УДК 004.855 DOI: 10.18698/2541-8009-2020-06-620

ПОДХОДЫ К УСОВЕРШЕНСТВОВАНИЮ МАШИННОГО ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ НА ОСНОВЕ ВНУТРЕННЕЙ МОТИВАЦИИ

А.В. Балицкая balitskayaanna@yandex.ru
SPIN-код: 8245-8151

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

Аннотация

На сегодняшний день обучение с подкреплением является одним из самых перспективных направлений машинного обучения. Однако возникает ряд задач (среди которых можно упомянуть абстрагирование от действий или изучение окружающей среды с редкими вознаграждениями), которые могут быть решены с помощью внутренней мотивации. Внутренняя мотивация побуждает агента участвовать в исследованиях, играх и других видах деятельности, вызванных любопытством, в отсутствие внешних вознаграждений. Способность эффективно самообучаться является одним из признаков интеллекта и позволяет агенту успешно функционировать в течение длительного периода времени в динамичных, сложных средах, о которых имеется мало априорных знаний. В статье представлен обзор о роли внутренней мотивации и описаны подходы по улучшению обучения агента на ее основе.

Ключевые слова

Машинное обучение с подкреплением, мультиагентное обучение, алгоритмы внутренней мотивации, глубокое обучение, нейронные сети, агенты, поведенческая психология, Starcraft, SMAC

Поступила в редакцию 20.05.2020

Введение. Термин «обучение с подкреплением» взят из работ известного русского физиолога, нобелевского лауреата Ивана Петровича Павлова (в англоязычной литературе принято сокращение RL — reinforcement learning). Активные исследования в области математического описания обучения с подкреплением начались в середине 50-х годов XX в. Особый вклад в эти исследования внесли Ричард Саттон и Эндрю Барто, которые разработали базовые алгоритмы и концепции на основании своего опыта изучения процесса обучения у животных. Именно они предложили классическую архитектуру «агент – критик» (*Actor – Critic*) [1]. В соответствии с ней агент (*Actor*) генерирует некоторые действия(*Actions*) в контексте состояний (*States*)окружающей среды (*Environment*), влияющие на изменение состояний этой среды с течением времени (рис. 1). Среда, в свою очередь, содержит критика (*Critic*), который предоставляет агенту на каждом временном шаге оценку (*Rewards*) его текущего поведения, т. е. с каждым процессом связан его алфавит восприятий и реакций. Критик сопоставляет состояния окружающей среды (или пару состояние – действие) с числовыми сигналами вознаграждения [2].

Цель агента — действовать в каждый момент времени таким образом, чтобы максимизировать меру общего количества вознаграждения, которое он ожидает получить в будущем. Эта мера может быть простой суммой сигналов вознаграждения, которую он ожидает получить в будущем, или чаще — дисконтированной суммой, в которой более поздние сигналы вознаграждения взвешиваются меньше, чем более ранние. Сигнал, который критик передает агенту, соответствует тому, что психологи называют первичным вознаграждением, обычно подразумевая вознаграждение, которое поощряет поведение, непосредственно связанное с выживанием и размножением организма [3].

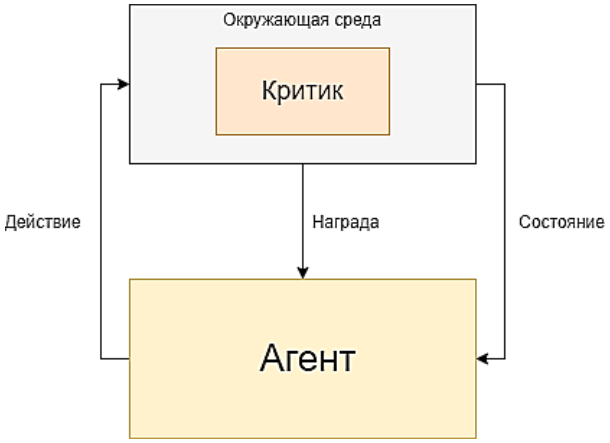

Рис. 1. Взаимодействие «агент – среда» в RL. Первичные сигналы поступают агенту от критика в его окружении

Однако данный подход сильно упрощен по сравнению с реальной моделью поведения живых организмов и ограничивает область применения такого агента на практике, поэтому появилась идея внедрить внутреннюю мотивацию.

Внутренняя мотивация в машинном обучении с подкреплением. Концепция внутренней мотивации была введена в 1950-х годах в психологии животных [4] и получила дальнейшее развитие в психологии человека [5], где она в настоящее время широко применяется. Например, студент, выполняющий свою домашнюю работу по математике, поскольку он считает, что это интересно, внутренне мотивирован, в то время как его одногруппник делает это, чтобы получить хорошую оценку, мотивирован внешне. Точно так же игра в компьютер для развлечения является внутренней мотивацией, а вот участие в телевизионном игровом шоу с целью заработать деньги является внешней мотивацией. Понятия внутренней и внешней мотивации относятся к причине действия, и не следует путать с интернальностью и экстернальностью, которые относятся к местоположению награды.

Смысл внутренней мотивации состоит в том, чтобы подтолкнуть агента к получению определенного поведения без какого-либо прямого вмешательства со стороны окружающей среды. Подобное поведение не направлено на достижение целей, связанных с базовыми потребностями организма. Речь идет о том,

чтобы сделать что-то для своего внутреннего удовлетворения (руководствуясь, например, любопытством), а не получить награду, назначенную окружающей средой. Сигналы вознаграждения агента определяются процессами в его мозге, которые контролируют не только его внешнее состояние, но и внутреннее (рис. 2). Критик в голове агента делает это путем разделения среды на внешнюю и внутреннюю среду. Внешняя среда представляет собой то, что находится вне агента или робота (будем называть это организмом, как обозначено на рисунке), тогда как внутренняя среда состоит из компонентов, находящихся внутри организма, и содержит критика, который определяет первичное вознаграждение.

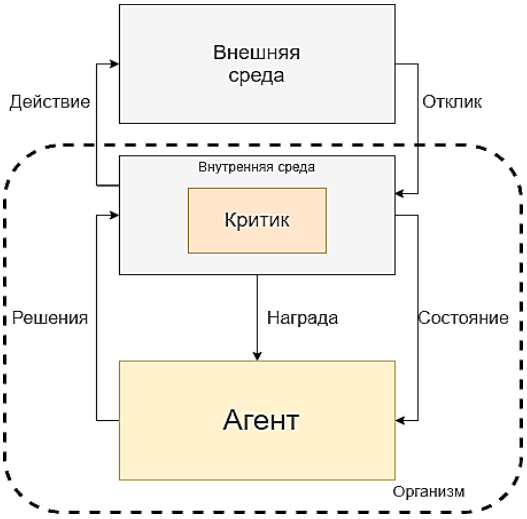

Рис. 2. Взаимодействие «агент – среда» в RL

Подобное уточнение классической структуры RL позволяет приблизить математическую модель к реальному поведению организма. Но внутренняя мотивация должна на чем-то основываться, поэтому были сформулированы три варианта мотивации: новизна, неожиданность и наращивание.

Новизна. Агент, взаимодействуя со средой, в итоге начинает переходить в принципиально новые состояния среды, в которые он обычно никогда не переходит. Подобное явление служит для агента проявлением новизны. Он получает внутренний бонус за это, тем самым побуждая себя к исследованию новых состояний среды [6]. Однако отметим, что по мере того как агент переходит в состояние, внутреннее вознаграждение, связанное с этим состоянием, уменьшается. Алгоритм внутренней мотивации на основе новизны использует подсчет посещений состояния [7]. Он может быть формализован с помощью следующей формулы:

$$R_{int}(s_t) = \frac{1}{N(s_t)},$$

где $N(s_t)$ — количество посещений данного состояния.

Данный метод довольно эффективен в табличной среде (среде с дискретным пространством состояний), но плохо применим, когда состояния многочисленны или непрерывны, поскольку агент с малой вероятностью возвращается в одно и то же состояние.

Неожиданность. Рассмотрим агента, который видит наблюдение x_t, совершает действие a_t и переходит в следующее состояние с наблюдением x_{t+1}. Мы хотим поощрить этого агента вознаграждением r_t, связанным с тем, насколько информативным был переход [8]. С учетом переходного кортежа $\{x_t, a_t, x_{t+1}\}$ вознаграждение за разведку тогда определяется как

$$r_t = -\log p(\varphi(x_{t+1}) \mid x_t, a_t).$$

В случае Q-learning, одном из алгоритмов реализации метода машинного обучения с подкреплением, агент, выполняя какое-то действие в окружающей среде, предсказывает последствие. Если реакция среды значительно отличается от реакции, предсказанной агентом, то агент запоминает комбинацию действий, которая привела к такому эффекту. Таким образом, подобное несоответствие может привлечь внимание агента, поскольку ставит под сомнение его предыдущие знания, и он начинает более активно вести себя в среде.

Наращивание. Подход, связанный с расширением возможностей агента, был разработан, чтобы ответить на вопрос о том, существует ли некоторая локальная функция полезности, которая делает возможным выживание организма в целом. Эта гипотетическая функция должна быть локальной в том смысле, что она не изменяет поведение организма в очень долгосрочной перспективе, и индуцированное поведение должно помогать выживанию видов [9]. Расширение возможностей измеряется как емкость канала, связывающего действия и наблюдения агента. Взаимная информация — это основная теоретико-информационная величина, которая действует как общая мера зависимости между двумя случайными переменными *x* и*y* и определяется как

$$\mathcal{L}(x, y) = \mathbb{E}_{p(y|x)p(x)}\left[\log\left(\frac{p(x,y)}{p(x)p(y)}\right)\right],$$

где $p(x, y)$ — совместное распределение по случайным переменным; $p(x)$, $p(y)$ — соответствующие предельные распределения.

Для внутренней мотивации используется внутренняя мера вознаграждения, называемая расширением возможностей, которая получается путем поиска максимальной взаимной информации $\mathcal{L}(x, y)$, обусловленной начальным состоянием *s* между последовательностью действий *K* и достигнутым конечным состоянием *s*′:

$$\mathcal{E}(s) = \max_w \mathcal{L}^w(a,\ s'|s) = \max_w \mathbb{E}_{p(s'|a,s)(a|s)}\left[\log\left(\frac{p(a,s' \mid s)}{w(a \mid s)p(s' \mid s)}\right)\right], \qquad (*)$$

где $a = \{a_1, \ldots, a_K\}$ — последовательность*K* примитивных действий a_k, приводящих к конечному состоянию *s*′; $p(s' \mid a, s)$ — вероятность перехода среды на

следующий шаг; $p(a, s' \mid s)$ — совместное распределение последовательностей действий и конечного состояния, $w(a \mid s)$ — распределение последовательностей действий на каждом шаге; $p(s' \mid s)$ — совместная вероятность, маргинальная по последовательности действий.

Формула (*) является определением пропускной способности канала в теории информации и представляет собой меру количества информации, содержащейся в последовательности действий относительно будущих состояний.

Максимизация возможностей — это поиск состояния, в котором агент имеет наибольший контроль над окружающей средой. Примером наращивания возможностей может являться стремление занять центральную позицию на карте или же получение навыка, который помогает агенту наиболее эффективно взаимодействовать с окружающей средой.

Эксперимент. Для тестирования подходов к мотивационному обучению была создана карта в игре Starcraft II, а для описания действий агентов использована библиотека SMAC. Игра StarCraftII часто используется в качестве исследовательской платформы для искусственного интеллекта и тестирования алгоритмов машинного обучения. Как правило, игра оформляется как конкурентная проблема: агент берет на себя роль человека-игрока, принимая управленческие решения и отдавая приказы отдельным подразделениям, зависящим от централизованного контроллера. Библиотека SMAC позволяет смоделировать качественно сложную среду, которая объединяет элементы частичной наблюдаемости, сложной динамики и высокомерных пространств наблюдения. SMAC состоит из набора микро-сценариев StarCraft II, которые направлены на оценку того, насколько хорошо независимые агенты могут научиться координации для решения сложных задач. Каждый такой сценарий — это противостояние двух армий. Исходное положение, количество и тип подразделений в каждой армии варьируются от сценария к сценарию, как и наличие либо отсутствие возвышенной или проходимой местности. На рис. 3 представлены скриншоты нескольких микросценариев SMAC.

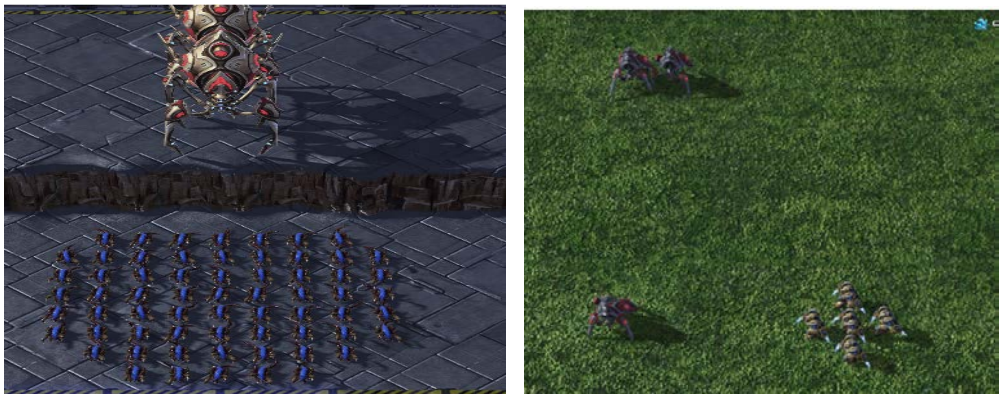

Рис.3. Скриншоты SMAC-сценариев

Первая армия контролируется обучаемыми союзными агентами. Вторая армия состоит из вражеских подразделений, управляемых встроенным игровым искусственным интеллектом (ИИ), который использует тщательно разработанные вручную необучаемые эвристики. В начале каждого эпизода игровой ИИ инструктирует свои подразделения атаковать союзных агентов, используя свои скриптовые стратегии. Эпизод заканчивается, когда все подразделения одной из армий погибли или же когда достигнут заранее заданный лимит времени (в этом случае игра считается поражением союзных агентов). Цель состоит в том, чтобы максимизировать коэффициент выигрыша, т. е. соотношение выигранных игр к сыгранным (win_rate) [10].

На каждом временном шаге агенты получают локальные наблюдения, сделанные в пределах их поля зрения. Наблюдения включают в себя информацию о карте в пределах круговой области вокруг каждой единицы измерения с радиусом, равным высотному диапазону. Дальность видимости делает окружающую среду частично наблюдаемой с точки зрения любого агента. Агенты могут наблюдать за другими агентами только в том случае, если они оба живы и находятся в пределах видимости.

Вектор признаков, наблюдаемый каждым агентом, содержит следующие атрибуты, одинаковые как для союзных, так и для вражеских единиц в пределах видимости: расстояние, относительные координаты *x*,относительные координаты *y*, состояние здоровья, щит и тип агента.

Дискретный набор действий, которые разрешено выполнять агентам, состоит из перемещения [направление], атаки [enemy_id], остановки и no-op (нет действий). Максимальное число действий, которое может предпринять агент, находится в диапазоне от 7 до 70 (в зависимости от сценария).

Общая цель состоит в том, чтобы максимизировать коэффициент выигрыша для каждого сценария битвы. Мастерство вражеской армии определяется конкретным типом боевых единиц врага. Зерги, зилоты, морпехи имеют разный уровень мастерства, и от этого зависит, какой размер награды будет получен от среды. По умолчанию используется награда, основанная на нанесенном вражескому агенту уроне и мастерстве вражеской армии, а также специальный бонус за победу в битве. Точные значения и масштабы для каждого из этих событий можно настроить с помощью диапазона флагов.

Для эксперимента в редакторе карт был создан простейший сценарий с участием двух агентов. Цель каждого из агентов — убить противника. Схема карты для этого сценария представлена на рис.4.

В данном эксперименте используется метод Q-learning. На основе получаемого от среды вознаграждения агент формирует функцию полезности Q, что впоследствии дает ему возможность уже не случайно выбирать стратегию поведения, а учитывать опыт предыдущего взаимодействия со средой.

Реализации изложенных выше подходов к улучшению обучения агента на основе мотивации были скорректированы с учетом выбранной окружающей среды и цели мини-игры. Награждение за новизну агент получал в том случае, если начинал посещать места на карте, которые до этого не посещал.

Рис.4. Экспериментальная карта

Принцип неожиданности был реализован как ошибка предсказания, т. е. как разница между таблицами Qpredict и Qtarget. Модуль предсказания был реализован с помощью простейшего фильтра Калмана.

Принцип наращивания был реализован как получение навыка, который помогает агенту быстрее убить врага. В данной игре агент получает поощрение, если у него открывается возможность убить противника и он действительно в этот момент осуществляет выстрел.

В каждом из этих случаев размер бонусного вознаграждения находился в пределах 0,15…0,35. Такой диапазон был получен эмпирическим путем. Обучение агентов происходило на 150 эпизодах игры с гиперпараметрами $\alpha = 0,5$ и $\gamma = 0,9$. Всего для каждого принципа было сформировано по пять итоговых Q-таблиц. Тестирование проводилось на 100 эпизодах и исходя из этого формировалась величина win_rate, представляющая собой отношение выигранных битв к общему числу игр. Результаты эксперимента представлены на рис. 5, где по оси *y*—величина win_rate, а по оси *x*— номер игры. Как видно из графика, win_rate обучения с использованием того или иного подхода мотивации лучше, чем без мотивации.

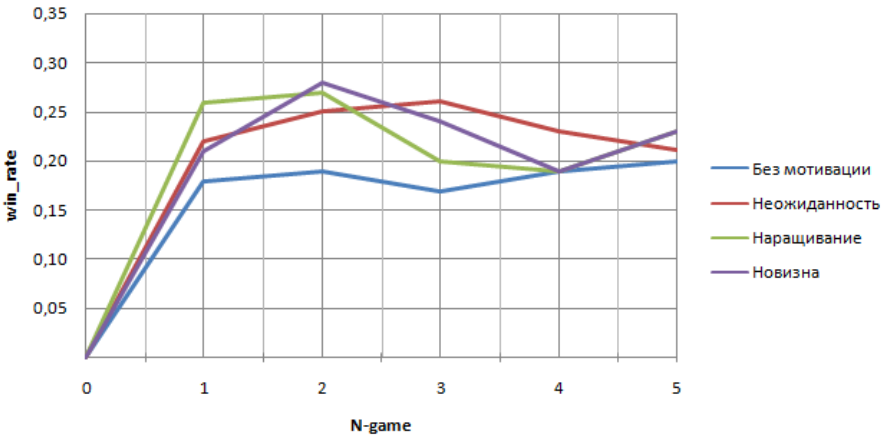

Рис.5. График эксперимента

Заключение. В настоящей статье описаны основные концепции и понятия в машинном обучении с подкреплением. Также описаны подходы по улучшению обучения агента с использованием внутренней мотивации, основанные на таких понятиях, как неожиданность, новизна и наращивание. Эти базовые концепции могут быть усложнены, например, при использовании других методов подсчета исследуемых величин или нейронных сетей.

Литература

[1] Алфимцев А.Н. Нечеткий процессно-ориентированный подход к недетерминированному проектированию интеллектуальных мультимодальных интерфейсов. *Наука и образование: научное издание*, 2012, № 11. URL: https://elibrary.ru/download/elibrary_18381185_41681497.pdf (дата обращения: 05.03.2020).

[2] Алфимцев А.Н. Декларативно-процессная технология разработки интеллектуальных мультимодальных интерфейсов. Автореф. дис. ... док. тех. наук. М., ИПУ РАН, 2016.

[3] Barto A.G., Sutton R.S. Landmark learning: an illustration of associative search. *Biol. Cybern.*, 1981, vol. 42, no. 1, pp. 1–8. DOI: https://doi.org/10.1007/BF00335152

[4] Harlow H.F. Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. *J. Comp. Physiol. Psychol.*, 1950, vol. 43, no. 4, pp. 289–294. DOI: https://doi.apa.org/doi/10.1037/h0058114

[5] Deci E. Intrinsic motivation. Plenum, 1975.

[6] Burda Y., Edwards H., Pathak D., et al. Large-scale study of curiosity-driven learning. *arxiv.org: веб-сайт*. URL: https://arxiv.org/abs/1808.04355 (дата обращения: 18.02.2020).

[7] Montúfar G., Ghazi-Zahedi K., Ay N. Information theoretically aided reinforcement learning for embodied agents. *arxiv.org: веб-сайт*. URL: https://arxiv.org/abs/1605.09735 (дата обращения: 18.02.2020).

[8] Achiam J., Sastry Sh. Surprise-based intrinsic motivation for deep reinforcement learning. *arxiv.org: веб-сайт*. URL: https://arxiv.org/abs/1703.01732 (дата обращения: 18.02.2020).

[9] Mohamed S., Rezende D.J. Variational information maximisation for intrinsically motivated reinforcement learning. *arxiv.org: веб-сайт*. URL: https://arxiv.org/abs/1509.08731 (дата обращения: 18.02.2020).

[10] Vinyals O., Ewalds T., Bartunov S., et al. StarCraft II: a new challenge for reinforcement learning. *arxiv.org: веб-сайт*. URL: https://arxiv.org/abs/1708.04782 (дата обращения: 18.02.2020).

Балицкая Анна Владимировна—магистрант 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:
Балицкая А.В. Подходы к усовершенствованию машинного обучения с подкреплением на основе внутренней мотивации. *Политехнический молодежный журнал*, 2020, № 06(47). http://dx.doi.org/ 10.18698/2541-8009-2020-06-620

APPROACHES TO IMPROVING MACHINE LEARNING WITH REINFORCEMENT BASED ON INTRINSIC MOTIVATION

A.V. Balitskaya

balitskayaanna@yandex.ru
SPIN-code: 8245-8151

Bauman Moscow State Technical University, Moscow, Russian Federation

Abstract

Today, reinforced learning is one of the most promising areas of machine learning. However, a number of problems arise (among which we can mention abstraction from actions or studying the environment with rare rewards), which can be solved with the help of intrinsic motivation. Intrinsic motivation encourages an agent to engage in research, games, and other activities caused by curiosity in the absence of external rewards. The ability to effectively self-learn is one of the hallmarks of intelligence and allows the agent to function successfully for a long period in dynamic, complex environments about which there is little prior knowledge. The article provides an overview of the role of intrinsic motivation and describes approaches to improving the training of an agent based on it.

Keywords

Reinforced machine learning, multi-agent learning, intrinsic motivation algorithms, deep learning, neural networks, agents, behavioral psychology, Starcraft, SMAC

Received 20.05.2020

References

[1] Alfimtsev A.N. Fuzzy process-oriented approach to nondeterministic design of intelligent multimodal interfaces. *Nauka i obrazovanie: nauchnoe izdanie* [Science and Education: Scientific Publication], 2012, no. 11 (in Russ.). URL: https://elibrary.ru/download/ elibrary_18381185_41681497.pdf (accessed: 05.03.2020).

[2] Alfimtsev A.N. Deklarativno-protsessnaya tekhnologiya razrabotki intellektual'nykh mul'timodal'nykh interfeysov. Avtoref. dis. dok. tekh. nauk [Declarative-process technology of developing intelligent multimode interfaces. Abs. doc. tech. sci. diss.]. Moscow, IPU RAS Publ., 2016 (in Russ.).

[3] Barto A.G., Sutton R.S. Landmark learning: an illustration of associative search. *Biol. Cyb*ern., 1981, vol. 42, no. 1, pp. 1–8. DOI: https://doi.org/10.1007/BF00335152

[4] Harlow H.F. Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. *J. Comp. Physiol. Psychol.*, 1950, vol. 43, no. 4, pp. 289–294. DOI: https://doi.apa.org/doi/10.1037/h0058114

[5] Deci E. Intrinsic motivation. Plenum, 1975.

[6] Burda Y., Edwards H., Pathak D., et al. Large-scale study of curiosity-driven learning. arxiv.org: website. URL: https://arxiv.org/abs/1808.04355 (accessed: 18.02.2020).

[7] Montúfar G., Ghazi-Zahedi K., Ay N. Information theoretically aided reinforcement learning for embodied agents. *arxiv.org: website*. URL: https://arxiv.org/abs/1605.09735 (accessed: 18.02.2020).

[8] Achiam J., Sastry Sh. Surprise-based intrinsic motivation for deep reinforcement learning. *arxiv.org: website*. URL: https://arxiv.org/abs/1703.01732 (accessed: 18.02.2020).

[9] Mohamed S., Rezende D.J. Variational information maximisation for intrinsically motivated reinforcement learning. *arxiv.org: website*. URL: https://arxiv.org/abs/1509.08731 (accessed: 18.02.2020).
[10] Vinyals O., Ewalds T., Bartunov S., et al. StarCraft II: a new challenge for reinforcement learning. *arxiv.org: website*. URL: https://arxiv.org/abs/1708.04782 (accessed: 18.02.2020).

Balitskaya A.V. — Master's Degree Student, Department of Information Systems and Telecommunications, Bauman Moscow State Technical University, Moscow, Russian Federation.

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

Please cite this article in English as:
Balitskaya A.V. Approaches to improving machine learning with reinforcement based on intrinsic motivation. *Politekhnicheskiy molodezhnyy zhurnal* [Politechnical student journal], 2020, no. 06(47). http://dx.doi.org/ 10.18698/2541-8009-2020-06-620.html (in Russ.).