таблицей, полученной по методу максимума, — 0,025;

– средняя награда при тестировании с первичными *Q*-таблицами — 3,105645;

– средняя награда при тестировании с *Q*-таблицей, полученной по методу максимума, — 3,414113.

По результатам экспериментов можно сделать следующие выводы.

1. Средний винрейт повысился на 33,3 %, средняя награда увеличилась на 9,9% относительно тестирования первичных *Q*-таблиц.

Одна из троек графиков зависимости винрейтов от одного коэффициента при двух фиксированных построена с помощью пакета Microsoft Excel представлена на рис. 4.

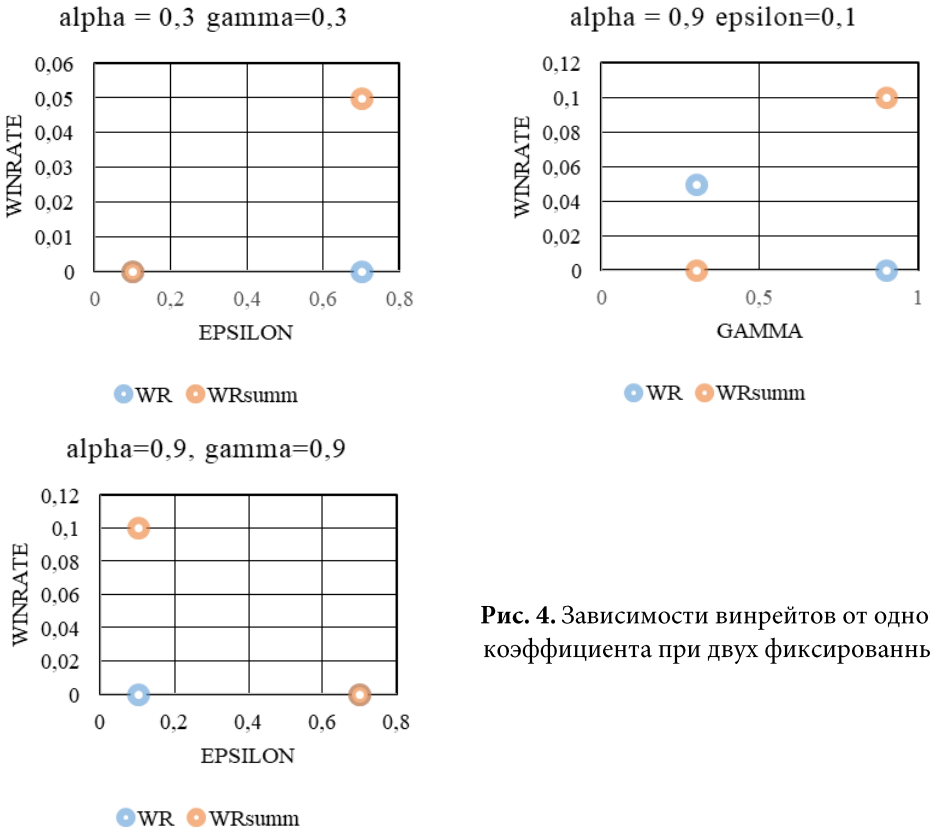

Рис. 4. Зависимости винрейтов от одного коэффициента при двух фиксированных

Для серии экспериментов с числом эпизодов 300 были подсчитаны и выявлены следующие критерии и закономерности.

1. Увеличение винрейта относительно первичного было зафиксировано при тройках коэффициентов $\alpha = 0{,}3$, $\gamma = 0{,}3$, $\varepsilon = 0{,}7$ и $\alpha = 0{,}9$, $\gamma = 0{,}9$, $\varepsilon = 0{,}1$, т. е. при одновременном увеличении коэффициентов α и γ. При этом винрейт при второй тройке выше.

2. В среднем увеличение ε отрицательно влияет на винрейт, однако и при малом ε наблюдался рост винрейта относительно тестирования первичных *Q*-таблиц. Это может быть объяснимо недостаточностью в заполненности *Q*-таблицы на первых эпизодах, ведь чем больше случайных попыток — тем лучше откалибрована *Q*-таблица вначале.

3. В среднем рост γ позитивно влияет на винрейт, однако и при малом γ наблюдался рост винрейта относительно результатов тестирования первичных *Q*-таблиц. С увеличением γ обеспечивает бóльшую ориентацию на значение будущей выгоды.

Результаты эксперимента с числом эпизодов 300 показали буквально нулевой винрейт при первичном обучении, однако использование данных предобучения позволило добиться винрейта, отличного от нуля. В соответствии с результатами эксперимента тройка коэффициентов $\alpha = 0{,}9$, $\gamma = 0{,}9$, $\varepsilon = 0{,}1$ определена как самая оптимальная, поскольку при данных значениях самый высокий винрейт наблюдается при обучении с использованием данных предобучения. Аналогичный вывод был сделан для 200 эпизодов обучения.

Выводы. На вопрос об эффективности использования данных предобученных структур в данной серии экспериментов был получен однозначный положительный ответ: для каждого числа эпизодов в итоге наблюдался рост среднего винрейта, в двух сериях из трех также наблюдалось и увеличение награды. Рассмотренные закономерности влияния ключевых коэффициентов обучения на винрейт можно применять для дальнейшего более эффективного обращения с предобученными структурами, по крайней мере, теми, в которые используется представленная композиция методов [10]. Вне всяких сомнений, мета-обучение как обучение высокоуровневых структур может поддерживать многообразие решений, также имеет смысл исследовать возможности применения этого подхода к решению масштабных задач.

Исходный код, карты и подробные результаты доступны по ссылке: https://github.com/karinoizerr/Meta-on-Q-learning

Литература

[1] Шарден Б., Массарон Л., Боскетти А. Крупномасштабное машинное обучение вместе с Python. М., DMK-Press, 2018.

[2] Stadie B.C., Yang G., Houthooft R., et al. Some considerations on learning to explore via meta-reinforcement learning. URL: https://arxiv.org/abs/1803.01118 (дата обращения: 15.02.2021).

[3] Коэльо Л.П., Ричард В. Построение систем машинного обучения на языке Python. М., ДМК Пресс, 2016.

[4] Nilsson N.J. Introduction to machine learning. URL: https://ai.stanford.edu/ ~nilsson/ MLBOOK.pdf (дата обращения: 15.02.2021).

[5] A StarCraft II bot API client library for Python 3. *github.com: веб-сайт.* URL: https://github.com/ Dentosal/python-sc2 (дата обращения: 15.02.2021).

[6] Getting started with Gym. *gym.openai.com: веб-сайт.* URL: https://gym.openai.com/docs (дата обращения: 15.02.2021).

[7] Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М., ДМК Пресс, 2015.

[8] Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. М., Диалектика, 2017.

[9] Саттон Р.С., Барто Э.Г. Обучение с подкреплением. М., Бином, 2012.

[10] Schweighofer N. Doya K. Meta-learning in reinforcement learning. *Neural Netw.*, 2003, vol. 16, no. 1, pp. 5–9. DOI: https://doi.org/10.1016/S0893-6080(02)00228-9

Смирнова Карина Дмитриевна — студентка кафедры «Информационные системы и телекоммуникации», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Научный руководитель — Алфимцев Александр Николаевич, доктор технических наук, профессор кафедры «Информационные системы и телекоммуникации», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Ссылку на эту статью просим оформлять следующим образом:
Смирнова К.Д. Метод глубокого метамультиагентного машинного обучения, основанный на принципе максимума. *Политехнический молодежный журнал*, 2021, № 04(57). http://dx.doi.org/ 10.18698/ 2541-8009-2021-04-694

DEEP METAMULTI-AGENT MACHINE LEARNING METHOD BASED ON THE MAXIMUM PRINCIPLE

K.D. Smirnova
iron.karina@gmail.com
SPIN-code: 7120-0782

Bauman Moscow State Technical University, Moscow, Russian Federation

Abstract

The concept of machine learning is considered and the target task is formulated as the search for the most universal teaching methods. The article gives a description of the discrete medium used in the experiment and the agents acting in it. The Q-learning method on which the developed algorithm is based is described. The results are presented of an experiment with a conceptual machine learning algorithm based on the maximum principle, where the algorithm used data from pretrained structures within the framework of meta-learning. The operation of the algorithm with different sets of parameters is analyzed. The regularities of the influence of the key training parameters on the result are investigated and the prospects for their use are considered.

Keywords

Machine learning, reinforcement learning, meta learning, high-level system, Q-learning, multi-agent learning, discount factor, learning rate

Received 08.04.2021

References

[1] Sjardin B., Massaron L., Boschetti A. Large scale machine learning with Python. Packt Publ., 2016. (Russ. ed.: Krupnomasshtabnoe mashinnoe obuchenie vmeste s Python. Moscow, DMK-Press, 2018.)

[2] Stadie B.C., Yang G., Houthooft R., et al. Some considerations on learning to explore via meta-reinforcement learning. URL: https://arxiv.org/abs/1803.01118 (accessed: 15.02.2021).

[3] Coelho L.P., Richert W. Building machine learning systems with Python. Packt Publ., 2015. (Russ. ed.: Postroenie sistem mashinnogo obucheniya na yazyke Python. Moscow, DMK Press Publ., 2016.)

[4] Nilsson N.J. Introduction to machine learning. URL: https://ai.stanford.edu/~nilsson/MLBOOK.pdf (accessed: 15.02.2021).

[5] A StarCraft II bot API client library for Python 3. *github.com: website.* URL: https://github.com/Dentosal/python-sc2 (accessed: 15.02.2021).

[6] Getting started with Gym. *gym.openai.com: website.* URL: https://gym.openai.com/docs (accessed: 15.02.2021).

[7] Flach P. Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press, 2012. (Russ. ed.: Mashinnoe obuchenie. Nauka i iskusstvo postroeniya algoritmov, kotorye izvlekayut znaniya iz dannych. Moscow, DMK Press Publ., 2015.)

[8] Müller A.C., Guido S. Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, 2016. (Russ. ed.: Vvedenie v mashinnoe obuchenie s pomoshch'yu Python. Rukovodstvo dlya spetsialistov po rabote s dannymi. Moscow, Dialektika Publ., 2017.)

[9] Sutton R.S. Reinforcement learning. MIT Press, 1998. (Russ. ed.: Obuchenie s podkrepleniem. Moscow, Binom Publ., 2012.)
[10] Schweighofer N. Doya K. Meta-learning in reinforcement learning. *Neural Netw.*, 2003, vol. 16, no. 1, pp. 5–9. DOI: https://doi.org/10.1016/S0893-6080(02)00228-9

Smirnova K.D. — Student, Department of Information Systems and Telecommunications, Bauman Moscow State Technical University, Moscow, Russian Federation.

Scientific advisor — Alfimtsev A.N., Dr. Sc. (Eng.), Assoc. Professor, Department of Information Systems and Telecommunications, Bauman Moscow State Technical University, Moscow, Russian Federation.

Please cite this article in English as:
Smirnova K.D. Deep metamulti-agent machine learning method based on the maximum principle. *Politekhnicheskiy molodezhnyy zhurnal* [Politechnical student journal], 2021, no. 04(57). http://dx.doi.org/ 10.18698/2541-8009-2021-04-694.html (in Russ.).