# ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ ЭВОЛЮЦИОННЫХ АЛГОРИТМОВ ПРИ ПОМОЩИ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ В НЕСТАЦИОНАРНОЙ СРЕДЕ[1]

***И. А. Петрова***
*студентка кафедры компьютерных технологий Университета ИТМО*
*E-mail: petrova@rain.ifmo.ru*

***А. С. Буздалова***
*студентка кафедры компьютерных технологий Университета ИТМО*
*E-mail: abuzdalova@gmail.com*

***М. В. Буздалов***
*ассистент кафедры компьютерных технологий Университета ИТМО*
*E-mail: mbuzdalov@gmail.com*

**Аннотация.** Одним из способов повысить эффективность алгоритмов однокритериальной оптимизации является использование вспомогательных критериев. В эволюционных алгоритмах для выбора вспомогательного критерия в процессе оптимизации существует метод EA+RL, основанный на обучении с подкреплением. В предыдущих исследованиях использовались алгоритмы обучения с подкреплением в стационарной среде. Однако если свойства вспомогательных критериев меняются в процессе оптимизации, может быть более эффективно использовать алгоритмы обучения с подкреплением в нестационарной среде.

В данной работе представлены результаты начальных исследований в области использования EA+RL в нестационарной среде. Предложен новый подход к обучению с подкреплением в нестационарной среде. Приводятся результаты его применения к модельной задаче, также проводится сравнение с ранее применявшимися алгоритмами обучения с подкреплением.

## Введение

Существуют методы повышения эффективности эволюционных алгоритмов при помощи вспомогательных критериев [1, 2]. Оптимизируемый критерий может выбираться случайным образом [1] или при помощи некоторой эвристики [3]. Первый метод может быть применен для решения любой задачи с конечным набором вспомогательным критериев, но при этом не учитывает специфику задачи, а второй может быть применен для решения лишь конкретной задачи. Метод EA+RL не имеет этих недостатков [4].

[1] Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01).

В методе EA+RL для выбора вспомогательного критерия, используемого в качестве функции приспособленности (ФП) на данном шаге алгоритма, применяется обучение с подкреплением [5]. Эволюционный алгоритм (ЭА) выступает в роли среды обучения. эффективность данного метода была продемонстрирована и теоретически доказана на ряде задач [4, 6]. Предполагалось, что среда стационарна и поэтому применялись алгоритмы обучения в стационарной среде. Однако в случае когда свойства вспомогательных критериев зависят от этапа оптимизации, возможно эффективнее использовать обучение с подкреплением в нестационарной среде.

В обучении с подкреплением агент обучения применяет действие к среде, в результате чего среда переходит в новое состояние и возвращает агенту численную награду. В методе EA+RL действием является выбор ФП — целевой или одной из вспомогательных. Цель обучения с подкреплением — максимизация суммарной награды. В EA+RL в качестве награды выбирается разность значений целевой функции на лучшей особи в текущем и предыдущем поколениях ЭА. Поэтому при максимизации награды также максимизируется и значение целевой функии. Стоит отметить, что задача оптимизации вспомогательных ФП не ставится, они используются лишь для более быстрой оптимизации целевой ФП.

Существуют различные методы обучения с подкреплением в нестационарной среде [7]. Одним из наиболее эффективных методов, который может быть применим к выбору вспомогательных критериев, является алгоритм RLCD [8]. Однако результаты применения алгоритма RLCD для решения описанной ниже задачи оказались хуже, чем результаты, полученные с помощью методов, применявшихся ранее. Поэтому был предложен новый метод, описанный в следующем разделе.

## Описание предлагаемого подхода

В новом подходе, в отличие от RLCD, используется алгоритм ε-жадного $Q$-обучения [5]. Также как в алгоритме классического $Q$-обучения, на каждой итерации агент применяет действие $a$ к среде, находящейся в состоянии $s$. Затем обновляется значение ожидаемой награды $Q(s, a)$. В отличие от классического алгоритма $Q$-обучения при условии, что $Q(s, a) - Q(s', a') < \eth$ для какой-то пары $(s, a)$ и $(s', a')$, обучение начинается заново. Перезапуск обучения связан с тем, что в описанном случае ожидаемая награда примерно одинакова для хотя бы одной пары действий и агент не может определить, какое из них более эффективно.

## Модельная задача

Рассмотрим постановку модельной задачи с двумя вспомогательными критериями, которые могут быть как эффективными так и неэффективными на разных этапах оптимизации. В этой задаче особи представляются битовыми строками длины $n$. Пусть $x$ — число бит, равных единице. Целевая

ФП задается формулой $g(x)=\left\lfloor \frac{x}{k} \right\rfloor$, где $k$ — константа, $k<n$. Необходимо максимизировать значение целевой ФП. Вспомогательные ФП имеют следующий вид:

$$h_1=\begin{cases} x,\ x \le p_1 \\ p_1,\ p_1 < x \le p_2 \\ x,\ p_2 < x \le p_3 \\ p_3,\ p_3 < x \le p_4 \\ \dots \\ x,\ p_s < x \le n \end{cases} \qquad h_2=\begin{cases} p_1,\ x \le p_1 \\ x,\ p_1 < x \le p_2 \\ x,\ p_3 < x \le p_3 \\ p_3,\ x < x \le p_4 \\ \dots \\ n,\ p_s < x \le n \end{cases}.$$

В точках $p_i$ вспомогательные ФП меняют свои свойства, будем называть их *точками переключения*. Вспомогательный критерий $h_1$ эффективен, когда $x \in [0, p_1]$, $(p_2, p_3]$, …, $(p_s, n]$, а $h_2$ эффективен во всех других случаях. Отметим, что использование правильного вспомогательного критерия позволяет различить особи с одинаковым значением целевой ФП, и выбрать ту, в которой содержится большее число единиц. Такая особь с большей вероятностью породит особь с более высоким значением целевой ФП.

## Описание эксперимента

В ходе экспериментов сравнивались результаты применения метода EA+RL с использованием различных алгоритмов обучения с подкреплением к различным конфигурациям модельной задачи. Результаты работы каждого алгоритма усреднялись за 100 запусков. Рассматривались конфигурации задачи с пятью и десятью точками переключения. Точки переключения располагались равномерно по длине особи. В качестве параметра $k$ были выбраны значения 10 и 25.

Поколение ЭА состояло из 100 особей. Оператор мутации изменял каждый бит с вероятностью 0.001. Оператор скрещивания[9] применялся с вероятностью 0.7. Выполнение ЭА останавливалось при достижении заданного числа итераций или максимального значения целевой ФП.

В качестве алгоритмов обучения с подкреплением использовались ε-жадное $Q$-обучение, отложенное $Q$-обучение[10] и предлагаемый подход. В качестве параметров для ε-жадного $Q$-обучения использовались: $\alpha=0.6$, $\gamma = 0.01$, $\varepsilon=0.03$ [4]. Отложенное $Q$-обучение использовалось с параметрами: $\alpha=0.6$, $\gamma=0.01$, $\varepsilon=0.4$, $m=5$ [4]. Параметры для предлагаемого подхода были выбраны в ходе предварительных экспериментов: $\alpha=0.6$, $\gamma=0.01$, $\varepsilon=0.0$ и $\delta=0.001$. Состояния представлялись в виде вектора порядковых номеров ФП, упорядоченных по значению $(f(x_c) — f(x_p))/f(x_c)$, где $f$ — ФП, $x_p$ — число бит, равных единице в лучшей особи предыдущего поколения, $x_c$ — число бит, равных единице в лучшей особи текущего поколения [1].

## Результаты экспериментов

Результаты экспериментов представлены в Таблице 1. В первой колонке указано число точек переключения, во второй и третьей — число итераций и длина особи соответственно. В последних трех колонках указаны средние значения целевой ФП, полученные при использовании предлагаемого подхода, ε-жадного $Q$-обучения и отложенного $Q$-обучения соответственно. Среднеквадратичное отклонение при использовании первых двух алгоритмов составило около 0.5%, при использовании отложенного $Q$-обучения около 20%. Для всех рассмотренных конфигураций модельной задачи результаты, полученные при использовании предложенного подхода, превосходят результаты существующих алгоритмов.

Для проверки того, что новый алгоритм отличим от предыдущих, был проведен статистический тест Уилкоксона [11]. Значения *p-value*, полученные при сравнении нового подхода с ε-жадным $Q$-обучением и отложенным $Q$-обучением, приведены в скобках в соответствующих колонках. Можно видеть, что в случаях, когда длина особи превышала 1000, *p-value*, полученные при сравнении нового подхода с ε-жадным $Q$-обучением, малы, что говорит о различимости этих подходов. Однако новый метод не всегда статистически различим с отложенным $Q$-обучением, несмотря на то, что среднее значение целевой ФП, полученное с помощью нового подхода, значительно выше. Это можно объяснить большим разбросом значений целевой ФП при применении отложенного $Q$-обучения.

Агент может выбирать на каждой итерации одну из вспомогательных ФП или целевую ФП. Наиболее эффективно выбирать ту ФП, которая в текущем поколении равна $x$. Будем называть выбор эффективной ФП *хорошим.* На Рис. 1 представлено число выборов ФП в ходе решения модельной задаче с пятью точками переключения, $k=10$, $n=750$. По горизонтали указан номер итерации, а ширина полосы соответствует числу выборов соответствующей

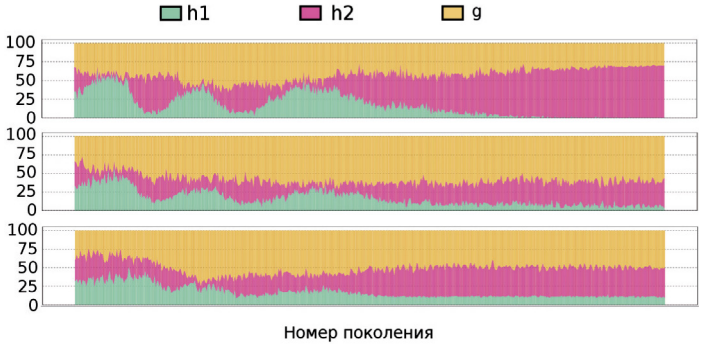


**Рис. 1.** Число выборов ФП при использовании нового метода (сверху), ε-жадного $Q$-обучения (в середине), отложенного $Q$-обучения (внизу)

Таблица 1

**Результаты экспериментов**

| **Число $p_i$** | **Число итераций** | **Длина** | **Новый метод** | **ε-жадное $Q$-обучение** | **Отложенное $Q$-обучение** |
|---|---|---|---|---|---|
| | | | $k = 10$ | | |
| 5 | 3000 | 750 | 74,52 | 74,49 (0,34) | 68,39 ($4{,}4\times10^{-10}$) |
| | | 1000 | 99,55 | 99,47 (0,13) | 87,39 ($2{,}2\times10^{-16}$) |
| | | 1250 | 124,44 | 124,19 ($8{,}3\times10^{-4}$) | 113,69 ($2{,}2\times10^{-16}$) |
| | | 1500 | 149,03 | 148,02 ($1{,}5\times10^{-3}$) | 133,5 ($8{,}1\times10^{-15}$) |
| | | 1750 | 173,98 | 173,63 ($8{,}4\times10^{-8}$) | 156,93 ($1{,}9\times10^{-13}$) |
| | 5000 | 2000 | 198,93 | 197,98 ($2{,}2\times10^{-16}$) | 186,37 ($9{,}5\times10^{-14}$) |
| | | 2250 | 222,01 | 220,23 ($7{,}2\times10^{-11}$) | 202,80 ($1{,}5\times10^{-3}$) |
| | | 2500 | 245,52 | 244,55 ($3\times10^{-3}$) | 232,81 (0,99) |
| 10 | 5000 | 2000 | 198,94 | 198,34 ($1{,}8\times10^{-12}$) | 170,59 ($2{,}7\times10^{-13}$) |
| | | 2250 | 223,36 | 220,79 ($2{,}2\times10^{-16}$) | 184,47 ($1{,}1\times10^{-12}$) |
| | | 2500 | 245,28 | 244,61 ($1{,}2\times10^{-4}$) | 204,42 (0,72) |
| | 9000 | 2750 | 269,38 | 269,14 ($5\times10^{-3}$) | 226,14 (0,57) |
| | | 3000 | 294,22 | 293,73 ($2{,}8\times10^{-5}$) | 249,96 (0,96) |
| | | 3250 | 318,92 | 318,70 (0,014) | 268,66 (0,97) |
| | | 3500 | 343,79 | 343,33 ($4\times10^{-5}$) | 285,76 (0,99) |
| | | 3750 | 368,52 | 367,90 ($1{,}3\times10^{-5}$) | 307,72 (0,99) |
| | | | $k = 25$ | | |
| 5 | 3000 | 750 | 29,45 | 29,40 (0,25) | 26,01 ($2{,}2\times10^{-16}$) |
| | | 1000 | 39,18 | 39,14 (0,22) | 36,20 ($8{,}9\times10^{-14}$) |
| | | 1250 | 49,02 | 49,00 (0,24) | 45,86 ($2{,}8\times10^{-9}$) |
| | | 1500 | 59,00 | 58,92 ($6\times10^{-3}$) | 54,77 ($4{,}3\times10^{-8}$) |
| | | 1750 | 68,96 | 68,02 ($4\times10^{-15}$) | 61,17 ($1{,}8\times10^{-11}$) |
| | 5000 | 2000 | 78,17 | 77,23 ($4{,}9\times10^{-16}$) | 70,54 (0,19) |
| | | 2250 | 87,30 | 87,12 ($8\times10^{-3}$) | 79,70 (0,98) |
| | | 2500 | 97,01 | 96,89 (0,004) | 84,62 (0,53) |

Таблица 2

**Число выборов ФП**

| Число $p_i$ | Длина | Число хороших выборов, % | | |
|---|---|---|---|---|
| | | Новый метод | ε-жадное $Q$-обучение | Отложенное $Q$-обучение |
| 5 | 750 | 62 | 50 ($1{,}3\times10^{-5}$) | 46 ($8\times10^{-3}$) |
| | 1000 | 55 | 51 (0,01) | 47 (0,26) |
| | 1250 | 51 | 43 ($7\times10^{-4}$) | 36 ($1{,}7\times10^{-5}$) |
| | 1500 | 50 | 39 ($2{,}2\times10^{-16}$) | 35 ($1{,}01\times10^{-10}$) |
| | 1750 | 48 | 39 ($2{,}2\times10^{-16}$) | 37 ($9{,}2\times10^{-10}$) |
| | 2000 | 46 | 39 ($2{,}2\times10^{-16}$) | 29 ($2{,}1\times10^{-15}$) |
| | 2250 | 37 | 23 ($2{,}2\times10^{-16}$) | 37 ($8{,}5\times10^{-7}$) |
| | 2500 | 30 | 17 ($2{,}2\times10^{-16}$) | 26 ($4{,}1\times10^{-14}$) |
| 10 | 2000 | 47 | 49 ($1{,}3\times10^{-8}$) | 33 ($6{,}9\times10^{-13}$) |
| | 2250 | 42 | 29 ($2{,}2\times10^{-16}$) | 36 ($3{,}6\times10^{-6}$) |
| | 2500 | 26 | 19 ($2{,}2\times10^{-16}$) | 38 ($1{,}2\times10^{-4}$) |
| | 2750 | 26 | 21 ($2{,}2\times10^{-16}$) | 41 ($3{,}9\times10^{-3}$) |
| | 3000 | 31 | 26 ($2{,}2\times10^{-16}$) | 33 ($5{,}1\times10^{-7}$) |
| | 3250 | 37 | 29 ($2{,}2\times10^{-16}$) | 36 ($1{,}6\times10^{-5}$) |
| | 3500 | 41 | 33 ($2{,}2\times10^{-16}$) | 38 ($4{,}6\times10^{-5}$) |
| | 3750 | 43 | 35 ($2{,}2\times10^{-16}$) | 37 ($4{,}6\times10^{-5}$) |

ФП в 100 запусках. Можно заметить, что новый подход делает хороший выбор чаще, чем ε-жадное $Q$-обучение, и ε-жадное $Q$-обучение делает хороший выбор чаще, чем отложенное $Q$-обучение.

В Таблице 2 представлен усредненный процент числа хороших выборов ФП для конфигураций со значением параметра $k=10$. При $k=25$ результаты аналогичны и для краткости не представлены. В первой колонке указано число точек переключения, во второй — длина особи. В последних трех колонках указаны средние значения числа выборов хорошей ФП в процентах от общего числа выборов ФП, полученного при использовании предлагаемого подхода, ε-жадного $Q$-обучения и отложенного $Q$-обучения соответственно. Среднеквадратичное отклонение при использовании первых двух алгоритмов составило около 8%, при использовании отложенного $Q$-обучения — около 100%.

Можно видеть, что предлагаемый подход делает хорошие выборы чаще, чем ε-жадное $Q$-обучение. Однако существуют конфигурации задачи, на которых отложенное $Q$-обучение делает больше хороших выборов, чем новый

подход. В то же время среднее значение целевой ФП, полученное при применении отложенного $Q$-обучения хуже, чем при применении нового метода. Это можно объяснить тем, что среднеквадратичное отклонение для отложенного $Q$-обучения гораздо больше, чем для нового подхода.

В последних двух колонках Таблицы 2 в скобках указаны результаты сравнения нового метода с соответствующими алгоритмами, проведенного с помощью теста Уилкоксона. Можно видеть, что новый метод отличим от существующих для всех рассмотренных конфигураций задачи.

## Заключение

В работе предложен новый подход к обучению с подкреплением, который может быть использован в методе EA+RL. Данный подход применим в случае нестационарности, заключающейся в изменении свойств вспомогательных критериев в зависимости от этапа оптимизации. Предлагаемый подход был применен для решения модельной задачи. Полученные результаты превосходят результаты работы ε-жадного $Q$-обучения и отложенного $Q$-обучения.

### Литература


1. *Jensen M. T*. Reducing the Run-time Complexity of Multiobjective EAs: The NSGA-II and Other Algorithms. Transactions on Evolutionary Computation. 2003. P. 503–515.
2. *Knowles J. D., Watson R. A., Corne D*. Reducing Local Optima in Single-Objective Problems by Multi-objectivization // In Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization. 2001. P. 269–283.
3. *Lochtefeld D. F., Ciarallo F. W*. Deterministic Helper-Objective Sequences Applied to Job-Shop Scheduling // In Proceedings of Genetic and Evolutionary Computation Conference. 2010. P. 431–438.
4. *Afanasyeva A., Buzdalov M*. Optimization with Auxiliary Criteria using Evolutionary Algorithms and Reinforcement Learning // In Proceedings of 18th International Conference on Soft Computing MENDEL. 2012. P. 58–63.
5. *Sutton R. S., Barto A. G*. Reinforcement Learning: An Introduction // MIT Press, Cambridge, MA, USA. 1998.
6. *Buzdalov M., Buzdalova A., Shalyto A*. A First Step towards the Runtime Analysis of Evolutionary Algorithm Adjusted with Reinforcement Learning // In Proceedings of the International Conference on Machine Learning and Applications. 2013. Vol. 1. P. 203–208.
7. *Granmo O.-C., Berg S*. Solving non-stationary bandit problems by random sampling from sibling kalman filters // In IEA/AIE. 2010. P. 199–208.
8. *B. C. da Silva, Basso E. W., Bazzan A. L. C., Engel P. M*. Dealing with non-stationary environments using context detection // In Proceedings of the 23rd International Conference on Machine Learning, ICML'06. 2006. P. 217–224.
9. *Strehl A. L., Li L., Wiewiora E., Langford J., Littman M. L*. PAC Model-free Reinforcement Learning // In Proceedings of the 23rd International Conference on Machine Learning. 2006. P. 881–888.

10. *Arkhipov V., Buzdalov M., Shalyto A*. Worst-Case Execution Time Test Generation for Augmenting Path Maximum Flow Algorithms using Genetic Algorithms // In Proceedings of the International Conference on Machine Learning and Applications. 2013. Vol. 2. P. 108–111.
11. *Derrac J., Garcia S., Molina D., Herrera F*. A practical tutorial on the use of non-parametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms // Swarm and Evolutionary Computation. 2011. P. 3–18.