УДК 519.85
DOI: https://doi.org/
Поступила в редакцию 15.01.2022
Подписана в печать 22.04.2022
ISSN 1995-5499

ИСПОЛЬЗОВАНИЕ МЕТОДОВ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ В ЗАДАЧАХ МЕДИЦИНСКОЙ ПРАКТИКИ

© 2022 М. В. Демченко, И. Л. Каширина✉, М. А. Фирюлина

Воронежский государственный университет
Университетская пл., 1, 394018 Воронеж, Российская Федерация

Аннотация. В статье обсуждаются особенности разработки современных методов обучения с подкреплением в задачах медицинской направленности. Методы обучения с подкреплением являются популярным инструментом машинного обучения, применяемым в задачах поиска оптимальных стратегий лечения пациентов, персонализированной медицины, а также интерактивных систем наблюдения за пациентами. При этом важной задачей является выбор оптимального алгоритма обучения с подкреплением из множества существующих на данный момент методов, обладающих своей спецификой применения, преимуществами и недостатками. Данная статья посвящена анализу алгоритмического аппарата наиболее популярных методов обучения с подкреплением и содержит примеры результатов работы рассматриваемых методов в контексте задачи поиска оптимальных схем лечения для кардиологических пациентов.
Ключевые слова: обучение с подкреплением, марковский процесс, динамическое программирование, уравнение Беллмана, итерация по стратегиям, итерация по значениям, Монте-Карло, метод временных различий, SARSA, Q-Learning.

ВВЕДЕНИЕ

Современные возможности сбора и хранения информации позволяют оперировать большими массивами данных, накопленными по всему миру и отражающими практический опыт и процессы, протекающие в различных сферах жизни общества. В частности, имеющееся в настоящее время в доступе количество медицинской информации позволяет в полном объеме получить сведения о течении тех или иных заболеваний, диагностике и медицинских предписаниях по лечению. Полученные сведения являются ценным источником данных для анализа, интерпретации обнаруженных закономерностей, валидации и совершенствования существующих методов диагностики, а также разработке новых эффективных способов лечения.

Оптимальным инструментом обработки таких массивов данных является применение методов машинного обучения. Подходы машинного обучения включают такие группы методов, как обучение с учителем, обучение без учителя и обучение с подкреплением. Методы обучения с учителем выполняют предсказательную функцию и нацелены на прогнозирование значений на основании обучения на размеченных наборах данных. Методы обучения без учителя носят описательный характер и ориентированы на раскрытие структуры и закономерностей в наборах данных.

Обучение с подкреплением, в свою очередь, представляет собой целенаправленное обучение, которое производится путем взаимодействия условного агента с некоторой средой, в процессе которого направление действий агента корректируется на основе полученного опыта. Последовательность действий агента с вероятностью их выбора в некотором состоянии формируют *стратегию* агента.

✉ Каширина Ирина Леонидовна
e-mail: kash.irina@mail.ru

В зависимости от цели агента выделяют 2 типа задач обучения с подкреплением.

1) Прогнозирование — задача оценки качества изучаемой стратегии (например, оценки эффективности назначаемой пациенту стратегии лечения).

2)Управление — задача поиска оптимальной стратегии. При этом, алгоритм управления подразумевает также решение задачи прогнозирования, т. е. включает вычисление как оптимальной стратегии, так и ожидаемого значения функции полезности. Примером задачи управления в медицине может являться разработка протокола индивидуального лечения пациента.

Для решения задач прогнозирования и управления было разработано множество алгоритмов обучения с подкреплением, наиболее популярными из которых являются методы динамического программирования (итерации по стратегиям и по значениям) и методы, основанные на семплинге (Монте-Карло, метод временных различий, Q-learning, SARSA, Expected SARSA).

Эффективность применения данных моделей в задачах медицинской направленности доказывают множество исследований и практических результатов, в частности, в таких крупных направлениях, как персонифицированная медицина, назначение оптимальных схем лечения и интерактивные системы наблюдения за пациентом.

Персонифицированная медицина является актуальным направлением современной медицины, преимуществом которой является учет индивидуальных характеристик и генетических особенностей пациентов при назначении лечения. Целью персонифицированной медицины является разработка и применение лечения, являющегося оптимальным для определенных фенотипов пациентов. Для решения этой задачи активно используются алгоритмы обучения с подкреплением. В частности, работа [1] посвящена исследованию индивидуальной дозировки эритропротеина при проведении гемодиализа с помощью такого метода обучения с подкреплением, как Q-Learning.

Назначение оптимальных схем лечения предполагает разработку индивидуального плана лечения определенного заболевания, представляющего последовательность предписаний (процедур, назначений препаратов и других рекомендаций), автоматически корректирующихся в зависимости от текущего состояния пациента. В частности, исследования [2, 3] посвящены применению алгоритмов обучения с подкреплением для динамического назначения лечения диабета и сепсиса.

Интерактивные системы динамического наблюдения за пациентом позволяют предупредить обострение хронических заболеваний. Современный уровень развития информационных технологий позволяет непрерывно осуществлять контроль значимых показателей состояния и получать обратную связь от пациента. Обучение с подкреплением используется в современных mobile health системах [4], с помощью чего становится возможным осуществлять непрерывный контроль за ходом лечения, физическим и эмоциональным состоянием пациента.

1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

1.1. Марковский процесс

Обучение с подкреплением является одной из наиболее активно развивающихся областей машинного обучения, представляющих собой вычислительный подход к обучению, при котором целью *агента* (алгоритм, искусственный интеллект) является максимизация суммарного *вознаграждения*, которое он получает во время взаимодействия с окружением, как правило, сложным и неопределенным. [5] На каждом шаге выбор агента из множества *действий* определен некоторым *состоянием*. Переход агента из одного состояния в другое определяется *вероятностью перехода*. Данный подход описывает марковский процесс (рис. 1).

Суммарное вознаграждение G_t является суммой вознаграждений, полученных после шага t. Если число шагов t является конеч-

Рис. 1. Марковский процесс
[Fig. 1. Markov decision process]

ным, т. е. $t = \overline{1,T}$, где T — финальный шаг, то процесс взаимодействия агента со средой естественным образом разбивается на *эпизоды*, т. е. суммарное вознаграждение определяется формулой (1):

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \ldots + R_T. \qquad (1)$$

В противном случае, задача является *непрерывной* и предполагает, что взаимодействие агента со средой производится бесконечное число шагов. Для того чтобы обеспечить сходимость суммы вознаграждений в непрерывных задачах, а также для учета значимости вознаграждений, полученных на различных шагах эпизода, вводится параметр дисконта γ, что отражено в формуле (2):

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = R_{t+1} + \gamma G_{t+1}. \quad (2)$$

Следовательно, если $\gamma = 0$, то учитываются только мгновенные вознаграждения, если $\gamma = 1$, то, напротив, большую значимость принимают вознаграждения, полученные в длительной перспективе.

1.2. Функции полезности и уравнения Беллмана

Целью агента в задаче обучения с подкреплением является максимизация функции полезности, представляющей собой условное математическое ожидание, которое отражает среднее вознаграждение агента, находящегося в состоянии $s \in S$ (3), или при выборе действия $a \in A(s)$ (4), где:

– S — заданное множество состояний, включающее промежуточные и конечные (*terminal*) состояния;

– $A(s)$ — множество доступных действий из состояния s;

– $p(s', r \mid s, a)$ — вероятность перехода из состояния s при выборе действия a в состояние s' с вознаграждением r;

– $\pi(a \mid s)$ — стратегия агента, т. е. вероятность выбора агентом действия $a \in A(s)$ из состояния $s \in S$: $\sum_{a \in A(s)} \pi(a \mid s) = 1,\ \pi(a \mid s) \geq 0.$

$$V_\pi(s) = E_\pi[G_t \mid S_t = s] = \\ = \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a)[r + \gamma v_\pi(s')], \qquad (3)$$

$$Q_\pi(s, a) = E_\pi[G_t \mid S_t = s, A_t = a] = \\ = \sum_{s'} \sum_r p(s', r \mid s, a)[r + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s', a')]. \qquad (4)$$

1.3. Стратегии обучения с подкреплением. Дилемма эксплуатации и использования

Одной из проблем обучения с подкреплением является поиск оптимального соотношения исследования и эксплуатации (дилемма исследования и эксплуатации).

Следование оптимальной на данном шаге последовательности действий является *жадной стратегией*. Жадный выбор действий предполагает принятие решений, обеспечивающих максимальную ожидаемую оценку результата, и является *политикой эксплуатации*, следуя которой можно получить максимально возможное на текущий момент вознаграждение.

Политика исследования, напротив, предполагает выбор действий, отличных от заведомо оптимальных, с целью того, чтобы извлечь новую информацию о ранее не изученных действиях (с возможным уменьшением текущего вознаграждения).

Например, в реальных условиях предписания врача по лечению пациента не являются случайными. Часто план лечения представляет собой стратегию, которая считается наилучшей, согласно медицинской практике, т. е. это жадная стратегия, которая является проверенной, безопасной и эффективной для пациента. Примером политики исследования является испытание нового способа лечения.

Недостаток эксплуатации в худшем случае может привести к тому, что лечение пациента

ранее не изученными способами не принесет положительного результата. В случае исключения исследования, назначения не будут включать потенциально более эффективные, но ранее мало изученные методы.

Примером стратегии, позволяющей оптимальным образом совмещать эксплуатацию и исследование, является ε-жадная стратегия (5), согласно которой с вероятностью ε может быть выбрана неоптимальная стратегия.

$$A^* \leftarrow \arg\max_a Q(s,a)$$
$$\pi(a \mid s) = \begin{cases} 1-\varepsilon, & a = A^* \\ \varepsilon, & a \neq A^* \end{cases} \quad (5)$$

2. МАТЕРИАЛЫ И МЕТОДЫ

2.1. Динамическое программирование

При условии наличия описанной модели марковского процесса, для решения задач управления с помощью алгоритмов обучения с подкреплением используются методы итерации по стратегиям и итерации по значениям.

Метод итерации по стратегиям включает фазы *оценивания* и *улучшения*. В процессе оценивания алгоритм осуществляет обновления значений функции полезности, а фаза улучшения подразумевает вычисление оптимальной стратегии. Таким образом, на каждой итерации оценивания и улучшения алгоритм сходится к оптимальной стратегии.

Данный алгоритм иллюстрирует рис. 2.

Алгоритм итерации по значениям во многом аналогичен итерации по стратегиям, за исключением того, что фаза улучшения фиксированной политики интегрируется в правило обновления значений функции полезности: обновления производятся в направлении действий, максимизирующих текущую оценку. Алгоритм этого метода обучения с подкреплением имеет следующий вид.

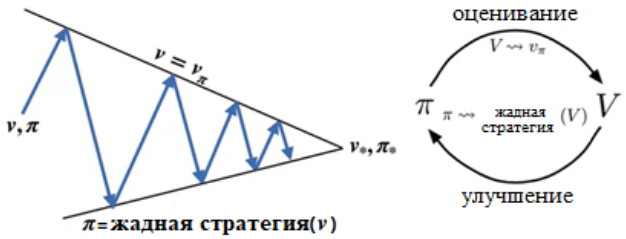

Рис. 2. Метод итераций по стратегиям
[Fig. 2. Policy iteration]

Шаг 1. Инициализация параметра $\theta > 0$ для определения точности оценки.

Шаг 2. Инициализация $V(s)$ $\forall s \in S$ случайными значениями, $V(terminal) = 0$.

Шаг 3. Выполнять:

$\Delta \leftarrow 0$;

Для каждого $s \in S$:

$v \leftarrow V(s)$;

$V(s) \leftarrow \max_a \sum_{s',r} p(s',r \mid s,a)[r + \gamma V(s')]$;

$\Delta \leftarrow \max(\Delta, | v - V(s) |)$;

Пока $\Delta < \theta$;

Шаг 4. Вычислить детерминированную стратегию $\pi \approx \pi_*$ по правилу:

$$\pi(s) = \operatorname{argmax}_a \sum_{s',r} p(s',r \mid s,a)[r + \gamma V(s')].$$

Алгоритмы обучения с подкреплением на основе *динамического программирования* зачастую демонстрируют высокую эффективность в задачах медицинской направленности. Например, данный подход применялся при разработке метода лечения эпилепсии в [6]. Также методы динамического программирования применялись для разработки стратегии лечения сепсиса в исследовании [7].

Необходимым условием применения методов динамического программирования в задачах обучения с подкреплением является наличие модели изучаемого процесса, включая вероятности перехода между состояниями $p(s',r \mid s,a)$ (например, вероятности изменения состояния пациента в контексте прикладных медицинских задач).

В случаях, когда полная модель процесса отсутствует, или формирование такой модели может привести к избыточным вычислениям, используется группа алгоритмов, основанных на семплинге.

2.2. Методы, основанные на семплинге

2.2.1. Метод Монте-Карло

Метод Монте-Карло в широком понимании является подходом к изучению случайных процессов. Данный метод подразумевает вычисление характеристик множества случайных выборок, на основании которых

возможно выявить вероятностные свойства изучаемого процесса. Свойства метода Монте-Карло:

– адаптируется к изменению вознаграждений и корректирует оценку функции полезности и используемую стратегию только в конце эпизода (т.к. использует усредненную сумму всех накопленных в течение эпизода вознаграждений);

– зачастую обеспечивает быструю сходимость, если эпизод состоит из короткой последовательности шагов.

Метод *Монте-Карло* применим в задачах, где конечный исход эпизода является более приоритетной оценкой, по сравнению с промежуточными вознаграждениями, наблюдаемыми в течение эпизода. Например, к таким относится задача испытаний действия препарата пролонгированного действия, где показателем эффективности является улучшение состояния пациента, наблюдаемое через некоторый интервал времени после начала лечения. Данный метод также используют в задачах разработки медицинских рекомендаций для пациентов интенсивной терапии [8], где приоритетной оценкой эффективности алгоритма является улучшение состояния пациента по итогу лечения.

Также метод Монте-Карло проявляет высокую производительность в задачах, в которых эпизод состоит из последовательности шагов и действий малой размерности. [9]. Алгоритмическая схема метода обучения с подкреплением для решения задачи прогнозирования на основе метода Монте-Карло имеет вид.

Шаг 1. Входные данные: оцениваемая стратегия π.

Шаг 2. Инициализация $V(s)$ $\forall s \in S$ случайными значениями.

Шаг 3. $Returns(s)$ — пустой список $\forall s \in S$.

Шаг 4.

Для каждого эпизода:

Сгенерировать эпизод, следуя стратегии π: $S_0,\ A_0,\ R_1,\ S_1,\ A_1,\ R_2,\ \ldots,\ S_{T-1},\ A_{T-1},\ R_T$;

$G \leftarrow 0$;

Для каждого шага эпизода, $t = T-1,\ T-2,\ \ldots,\ 0$:

$G \leftarrow \gamma G + R_{t+1}$;

Добавить G в список $Returns(S_t)$;

$V(S_t) \leftarrow \text{average}(Returns(S_t))$;

Алгоритмическая схема метода обучения с подкреплением для решения задачи управления на основе метода Монте-Карло имеет вид.

Шаг 1. Инициализация входного параметра алгоритма: $\varepsilon > 0$.

Шаг 2. Инициализация π случайными значениями (ε-гибкая стратегия).

Шаг 3. Инициализация $Q(s,a) \in \mathbb{R}$ $\forall s \in S$, $a \in A(s)$ случайными значениями.

Шаг 4. $Returns(s)$ — пустой список $\forall s \in S$, $a \in A(s)$;

Шаг 5.

Для каждого эпизода:

Сгенерировать эпизод, следуя стратегии π: $S_0,\ A_0,\ R_1,\ S_1,\ A_1,\ R_2,\ \ldots,\ S_{T-1},\ A_{T-1},\ R_T$;

$G \leftarrow 0$;

Для каждого шага эпизода, $t = T-1,\ \ldots,\ 0$:

$G \leftarrow \gamma G + R_{t+1}$;

Добавить G в список $Returns(S_t, A_t)$;

$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$;

$A^* \leftarrow \text{argmax}_a Q(S_t, a)$;

Для каждого $a \in A(S_t)$:

$$\pi(a \mid S_t) \leftarrow \begin{cases} 1 - \varepsilon / |A(S_t)|, & \text{если } a = A^* \\ \varepsilon / |A(S_t)|, & \text{если } a \neq A^* \end{cases}.$$

2.2.2. Метод временных различий

Данный метод [10], является одним из фундаментальных подходов и центральных идей теории обучения с подкреплением.

Свойства метода временных различий:

1) Корректировка функции полезности производится не в конце эпизода, а на каждом шаге, на основании оценок, полученных на предыдущем шаге. Данный подход называется бутстрэппинг:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = R_{t+1} + \gamma G_{t+1}$$

$$v_\pi(s) = E_\pi[G_t \mid S_t = s] = \tag{6}$$

$$= E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] = R_{t+1} + \gamma v_\pi(S_{t+1});$$

2) Значения функции полезности обновляются на каждом шаге эпизода (обучение онлайн) с использованием правила временных различий (7), где ошибка определяется формулой (8):

$$V(S_t) = V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)] \quad (7)$$

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \quad (8)$$

3) В большинстве случаев сходится быстрее, чем алгоритм Монте-Карло.

Метод временных различий используется в задачах прогнозирования для оценки функции полезности, в частности, может быть использован для оценки стратегии лечения пациентов. Например, оценка стратегии лечения сепсиса с помощью метода временных различий продемонстрирована в исследовании [7].

Алгоритм метода временных различий.

Шаг 1. Входные данные: оцениваемая стратегия π.

Шаг 2. Инициализация входного параметра $\alpha \in (0,1]$ (размер шага).

Шаг 3. Инициализация $V(s)$ $\forall s \in S$ случайными значениями, $V(terminal) = 0$.

Шаг 4.

Для каждого эпизода:

Выбрать состояние S;

Для каждого шага эпизода:

Выбрать действие A из состояния S, согласно стратегии π;

A — действие, выбранное путем следования стратегии π для S;

Выполнить A, получить R, S';

$V(S) \leftarrow V(S) + \alpha[R + \gamma V(S') - V(S)]$;

$S \leftarrow S'$;

Повторять, пока S не является терминальным состоянием.

Метод временных различий определяет базовый алгоритмический механизм для методов SARSA, Q-Learning и Expected SARSA. Следовательно, задачи, решаемые данными алгоритмами, в целом являются общими, однако существуют особенности поведения данных алгоритмов.

2.2.3. SARSA

Данный алгоритм [11], совмещает подход *итерации по стратегиям* с правилом обновления метода временных различий. Таким образом, данный алгоритм позволяет решить задачу управления, путем последовательных итераций между оценкой функции полезности согласно алгоритму временных различий и улучшением стратегии на каждой итерации (путем вычисления жадной стратегии).

Правило обновления функции полезности SARSA:

$$Q(S, A) \leftarrow Q(S, A) + \\ +\alpha[R + \gamma Q(S', A') - Q(S, A)]. \quad (9)$$

Достоинством метода SARSA является то, что стратегия, вычисленная с использованием данного алгоритма, является наиболее безопасной и связана с наименьшим риском негативного исхода. Однако недостатком данного подхода является невысокая скорость обучения.

Таким образом, в контексте задач медицинской направленности стратегия обучения с подкреплением SARSA связана с наименьшим риском осложнений и неудачных исходов, при этом является наиболее длительной и, в то же время, безопасной.

Эффективность данного алгоритма, в частности, продемонстрирована, в исследовании по управлению уровнем глюкозы в крови [12]. Также с помощью алгоритма SARSA было проведено исследование [13] индивидуальных подходов лечения хронических заболеваний с помощью лекарственных препаратов.

2.2.4. Q-Learning.

Алгоритм Q-Learning [14] в настоящее время является наиболее популярным алгоритмом обучения с подкреплением в целом и, в частности, при решении задач управления в медицине. Как правило, стратегия обучения с подкреплением Q-Learning связана с оптимальным и максимально эффективным по длительности лечением, однако недостатком данного подхода является высокий риск возможных осложнений (критичных терминальных состояний).

С учетом известных достоинств и недостатков данного алгоритма, Q-Learning широко применяется при решении задач поиска оптимальных стратегий лечения [15], в частности, использовался в задаче определения оптимальной дозы инъекции инсулина [16].

Также данный метод доказал свою применимость в сфере персонализированной медицины [17].

Алгоритм Q-Learning

Шаг 1. Инициализация входных параметров $\alpha \in (0,1]$, $\varepsilon > 0$;

Шаг 2. Инициализация $Q(s,a) \in \mathbb{R}$ $\forall s \in S$, $a \in A(s)$ случайными значениями, $Q(terminal,\cdot) = 0$;

Шаг 3. Для каждого эпизода:

Выбрать состояние S;

Для каждого шага эпизода:

Выбрать действие A из состояния S, согласно стратегии Q (например, ε-жадной);

Выполнить A, получить R, S';

$$Q(S,A) \leftarrow Q(S,A) +$$

$$+\alpha[R + \gamma \max_a Q(S',A') - Q(S,A)];$$

$$S \leftarrow S'.$$

Повторять пока S не является терминальным состоянием.

2.2.5. Expected SARSA

Алгоритм Expected SARSA во многом идентичен Q-learning, за исключением того, что внутри правила обновления Q-функции вместо ее максимума используется математическое ожидание. Данная модификация приводит к тому, что целевая функция обновлений алгоритма Expected Sarsa является более устойчивой, чем обновления Sarsa.

$$\begin{gathered} Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \\ +\alpha[R_{t+1} + \gamma E_\pi[Q(S_{t+1},A_{t+1}) \mid S_{t+1}] - Q(S_t,A_t)] \leftarrow \\ \leftarrow Q(S_t,A_t) + \\ +\alpha[R_{t+1} + \gamma \sum_a \pi(a \mid S_{t+1}) Q(S_{t+1},a) - Q(S_t,A_t)]. \end{gathered} \quad (10)$$

Приведенные алгоритмы являются наиболее часто используемыми при решении задач медицинской практики. Однако важным вопросом является проблема выбора оптимального алгоритма в контексте конкретных задач.

2.3. Сравнение и оценка производительности рассматриваемых алгоритмов

Методы SARSA, Q-Learning и Expected SARSA имеют аналогичный контекст применения.

Алгоритм Q-Learning является одним из наиболее популярных и повсеместно используемых методов обучения с подкреплением, что, в основном, связано с зачастую лучшей производительностью данного алгоритма относительно методов SARSA/Expected SARSA. Однако существуют сценарии, в которых Q-Learning уступает SARSA и Expected SARSA, при этом результаты сравнительного анализа SARSA, Q-Learning и Expected SARSA зачастую демонстрируют лучшую производительность Expected SARSA среди ряда алгоритмов, основанных на методе временных различий [18].

Примером задачи, где применение алгоритма Expected SARSA является предпочтительным, является задача лечения заболевания, связанного с риском возможных осложнений. На рис. 3 продемонстрировано поведение алгоритмов Q-Learning, SARSA и Expected SARSA в условиях данной задачи.

Expected SARSA представляет собой наиболее оптимальную и сбалансированную относительно длительности лечения и рисками осложнений стратегию. Следовательно, в слу-

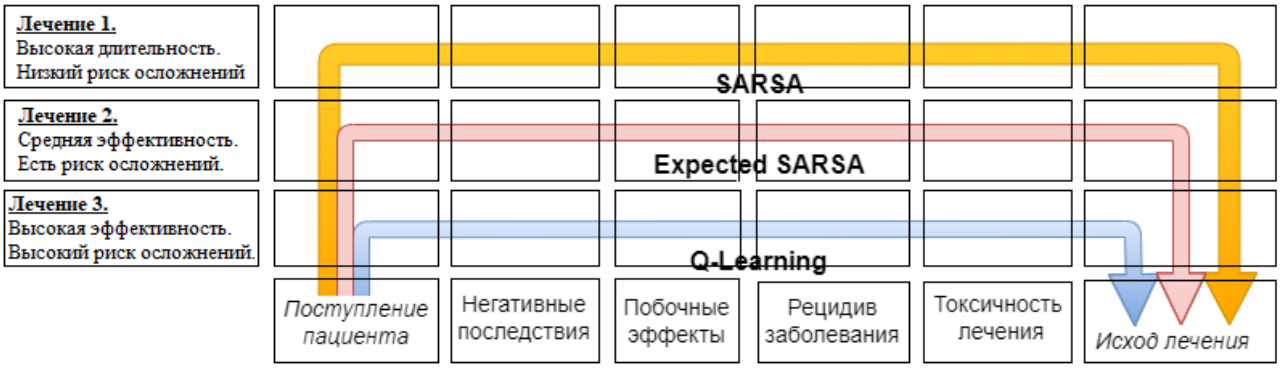

Рис. 3. Стратегии SARSA, Expected SARSA и Q-Learning
[Fig. 3. SARSA, Expected SARSA, Q-Learning]

чае, если оптимальные действия стратегии совмещены с риском осложнений и возможными негативными последствиями, алгоритм Expected SARSA в целом обеспечивает более стабильное поведение и успешно реализует безопасную стратегию.

Одной из широких областей применения обучения с подкреплением является задача разработки оптимальных стратегий лечения пациентов.

Общая схема принятия решений о выборе оптимального алгоритма обучения с подкреплением в решении задач назначения оптимальных схем лечения проиллюстрирована на рис. 4.

2.4. Валидация алгоритмов обучения с подкреплением

Значимым этапом решения задачи разработки стратегий лечения является оценка качества полученной модели. Рассмотренные ранее методы решения задач управления позволяют получить оценку качества построенной модели обучения с подкреплением. В общем случае, оценка эффективности разработанной модели может быть получена путем ее внедрения в реальное окружение и оценки результатов, полученных в ходе ее работы.

Однако контекст задач медицинской практики подразумевает отсутствие возможности интегрирования разработанной модели в реальную среду и тестирование разработанных стратегий лечения в клинических условиях.

Таким образом, возникает необходимость предварительной валидации модели обучения с подкреплением без ее интеграции в реальную среду.

Одним из способов решения данной проблемы является применение методов с разделенной оценкой ценности стратегий (off-policy), где целевая стратегия отличается от стратегии поведения, которая используется для выбора действий.

Например, в качестве стратегии поведения b можно выбрать реальную стратегию, которая используется врачом, а в качестве целевой стратегии π-оптимальную стратегию, теоретически рассчитанную одним из методов обучения с подкреплением. Изучаемая стратегия и стратегия, которую использует агент при выборе действий, являются стратегиями цели и поведения, соответственно.

Таким образом, возможно оценить качество разработанной алгоритмом обучения с подкреплением теоретической стратегии π на реальных медицинских данных, полученных в результате накопленного практическо-

Рис. 4. Схема выбора алгоритма обучения с подкреплением при решении задач назначения оптимальных схем лечения
[Fig. 4. Reinforcement learning method selection scheme for the task of searching for optimal treatment strategies]

го медицинского опыта (практически используемой стратегии b), с помощью метода выборки по значимости (11)–(12).

$$V^{\pi} = \frac{1}{N}\sum_{n=1}^{N} w^{H_n} R^{H_n}; \qquad (11)$$

$$w^{H_n} = \prod_{t=0}^{T_{H_n}} \frac{\pi(a_t^{H_n} \mid s_t^{H_n})}{b(a_t^{H_n} \mid s_t^{H_n})}; \qquad (12)$$

где H_n — история лечения пациента в течение госпитализации (эпизода) n, N — число эпизодов лечения.

3. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

С учетом рассмотренных особенностей решения задачи разработки стратегий лечения пациентов с помощью методов машинного обучения, была решена задача поиска оптимальных схем лечения (в условиях медицинского стационара) такого заболевания, как атеросклероз.

Основными этапами данного исследования являлись:

1. Разработка модели марковского процесса на основании набора исходных данных, представленного в виде множества эпизодов, состояний и действий.

2. Использование алгоритма итераций по значениям для поиска и оценки оптимальной стратегии лечения на основании построенной модели марковского процесса.

3. Оценка фактической стратегии лечения (стратегии лечения, применяемой на практике в госпитале) с использованием метода временных различий.

4. Оценка обобщающей способности модели на основании тестовой выборки с использованием алгоритма выборки по значимости.

Учитывая, что методы обучения с подкреплением, как правило, требуют большого объема входных данных, выборка стационарных пациентов, у которых был выявлен данный диагноз, была сформирована из базы данных MIMIC-III [19].

Исходный набор данных представлял собой массив измерений различных показателей пациентов – систолического, диастолического артериального давления, венозного давления и др.

С помощью проведения кластерного анализа на исходном наборе измерений пациентов был построен набор из 19 кластеров, представляющих собой основные состояния модели обучения с подкреплением. В [20] приведено описание решения задачи кластеризации пациентов с атеросклерозом для дальнейшего применения в модели обучения с подкреплением.

Помимо основных состояний, были добавлены 4 терминальных (конечные) состояния, соответствующих результатам госпитализации пациента: выписка (кластер № 20), выписка с назначением специализированного ухода (кластер № 21), назначение специализированного медицинского ухода (кластер № 22), летальный исход (кластер № 23).

Изменение состояния пациента подразумевает переход из одного состояния (кластера) в другой и сопровождается назначением вознаграждения. При переходе между основными состояниями агент получает отрицательное вознаграждение от –1 до 0, пропорционально степени тяжести состояния здоровья пациента. При переходе в одно из терминальных состояний агент получает вознаграждения от –2 до 2.

На основании набора обучающих данных эпизодов лечения была сформирована модель марковского процесса, описываемая следующими компонентами.

1. *Эпизод* лечения представляет собой историю госпитализации пациента, в течение периода которой фиксируются измерения показателей пациентов и производится назначение лечения (лекарственных препаратов). Исходный набор содержит 1470 эпизодов лечения, среди которых была выделена обучающая часть ~67 % (945 эпизодов) для тренировки алгоритма, тестовая часть ~ 33 % (466 эпизодов) для валидации алгоритма.

2. *Состояния*: $S: \ |S| = N$, где $N = 23$ (19 нетерминальных, 4 терминальных). Основные (нетерминальные) состояния соответствуют кластерам состояний здоровья пациентов, терминальные состояния соответствуют исходу лечения (выписке, выписке со специа-

лизированным домашним лечением, перевод пациента в другое отделение с целью специализированного ухода, летальный исход).

3. *Действия*: $A(S)$ — набор назначаемых пациентам медицинских предписаний в состоянии S, являющихся множеством наиболее частых комбинаций препаратов, применяемых в терапии атеросклероза.

4. *Вознаграждения*: $-1 \le R \le 0$ в нетерминальных состояниях (соответственно степени тяжести состояния), $R = 2$ в случае выписки пациента, $R = -2$ в случае летального исхода, $R = 0.5$ в случае выписки со специализированным лечением, $R = 0$ в случае перевода в отделение специализированного ухода.

5. *Вероятности перехода*: $p(s', r \mid s, a)$ — вероятность перехода из состояния s в состояние s' с вознаграждением r при условии применения действия a рассчитывается как доля переходов из состояния s в состояние s' при воздействии a относительно суммарного числа переходов из состояния s при воздействии a.

Следуя диаграмме на рис. 4, в качестве оптимального метода для решения данной задачи, выбран метод *динамического программирования* — алгоритм итерации по значениям.

Результатом выполнения алгоритма являлось получение оптимальной стратегии лечения пациентов, а также оценка стратегии для каждого из возможных состояний.

Фактическая стратегия лечения пациентов, используемая в медицинской практике, оценивалась с использованием алгоритма временных различий. Сравнение оценок оптимальной теоретической и фактической стратегий приведено на рис. 5.

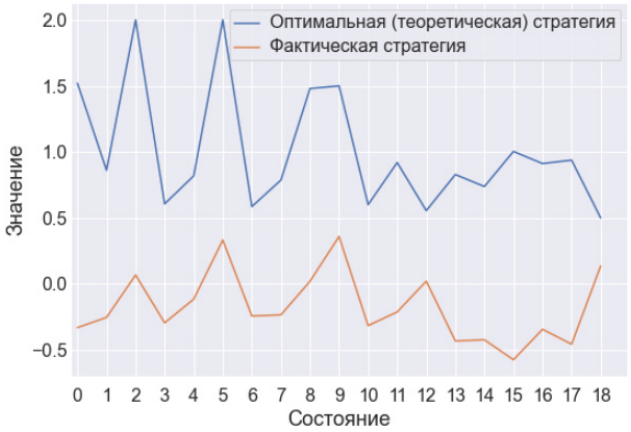

Рис. 5. Функции значимости теоретической и фактической стратегий
[Fig. 5. Value functions for theoretic and clinician strategies]

Из данного графика можно сделать вывод, что разработанная теоретическая стратегия превосходит фактическую, т. к. в каждом из состояний значение функции значимости теоретической стратегии превосходит значение для фактической стратегии.

Для валидации теоретической стратегии на тестовом наборе данных был применен метод выборки по значимости. При этом взвешенная оценка выборки по значимости была проведена с использованием различных значений параметра дисконта, что отражено в табл. 1.

Таблица 1. Взвешенная оценка выборки по значимости
[Table 1. Weighted importance sampling]

Параметр дисконта	Взвешенная оценка выборки по значимости	Среднее накопленное вознаграждение фактической стратегии
1	0.059	–0.17
0.9	0.059	–0.07
0.8	0.423	–0.04
0.7	0.438	–0.02
0.6	0.442	–0.01
0.5	0.448	0
0.4	0.448	0.009
0.3	0.6	0.01
0.2	0.6	0.02

Из табл. 1 следует, что при различных значениях параметра дисконта взвешенная оценка выборки по значимости превосходит среднее накопленное вознаграждение фактической стратегии и принимает положительные значения, что свидетельствует о высокой эффективности теоретической оптимальной стратегии лечения на тестовом наборе данных, в сравнении с фактически применяемой стратегией.

ЗАКЛЮЧЕНИЕ

В ходе данного исследования были рассмотрены основные теоретические аспекты такой области машинного обучения, как обучение с подкреплением.

Описанные подходы проиллюстрированы на примере задачи выбора стратегий лечения пациентов с атеросклерозом. Приведены результаты решения задачи поиска оптимальной стратегии лечения пациента с атеросклерозом на основе набора данных MIMIC-III.

БЛАГОДАРНОСТИ

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-37-90029 Аспиранты.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients / J. D. Martín-Guerrero, F. Gomez, E. Soria-Olivas, J. Schmidhuber, M. Climente-Martí, N. V. Jiménez-Torres // Expert Systems with Applications. – 2009. – Vol. 36, № 6. – P. 9737–9742. DOI: https://doi.org/10.1016/j.eswa.2009.02.041

2. Supervised Reinforcement Learning with Recurrent Neural Network for Dynamic Treatment Recommendation / L. Wang, W. Zhang, X. He, H. Zha // arXiv:1807.01473: электронный ресурс. – 2018. URL: https://arxiv.org/abs/1807.01473 (дата обращения 13.01.2022). DOI: https://doi.org/10.48550/arXiv.1807.01473

3. Learning the Dynamic Treatment Regimes from Medical Registry Data through Deep Q-network / N. Liu, Y. Liu, B. Logan, Z. Xu, J. Tang, Y. Wang // Scientific Reports. – 2019. – Vol. 9, № 1. DOI: https://doi.org/10.1038/s41598-018-37142-0

4. *Istepanian, R.S.H.* m-Health 2.0: New perspectives on mobile health, machine learning and big data analytics / R.H.S. Istepanian, T. Al-Anzi // Methods. – 2018. – № 151. – P. 34–40. DOI: 10.1016/j.ymeth.2018.05.015

5. *Саттон, Р.* Обучение с подкреплением: Введение / Р. Саттон, Р., Э. Д. Барто; пер. с англ. А. А. Слинкина. – М : ДМК Пресс, 2020. – 552 с.

6. Treating epilepsy via adaptive neurostimulation: a reinforcement learning approach / J. Pineau, A. Guez, R. Vincent, G. Panuccio, M. Avoli // International Journal of Neural Systems. – 2009. – Vol. 19, № 04. – P. 227–240. DOI: 10.1142/S0129065709001987

7. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care / M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, A. A. Faisal // Nature Medicine. – 2018. – Vol. 24, № 11. – P. 1716–1720. DOI: 10.1038/s41591-018-0213-5

8. *Utomo, C. P.* Treatment Recommendation in Critical Care: A Scalable and Interpretable Approach in Partially Observable Health States / C. P. Utomo, X. Li, W. Chen // ICIS 2018 Proceedings: электронный ресурс. – 2018. URL: https://aisel.aisnet.org/icis2018/healthcare/Presentations/9/ (дата обращения 13.01.2022).

9. *Jaimes, L. G.* CALMA, an algorithm framework for mobile just in time interventions / L. G. Jaimes, M. Llofriu, A. Raij // SoutheastCon 2015. DOI: 10.1109/SECON.2015.7133041

10. *Sutton, R. S.* Learning to predict by the methods of temporal differences / R. S. Sutton // Machine Learning. – 1988. – Vol. 3, № 1. – P. 9–44. DOI: https://doi.org/10.1007/BF00115009

11. *Rummery, G.* On-Line Q-Learning Using Connectionist Systems / G. Rummery, M. Niranjan // Technical Report CUED/F-INFENG/TR 166. – 1994.

12. *Noori, A.* Glucose level control using Temporal Difference methods / A. Noori, M. A. Sadrnia, M. Sistani, M. bagher N. // IEEE Xplore: электронный ресурс. – 2017. URL: https://ieeexplore.ieee.org/document/7985166 (дата обращения 13.01.2022). DOI: 10.1109/IranianCEE.2017.7985166

13. Reinforcement learning approach to individualization of chronic pharmacotherapy / A. E. Gaweda, M. K. Muezzinoglu, G. R. Aronoff, A. A. Jacobs, J. M. Zurada, M. E. Brier // Proceedings. 2005 IEEE International Joint Conference on Neural Networks. – 2005. DOI:10.1109/IJCNN.2005.1556455

14. *Watkins, C.J.C.H.* Learning from delayed rewards. – 1989. – P.234.

15. *Baniya, A.* Adaptive Interventions Treatment Modelling and Regimen Optimization Using Sequential Multiple Assignment Randomized Trials (Smart) and Q-Learning. – 2018. – P. 107.

16. Control of Blood Glucose for Type-1 Diabetes by Using Reinforcement Learning with Feedforward Algorithm / P. D. Ngo, S. Wei, A. Holubová, J. Muzik, F. Godtliebsen// Computational and Mathematical Methods in Medicine. – 2018. – P. 1–8. DOI: https://doi.org/10.1155/2018/4091497

17. Tools for the Precision Medicine Era: How to Develop Highly Personalized Treatment Recommendations From Cohort and Registry Data Using Q-Learning / E. F. Krakow, M. Hemmer, T. Wang, B. Logan, M. Arora, S. Spellman, D. Couriel, A. Alousi, J. Pidala, M. Last, S. Lachance, E.E.M. Moodie // American Journal of Epidemiology. – 2017. – Vol. 186, № 2. – P. 160–172. DOI: 10.1093/aje/kwy215

18. *Seijen, H. A.* Theoretical and Empirical Analysis of Expected Sarsa / H. van Seijen, H. van Hasselt, S. Whiteson, M. Wiering // Proceedings of the IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning: ADPRL: электронный ресурс. – 2009. URL: https://research.rug.nl/en/publications/a-theoretical-and-empirical-analysis-of-expected-sarsa (дата обращения 13.01.2022). DOI: 10.1109/ADPRL.2009.4927542

19. *Johnson, A.* OPEN SUBJECT CATEGORIES Background & Summary / A. Johnson, T. Pollard, L. Shen, L. W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, R. Mark: электронный ресурс. – 2016. URL: https://lcp.mit.edu/pdf/JohnsonSD2016.pdf.

20. *Демченко, М. В.* Кластеризация состояний пациентов для модели назначения схем лечения атеросклероза / М. В. Демченко, И. Л. Каширина, М. А. Фирюлина // Вестник ВГУ. Серия: Системный анализ и информационные технологии. – 2021. – № 2. – С. 126–137. DOI: https://doi.org/10.17308/sait.2021.2/3509

Демченко Мария Владиславовна — аспирант факультета ПММ Воронежского государственного университета.
E-mail: masha-vrn@yandex.ru
ORCID: https://orcid.org/0000-0002-6439-8957

Каширина Ирина Леонидовна — д-р техн. наук, профессор кафедры математических методов исследования операций факультета ПММ Воронежского государственного университета.
E-mail: kash.irina@mail.ru
ORCID: https://orcid.org/0000-0002-8664-9817

Фирюлина Мария Андреевна — аспирант факультета ПММ Воронежского государственного университета.
E-mail: mashafiryulina@mail.ru
ORCID: https://orcid.org/0000-0003-3468-5514

DOI: https://doi.org/
Received 15.01.2022
Accepted 22.04.2022

ISSN 1995-5499

THE USE OF REINFORCEMENT LEARNING METHODS IN MEDICAL PROBLEMS

© 2022 M. V. Demchenko, I. L. Kashirina✉, M. A. Firyulina

Voronezh State University
1, Universitetskaya Square, 394018 Voronezh, Russian Federation

Annotation. In this article the features of the modern reinforcement learning methods development for the medical tasks are discussed. Reinforcement learning methods are a popular machine learning tool used in the problems of finding optimal patient treatment strategies, personalized medicine, as well as interactive patient monitoring systems. One of the important task is to choose the optimal reinforcement learning algorithm from a variety of currently existing methods that have their own application specifics, advantages and disadvantages. This article is devoted to the analysis of the algorithmic apparatus of the most popular reinforcement learning methods and contains examples of models and results of the methods under consideration in the context of the problem of finding optimal treatment regimens for cardiac patients.
Keywords: Reinforcement learning, Markov decision process, dynamic programming, Bellman's equation, iteration over strategies, iteration over values, Monte Carlo, time difference method, SARSA, Q-Learning.

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Martín-Guerrero J. D., Gomez F., Soria-Olivas E., Schmidhuber J., Climente-Martí M. and Jiménez-Torres N. V.* (2009). A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients. *Expert Systems with Applications*. 36(6). P. 9737–9742.

2. *Wang L., Zhang W., He X. and Zha H.* (2018). Supervised Reinforcement Learning with Recurrent Neural Network for Dynamic Treatment Recommendation. arXiv:1807.01473. [online] Available at: https://arxiv.org/abs/1807.01473 (accessed 13 Jan. 2022).

3. *Liu N., Liu Y., Logan B., Xu Z., Tang J. and Wang Y.* (2019). Learning the Dynamic Treatment Regimes from Medical Registry Data through Deep Q-network. *Scientific Reports*. 9(1).

4. *Istepanian R.S.H. and Al-Anzi T.* (2018). m-Health 2.0: New perspectives on mobile health, machine learning and big data analytics. *Methods*. 151. P. 34–40.

5. *Sutton R. and Barto A. G.* (2020). Reinforcement learning: introduction. 552 p. (In Russian)

6. *Pineau J., Guez A., Vincent R., Panuccio G. and Avoli M.* (2009). Treating epilepsy via adaptive neurostimulation: a reinforcement learning approach. *International Journal of Neural Systems*. 19(04). P. 227–240.

7. *Komorowski M., Celi L. A., Badawi O., Gordon A. C. and Faisal A. A.* (2018). The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, [online] 24(11). P. 1716–1720. Available at: https://www.nature.com/articles/s41591-018-0213-5/.

8. *Utomo C. P., Li X. and Chen W.* (2018). Treatment Recommendation in Critical Care: A Scalable and Interpretable Approach in Partially Observable Health States. *ICIS 2018 Proceedings*. [online] Available at: https://aisel.aisnet.org/icis2018/healthcare/Presentations/9/ (accessed 13 Jan. 2022).

✉ Kashirina Irina L.
e-mail: kash.irina@mail.ru

9. *Jaimes L. G., Llofriu M. and Raij A.* (2015). CALMA, an algorithm framework for mobile just in time interventions. *SoutheastCon 2015*.

10. *Sutton R. S.* (1988). Learning to predict by the methods of temporal differences. *Machine Learning*. 3(1). P. 9–44.

11. *Rummery G. and Niranjan Mahesan.* (1994). On-Line Q-Learning Using Connectionist Systems. Technical Report CUED/F-INFENG/TR 166.

12. *Noori A., Sadrnia M. A., Sistani M. and Bagher N.* (2017). Glucose level control using Temporal Difference methods. [online] *IEEE Xplore*. Available at: https://ieeexplore.ieee.org/document/7985166 (accessed 13 Jan. 2022).

13. *Gaweda A. E., Muezzinoglu M. K., Aronoff G. R., Jacobs A. A., Zurada J. M. and Brier M. E.* (n.d.). Reinforcement learning approach to individualization of chronic pharmacotherapy. Proceedings. *2005 IEEE International Joint Conference on Neural Networks*, 2005.

14. *Watkins C.J.C.H.* (1989). Learning from delayed rewards. P. 234.

15. *Baniya A.* (2018). Adaptive Interventions Treatment Modelling and Regimen Optimization Using Sequential Multiple Assignment Randomized Trials (Smart) and Q-Learning. p.107.

16. *Ngo P. D., Wei S., Holubová A., Muzik J. and Godtliebsen F.* (2018). Control of Blood Glucose for Type-1 Diabetes by Using Reinforcement Learning with Feedforward Algorithm. *Computational and Mathematical Methods in Medicine*. 2018. P. 1–8.

17. *Krakow E. F., Hemmer M., Wang T., Logan B., Arora M., Spellman S., Couriel D., Alousi A., Pidala J., Last M., Lachance S. and Moodie E.E.M.* (2017). Tools for the Precision Medicine Era: How to Develop Highly Personalized Treatment Recommendations From Cohort and Registry Data Using Q-Learning. *American Journal of Epidemiology*. 186(2). P. 160–172.

18. *Seijen H. van, Hasselt H. van, Whiteson S. and Wiering M.* (2009). A Theoretical and Empirical Analysis of Expected Sarsa. *Proceedings of the IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning: ADPRL*. [online] Available at: https://research.rug.nl/en/publications/a-theoretical-and-empirical-analysis-of-expected-sarsa (accessed 13 Jan. 2022).

19. *Johnson A., Pollard T., Shen L., Lehman L.-W., Feng M., Ghassemi M., Moody B., Szolovits P., Celi L. and Mark R.* (2016). OPEN SUBJECT CATEGORIES Background & Summary. [online] Available at: https://lcp.mit.edu/pdf/JohnsonSD2016.pdf.

20. *Demchenko M. V., Kashirina I. L. and Firyulina M. A.* (2021). Clustering of patients' states for the development of atherosclerosis treatment model. *Proceedings of VSU. Series: Systems analysis and information technologies*. (2). P. 126–137.

Demchenko Maria V. — post-graduate student at Applied Mathematics and Mechanics faculty, Voronezh State University.
E-mail: masha-vrn@yandex.ru
ORCID: https://orcid.org/0000-0002-6439-8957

Kashirina Irina L. — DSc in Technical Sciences, Professor of the Department of Mathematical Methods of Operations Research at Applied Mathematics and Mechanics faculty, Voronezh State University.
E-mail: kash.irina@mail.ru
ORCID: https://orcid.org/0000-0002-8664-9817

Firyulina Maria A. – post-graduate student at Applied Mathematics and Mechanics faculty, Voronezh State University.
E-mail: mashafiryulina@mail.ru
ORCID: https://orcid.org/0000-0003-3468-5514