НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2024 Том 24 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2024 Vol. 24 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2024-24-3-474-482
УДК 004.048

Метод формирования сегментов информационной последовательности с использованием функционала качества моделей обработки

Даниил Дмитриевич Тихонов[1]✉, Илья Сергеевич Лебедев[2]

[1,2] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация

[1] tikhonovdanil@gmail.com✉, https://orcid.org/0009-0008-0128-4144
[2] isl_box@mail.ru, https://orcid.org/0000-0001-6753-2181

Аннотация

Введение. Постоянно возникающая потребность увеличения эффективности решения задач классификации и предсказания поведения объектов наблюдения вызывает необходимость совершенствования методов обработки данных. В работе предлагается метод повышения показателей качества моделей машинного обучения в задачах регрессии и прогнозирования. **Метод.** Предложенная обработка информационных последовательностей предполагает применение сегментации входных данных. В результате разделения данных образуются сегменты с различными свойствами объектов наблюдения. Новизна метода заключается в разделении последовательности на сегменты с использованием функционала качества моделей обработки на подвыборках данных. Это позволяет применять лучшие по качественным показателям модели на разных сегментах данных. Полученные сегменты являются отдельными подвыборками, на которые назначаются лучшие по качественным показателям модели и алгоритмы машинного обучения. **Основные результаты.** Для оценки качества предлагаемого решения выполнен эксперимент с использованием модельных данных и множественной регрессии. Рассчитанные значения показателя качества Root Mean Squared Error (RMSE) для выбранных алгоритмов на экспериментальной выборке и при различном количестве сегментов продемонстрировали повышение качественных показателей отдельных алгоритмов при увеличении количества сегментов. Предлагаемый метод позволяет повысить показатели RMSE в среднем на 7 % за счет сегментации и назначения моделей, которые имеют наилучшие показатели в отдельных сегментах. **Обсуждение.** Результаты метода могут применяться дополнительно при разработке моделей и методов обработки данных. Представленное решение направлено на дальнейшее усовершенствование и расширение ансамблевых методов. Формирование многоуровневых модельных структур, осуществляющих обработку, анализ поступающих информационных потоков и назначение наиболее подходящей модели для решения текущей задачи, позволяет уменьшить сложность и ресурсоемкость классических ансамблевых методов. В результате уменьшено влияние проблемы переобучения, снижена зависимость результатов обработки от базовых моделей, повышена оперативность настройки базовых алгоритмов в случае трансформации свойств данных и улучшена интерпретируемость результатов.

Ключевые слова

информационная последовательность данных, многоуровневая модель обработки данных, сегментация данных, повышение показателей качества

Ссылка для цитирования: Тихонов Д.Д., Лебедев И.С. Метод формирования сегментов информационной последовательности с использованием функционала качества моделей обработки // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 3. С. 474–482. doi: 10.17586/2226-1494-2024-24-3-474-482

Method for generating information sequence segments using the quality functional of processing models

Daniil D. Tikhonov[1]✉, Ilya S. Lebedev[2]

[1,2] St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation

[1] tikhonovdanil@gmail.com✉, https://orcid.org/0009-0008-0128-4144
[2] isl_box@mail.ru, https://orcid.org/0000-0001-6753-2181

Abstract

The constantly emerging need to increase the efficiency of solving classification problems and predicting the behavior of objects under observation necessitates improving data processing methods. This article proposes a method for improving the quality indicators of machine learning models in regression and forecasting problems. The proposed processing of information sequences involves the use of input data segmentation. As a result of data division, segments with different properties of observation objects are formed. The novelty of the method lies in dividing the sequence into segments using the quality functional of processing models on data subsamples. This allows you to apply the best quality models on various data segments. The segments obtained in this way are separate subsamples to which the best quality models and machine learning algorithms are assigned. To assess the quality of the proposed solution, an experiment was performed using model data and multiple regression. The obtained values of the quality indicator RMSE for various algorithms on an experimental sample and with a different number of segments demonstrated an increase in the quality indicators of individual algorithms with an increase in the number of segments. The proposed method can improve RMSE performance by an average of 7 % by segmenting and assigning models that have the best performance in individual segments. The results obtained can be additionally used in the development of models and data processing methods. The proposed solution is aimed at further improving and expanding ensemble methods. The formation of multi-level model structures that process, analyze incoming information flows and assign the most suitable model for solving the current problem makes it possible to reduce the complexity and resource intensity of classical ensemble methods. The impact of the overfitting problem is reduced, the dependence of processing results on the basic models is reduced, the efficiency of setting up basic algorithms in the event of transformation of data properties is increased, and the interpretability of the results is improved.

Keywords

information sequence of data, multi-level data processing model, data segmentation, improving quality indicators

For citation: Tikhonov D.D., Lebedev I.S. Method for generating information sequence segments using the quality functional of processing models. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 3, pp. 474–482 (in Russian). doi: 10.17586/2226-1494-2024-24-3-474-482

Введение

Повышение качественных показателей моделей обработки данных при решении задач классификации, регрессии и предсказания поведения системы является одним из фундаментальных вопросов развития методов машинного обучения. Эффективность алгоритмов зависит не только от выбранных способов обработки, но и от качества самих данных. Наличие ошибок, шумовых составляющих, выбросов, появление избыточных и зависимых переменных в выборках приводят к снижению качественных показателей обработки при задачах прогнозирования, регрессии и классификации [1]. В итоге возникает задача формирования оптимальных выборок данных для обучения моделей и их последующего использования.

С другой стороны, не менее важным для достижения высоких качественных показателей является использование эффективной модели обработки данных. В современных исследованиях для ее построения применяются как базовые алгоритмы, так и различные нейросетевые и ансамблевые структуры. Достижение заданных показателей в этих методах в большой степени зависит от свойств выборки данных, таких как распределение, размерность, частота появления объектов наблюдения [2]. Кроме того, в зависимости от предметной области на модели накладываются разные ограничения, связанные с быстродействием и ресурсоемкостью, возможностью адаптации при возникновении трансформации свойств данных [3]. Различные модели могут быть оптимизированы под одни свойства данных, но терять свою адекватность при изменении входных параметров анализируемой последовательности. В связи с этим в настоящей работе рассматривается метод, использующий разделение входной последовательности данных и назначение на отдельные сегменты моделей обработки, имеющих лучшие качественные показатели для полученных при сегментации подвыборок данных.

Существующие подходы

Процессы оптимизации методов и моделей обработки данных происходят по двум основным направлениям. Во-первых, осуществляются процессы «повышения качества» обрабатываемых данных, а во-вторых, выполняется построение эффективной модели обработки [4].

К первому направлению относятся методы формирования пространства признаков. Среди них можно выделить подходы на основе кластеризации, поиска точек разладки временных рядов, обнаружения «дрейфа концепта» при трансформации свойств данных. В машинном обучении такие подходы используются для решения ряда задач разделения последовательностей, автоматической генерации дополнительных параметров

алгоритмов машинного обучения, выявления неявных информационных структур.

Основополагающие работы [5–7] определили ряд параметрических методов поиска точек, где изменяются свойства. Несмотря на довольно длительное время существования, такие методы продолжают развиваться. Например, в [8, 9] предложены байесовские подходы для регрессий по точкам изменения. Однако они имеют высокую вычислительную сложность и требуют большое количество итераций.

Для ускорения обработки данных, при большом размере пространства признаков, для сегментации в [10–12] предложены эволюционные алгоритмы. В [13] рассмотрен метод, использующий процедуру растущего окна и последовательного сравнения свойств сегментов. В [14] представлено решение, основанное на оценивании характеристик и свойств информационных последовательностей.

Сегментирование данных — часто используемый метод для последующего анализа. В работе [15] выделен ряд решений для методов разделения: настройка существующих традиционных алгоритмов на свойства объектов наблюдения, преобразование данных временных рядов в статические выборки для дальнейшей обработки алгоритмами машинного обучения, использование паттернов на основе формы, признаков и моделей, с последующей их обработкой алгоритмами.

Основной целью этих методов является создание сегментов информационной последовательности и временного ряда в целях уменьшения сложности обработки и анализа. Однако такие подходы имеют проблемы с масштабируемостью, а их производительность зависит от свойств объектов наблюдения внутри сегментов.

В случае относительно «простых» данных сегментирование часто становится одним из основных решений, направленных на оптимизацию информационных последовательностей. Разделение выборки на сегменты позволяет определить внутреннюю структуру данных для дальнейшего анализа и обработки, исследовать вероятные связи между объектами наблюдений [16, 17]. В настоящее время сегментирование становится важным инструментом для решения ряда практических задач поиска знаний, обнаружения сбоев и аномалий [18, 19]. В методах машинного обучения сегменты могут использоваться для формирования меток информации, содержащейся в немаркированных образцах, подвыборок данных, объединения схожих объектов наблюдения, выбора объектов [20, 21].

Второе направление связано с поиском наиболее эффективной модели обработки последовательности. В простейших случаях применяются базовые алгоритмы, например: наивный байесовский классификатор (NB), линейный дискриминант (LD), метод опорных векторов (SVM), деревья решений (DT). Достигаемые ими значения показателей качества обработки зависят от свойств обрабатываемых данных. Линейные модели более устойчивы к шуму, лучше работают на краткосрочных периодах, но в случае нелинейности данных непригодны для долгосрочных прогнозов [22]. SVM показывает плохие результаты при наличии выбросов и шумов [23]. LD чувствителен к распределению данных. NB использует не всегда корректное предположение о независимости признаков [24]. DT подвержены неконтролируемому росту в случае наличия большого количества вариантов [25, 26]. В целях преодоления обозначенных проблемных вопросов используется многомодельный подход, направленный на формирование ансамбля моделей и алгоритмов, сочетающий несколько методов машинного обучения. Применение ансамблей алгоритмов направлено на повышение качественных показателей обработки во многих задачах анализа данных. Оно основано на оценке результатов различных методов обработки, что дает возможность создать более точную модель, агрегирующую выходные результаты. Это позволяет улучшить результат предсказания и уменьшить зависимость модели от конкретного набора данных [21]. В настоящее время такому подходу уделяется большое внимание в научных работах по повышению качества обработки последовательности. Ансамблевые методы используют разнородные модели, различное представление данных, подпространств, подвыборок, аппроксимаций параметров, что дает возможность добиваться повышения качественных показателей полноты и точности обработки. Модели и алгоритмы, объединенные в группы, легко распараллеливаются, что позволяет их использовать в высокопроизводительных вычислениях. А парадигма их применения сочетает простые зарекомендовавшие себя модели с более сложными моделями глубоких нейронных сетей. Все эти методы в той или иной степени улучшают отдельные качественные показатели, однако основными их недостатками являются сложность формирования выборки для обучения. Кроме того, часто возникают ситуации, когда неправильно подобранные модели и способы агрегации их результатов ухудшают общий прогноз. А в случае трансформации свойств данных могут быть затруднены процессы обучения [27].

Таким образом, сегментация информационных последовательностей и развитие методов, учитывающих локальные свойства данных, являются актуальными проблемными вопросами для методов машинного обучения.

Постановка задачи

Формирование временных рядов и информационных последовательностей осуществляются в целях оценки, контроля состояния, режимов работы и характеристик системы. При возникновении различных воздействий на анализируемый объект в определенные моменты времени возможно резкое изменение значений отслеживаемых параметров. Обнаружение таких точек дает возможность выделить сегменты. Для решения этой задачи могут быть использованы различные методы кластеризации, сегментации и разделения выборки.

Результаты таких методов зависят от настроек, метрик расстояния, точности вычисления точек разладки в последовательностях. Изменение отдельных параметров приводит к разным результатам. Кроме того, возникают проблемы определения количества сегментов. Свойства объектов наблюдения в сегментах отличаются, что приводит к тому, что различные модели обработ-

ки могут иметь разные значения показателей качества на сегментах, а усредненные значения существенно отличаться в зависимости от применяемых методов разбиения и количества сегментов.

В связи с этим возможно использование показателя качества моделей обработки для выбора способа сегментирования и количества сегментов.

Формальную постановку задачи представим следующим образом.

Имеется информационная последовательность объектов наблюдения X. Определены модели обработки $\{a_1, \ldots, a_N\} \in A$ и методы сегментации данных $\{\mu_1, \ldots, \mu_L\} \in \mu$.

Целью является поиск метода μ^* и его характеристик разделения последовательности на сегменты $X^{\mu^*} = \{X^{\mu^*}_{1\mu^*}, \ldots, X^{\mu^*}_{m\mu^*}\}$, при котором функционал качества каждой модели обработки $a_i \in A$, назначенной на определенный сегмент, имеет лучшее значение $Q(a_i(x), X^{\mu^*}_{i\mu^*}) \to \max\limits_{a_i \in A,\ \mu^* \in \mu}$.

В результате возникает задача разработки метода формирования сегментов информационной последовательности. Метод должен отличаться от известных использованием функционала качества моделей обработки на подвыборках данных, что позволит сформировать агрегационную модель, осуществляющую назначение лучших по качественным показателям моделей на сегменты.

Предлагаемый метод

В качестве моделей могут выступать базовые алгоритмы обработки данных, например: линейная регрессия, деревья решений или машина опорных векторов. Методы сегментации определяются исходя из свойств последовательностей данных. Для решения регрессионных задач могут быть, например, алгоритмы кластеризации или поиска точек разладки. Выбор метода разбиения ограничивается вычислительной сложностью и ресурсоемкостью.

Последовательность объектов наблюдения сегментируется методами $\{\mu_1, \ldots, \mu_L\} \in \mu$. Все модели $\{a_1, \ldots, a_N\} \in A$, обучаются на всех сегментах. На каждый сегмент назначается модель a_i, которая имеет лучшие значения выбранного показателя качества. Для полученного разбиения выборки строится агрегированная модель обработки, состоящая из алгоритмов $\{a_1, \ldots, a_N\} \in A$, в которой алгоритм $a_i \in A$ выбирается и назначается на тот сегмент, который имеет лучшие значения показателя качества по сравнению с другими алгоритмами. В дальнейшем a_i обрабатывает только данные, принадлежащие этому сегменту.

Реализация метода предполагает выполнение следующих шагов.

Шаг 1. Формируется тренировочный датасет X, содержащий обучающую последовательность.

Шаг 2. Определяется L методы $\{\mu_1, \ldots, \mu_L\} \in \mu$ разбиения выборки X.

Шаг 3. Определяется N модели $\{a_1, \ldots, a_N\} \in A$ обработки данных выборки X.

Шаг 4. Задается функционал качества $Q(a(x), X)$.

Шаг 5. Определяется M максимальное количество сегментов.

Шаг 6. Выполняется цикл перебора методов разбиения выборки $l = 1, \ldots, L$.

Шаг 7. Выполняется цикл, увеличивающий количество сегментов на каждом шаге $m = 1, \ldots, M$.

Шаг 8. Выборка X обрабатывается методом разбиения μ_l.

Шаг 9. Формируются сегменты $\{X^{\mu_l}_{1\mu_l}, \ldots, X^{\mu_l}_{j\mu_l}, \ldots, X^{\mu_l}_{m\mu_l}\} \in X^{\mu_l}$ для текущего метода разбиения μ_l и количества сегментов m.

Шаг 10. Выполняется цикл перебора сегментов $j = 1, \ldots, m$.

Шаг 11. Выполняется цикл перебора моделей $i = 1, \ldots, N$.

Шаг 12. Выполняется обучение модели a_i на сегменте $X^{\mu_l}_{j\mu_l}$.

Шаг 13. Ожидание окончания цикла перебора моделей (если нет окончания цикла, то переход к шагу 11).

Шаг 14. На сегменте $X^{\mu_l}_{j\mu_l}$ определяется лучшая из моделей $\{a_1, \ldots, a_N\} \in A$ по значению показателя качества $a^{j}{}_{\mu_l} = \arg\max\limits_{a_i \in A} Q(a_i(x, X^{\mu_l}_{j\mu_l}))$.

Шаг 15. Ожидание окончания цикла перебора сегментов (если нет окончания цикла, то переход к шагу 10).

Шаг 16. Определяется выборка $X^{\mu_l}_m = \{X^{\mu_l}_{1\mu_l}, \ldots, X^{\mu_l}_{m\mu_l}\}$.

Шаг 17. Формируется модель из моделей, определенных на шаге 14 $\{a^1{}_{\mu_l}, \ldots, a^m{}_{\mu_l}\} \in A$, показывающая лучшие результаты по значению показателя качества, после обработки методом μ_l и содержащая m сегментов на выборке $X^{\mu_l}_m$

$$a^{\mu_l}_m(x, X^{\mu_l}_m) = \begin{cases} a^1{}_{\mu_l}(x, X^{\mu_l}_{1\mu_l}), x \in X^{\mu_l}_{1\mu_l} \\ \ldots \\ a^m{}_{\mu_l}(x, X^{\mu_l}_{m\mu_l} l), x \in X^{\mu_l}_{m\mu_l} \end{cases}.$$

Шаг 18. Ожидание окончания цикла увеличения количества сегментов (если нет окончания цикла, то переход к шагу 7).

Шаг 19. Определяется количество сегментов при разбиении методом μ_l, на котором был достигнут лучший показатель качества $m^{\mu_l} = \arg\max\limits_{m \in \{1,\ldots,M\}} Q(a^{\mu_l}_m(x, X^{\mu_l}_m))$.

Шаг 20. Определяется выборка $X^{\mu_l}_{m^{\mu_l}} = \{X^{\mu_l}_{1^{\mu_l}}, \ldots, X^{\mu_l}_{m^{\mu_l}}\}$.

Шаг 21. Определяется модель при разбиении методом μ_l, которая достигает лучшего показателя качества $a^{\mu_l}_{m^{\mu_l}} = (x, X^{\mu_l}_{m^{\mu_l}}) = \arg\max\limits_{a^{\mu_l}_m \in A} Q(a^{\mu_l}_m(x, X^{\mu_l}_m))$.

Шаг 22. Ожидание окончания цикла перебора методов разбиения выборки (если нет окончания цикла, то переход к шагу 6).

Шаг 23. Выбирается метод разбиения выборки, где достигается максимальный показатель качества $\mu^* = \arg\max\limits_{\mu_l \in \mu} Q(a^{\mu_l}_{m^{\mu_l}}(x, X^{\mu_l}_{m^{\mu_l}}))$.

Шаг 24. Определяется количество сегментов $m^{\mu^*} = \arg\max\limits_{m \in \{1,\ldots,M\}} Q(a^{\mu_l}_{m^{\mu_l}}(x, X^{\mu_l}_{m^{\mu_l}}))$.

Шаг. 25. Определяются сегменты выборки, обработанной выбранным методом разбиения $X^{\mu^*} = \{X_{1^{\mu^*}}^{\mu^*}, \ldots, X_{m^{\mu^*}}^{\mu^*}\}$.

Шаг. 26. Формируется модель обработки

$$a^{\mu^*}(x, X^{\mu^*}) = \begin{cases} a_{1^{\mu^*}}^{\mu^*}(x, X_{1^{\mu^*}}^{\mu^*}), x \in X_{1_{\mu^*}}^{\mu^*} \\ \ldots \\ a_{m^{\mu^*}}^{\mu^*}(x, X_{m^{\mu^*}}^{\mu^*}), x \in X_{m^{\mu^*}}^{\mu^*} \end{cases}.$$

Представленная алгоритмическая последовательность действий дает возможность определить лучший из заранее выбранных метод сегментации и количество сегментов, сформировать модель обработки, где на каждый сегмент назначается свой алгоритм, показывающий в процессе обучения лучший результат на данном сегменте.

Экспериментальное исследование метода

Цель проведения эксперимента состояла в оценке повышения качественных показателей обработки информационных последовательностей при применении рассматриваемого метода. Так как он использует сегментацию выборки данных, то были рассмотрены два основных подхода к разделению. При первом подходе последовательность обрабатывалась алгоритмом кластеризации k-ближайших соседей (KNN), границы сегмента определялись принадлежностью кластеру. Во втором — применен подход, разделяющий последовательность на равные по количеству наблюдений сегменты.

Значения показателя качества обработки данных определялись метрикой Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \quad (1)$$

где y_i — реальное значение; $\hat{y}_i$ — предсказанное значение; n — количество объектов наблюдения.

Для обработки данных выполнялось разделение на последовательности на m частей обоими подходами. На каждом сегменте выполнялось обучение алгоритма линейной регрессии. После этого для входной последовательности определялась принадлежность объекта наблюдения и значение $\hat{y}_i$ предсказывалось алгоритмом, назначенным на сегмент.

На рис. 1 показаны модельные данные эксперимента: зависимость для всей выборки; разбиение на три равных по количеству объекта наблюдения сегмента и сегменты, определенные алгоритмом KNN с тремя заданными кластерами.

На рис. 2 представлена диаграмма значений функции потерь (1) для всей выборки целиком (RMSE all) и при делении на три сегмента двумя подходами (RMSE seg, RMSE knn). Вычисленные значения выражения (1) показывают преимущество использования методов разделения последовательности данных. Причем разбиение на равные части для рассматриваемого случая оказалось предпочтительнее использования метода KNN.

Выполним увеличение m числа сегментов двумя выбранными подходами. На рис. 3 приведена зависимость значений RMSE линейной регрессии от количества сегментов, полученных при делении последовательности на равные части и при использовании KNN.

Из рис. 3 видно, что значения функции потерь уменьшаются при увеличении количества сегментов. Причем на представленных данных разделение сегментов на равные части при малых значениях числа сегментов m позволяет получить значения функции потерь лучше, чем методом KNN. В дальнейшем при уменьшении размера кластера выбор метода почти не влияет на результат.

Далее для проведения эксперимента был выбран набор данных, позволяющий решать задачи множественной регрессии. В качестве базовых алгоритмов определены линейная регрессия (LR), регрессия гауссова процесса (GR), машина опорных векторов (SVM), деревья решений (DT).

На рис. 4 представлены графики RMSE (m) выражения (1) для алгоритмов LR, GR, SVM, DT.

Графики рис. 4 показывают, что в ряде случаев, например для алгоритмов LR, SVM, GR уменьшение размера сегмента без учета свойств содержащейся в нем информации (направление тренда, разброс данных) может приводить к улучшению качественных показателей обработки.

Несмотря на полученные высокие результаты для различных датасетов, где применение предлагаемого метода позволяет повысить качественные показатели отдельных алгоритмов и уменьшить вычислительную сложность, для его использования необходим предварительный анализ данных, направленный на оценку репрезентативности, однородности и адекватности выборки. Отметим, что не для всех моделей предложенный метод может быть эффективен. Например, на алгоритм DT уменьшение размера кластера почти не оказывает влияния, а значение показателя RMSE для него остается на одном уровне.

Другая особенность предлагаемого метода состоит в возможности построения агрегационной функции применения различных алгоритмов на разных сегментах. В рамках эксперимента на выбранном датасете при разделении на 10 равных по размеру сегментов — на каждом сегменте выбранные алгоритмы достигают разных результатов. На рис. 5 приведены значения RMSE алгоритмов на каждом из 10 сегментов. При анализе гистограмм видно, что алгоритм GR демонстрирует лучшие значения, но на сегментах № 3 и № 7 лучшие значения показывает алгоритм LR, а на сегменте № 9 — SVM.

Заметим, что может быть целесообразно в случае определения свойств данных внутри сегментов решение задачи по назначению лучшего алгоритма на сегмент, где он показывает лучшие значения.

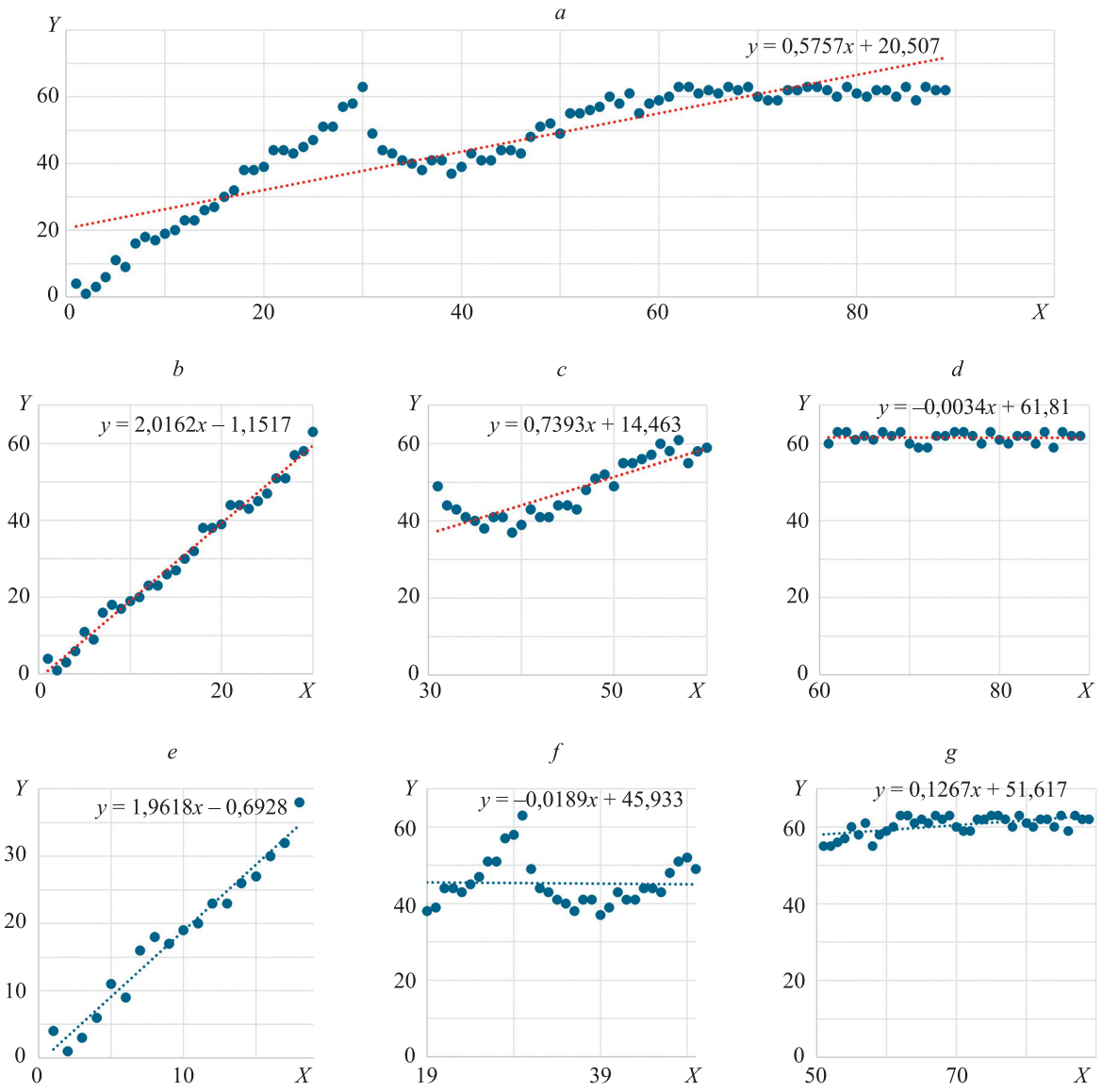

Рис. 1. Назначение моделей на сегменты: вся последовательность целиком (*a*); разделение последовательности на равные части (*b–d*) и методом KNN (*e–g*)

Fig. 1. Assignment of models to segments: the entire sequence (*a*); sequence division into equal parts (*b–d*) and KNN method (*e–g*)

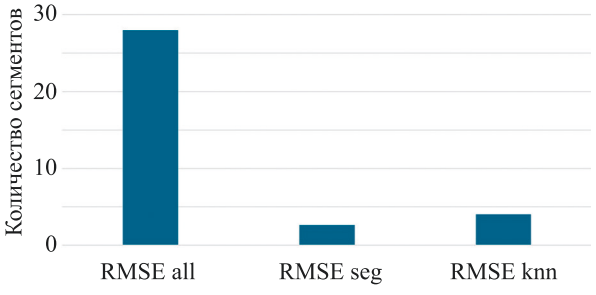

Рис. 2. Функция потерь RMSE на всей выборке (RMSE all) и при делении на три сегмента равным количеством объектов наблюдения (RMSE seg) и методом KNN (RMSE knn)

Fig. 2. Loss function RMSE on the entire sample (RMSE all) and when divided into three segments by an equal number of observation objects (RMSE seg) and the KNN method (RMSE knn)

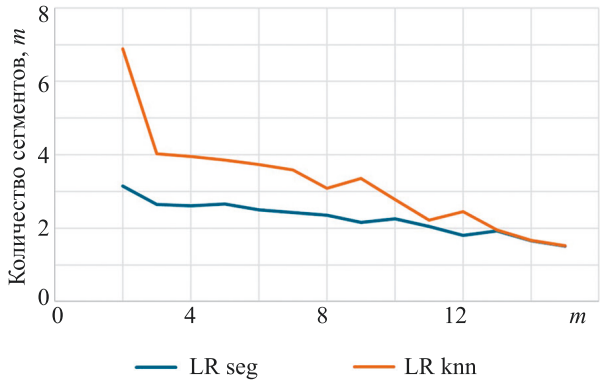

Рис. 3. Зависимость значений RMSE от количества сегментов *m* для линейной регрессии при разделении на равные части по количеству объектов (LR seg) и методом KNN (LR knn)

Fig. 3. Dependence of RMSE values vs. the number of segments *m* for linear regression when divided into equal parts by the number of objects (LR seg) and the KNN method (LR knn)

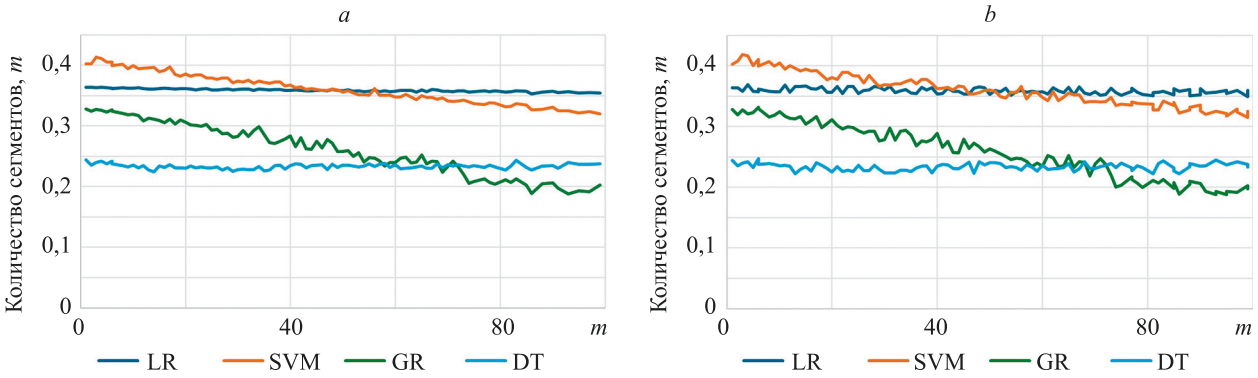

Рис. 4. Зависимость значений RMSE различных алгоритмов от количества сегментов *m* для множественной регрессии при сегментировании делением на равные отрезки (*a*) и методом KNN (*b*)

Fig. 4. RMSE values dependence of various algorithms vs. the *m* segments number for multiple regression when segmented by division into equal segments (*a*) and the KNN method (*b*)

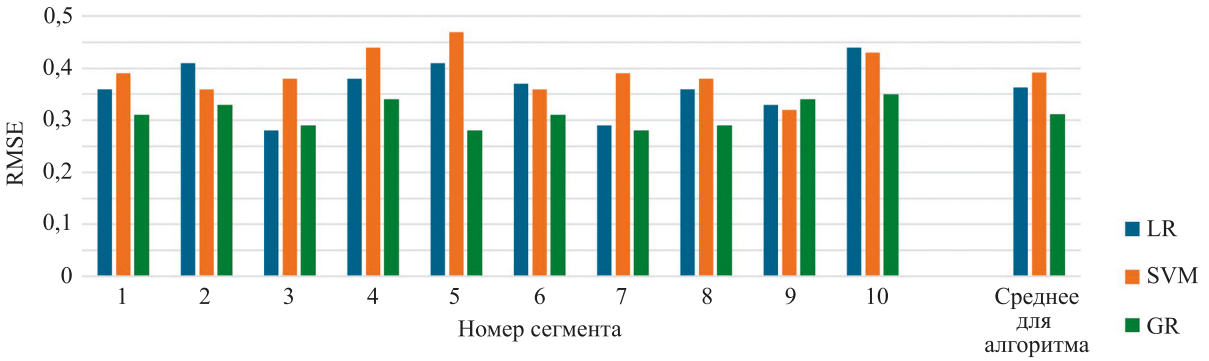

Рис. 5. Значения RMSE алгоритмов LR, SVM, GR для разных сегментов

Fig. 5. RMSE values of LR, SVM, GR algorithms for different segments

Заключение

Предложенный метод позволяет совершенствовать ансамблевые методы. Он направлен на улучшение показателей качества обработки информационных потоков и выборок данных при ограничении ресурсов. Новизна метода заключается в использовании функционала качества моделей обработки при сегментации информационных последовательностей, что позволяет формировать агрегационную модель, использующую назначение лучших по качественным показателям алгоритмов на сегменты. На каждом сегменте по отдельности происходит обучение, а затем выбирается и назначается алгоритм с лучшими качественными показателями для данного сегмента.

Применение метода позволяет использовать менее ресурсоемкие модели обработки данных, что дает возможность снизить вычислительные затраты на переобучение в случае изменения свойств данных.

Литература

1. Marques H.O., Swersky L., Sander J., Campello R., Zimek A. On the evaluation of outlier detection and one-class classification: a comparative study of algorithms, model selection, and ensembles // Data Mining and Knowledge Discovery. 2023. V. 37. N 4. P. 1473–1517. https://doi.org/10.1007/s10618-023-00931-x
2. Mishra S., Shaw K., Mishra D., Patil S., Kotecha K., Kumar S., Bajaj S. Improving the accuracy of ensemble machine learning classification models using a novel bit-fusion algorithm for healthcare AI systems // Frontiers in Public Health. 2022. V. 10. P. 1–17. https://doi.org/10.3389/fpubh.2022.858282
3. Ren J., Tapert S., Fan C.C., Thompson W.K. A semi-parametric Bayesian model for semi-continuous longitudinal data // Statistics in Medicine. 2022. V. 41. N 13. P. 2354–2374. https://doi.org/10.1002/sim.9359
4. Zhang Y., Liu J., Shen W. A review of ensemble learning algorithms used in remote sensing applications // Applied Sciences. 2022. V. 12. N 17. P. 8654. https://doi.org/10.3390/app12178654

References

1. Marques H.O., Swersky L., Sander J., Campello R., Zimek A. On the evaluation of outlier detection and one-class classification: a comparative study of algorithms, model selection, and ensembles. *Data Mining and Knowledge Discovery*, 2023, vol. 37, no. 4, pp. 1473–1517. https://doi.org/10.1007/s10618-023-00931-x
2. Mishra S., Shaw K., Mishra D., Patil S., Kotecha K., Kumar S., Bajaj S. Improving the accuracy of ensemble machine learning classification models using a novel bit-fusion algorithm for healthcare AI systems. *Frontiers in Public Health*, 2022, vol. 10, pp. 1–17. https://doi.org/10.3389/fpubh.2022.858282
3. Ren J., Tapert S., Fan C.C., Thompson W.K. A semi-parametric Bayesian model for semi-continuous longitudinal data. *Statistics in Medicine*, 2022, vol. 41, no. 13, pp. 2354–2374. https://doi.org/10.1002/sim.9359
4. Zhang Y., Liu J., Shen W. A review of ensemble learning algorithms used in remote sensing applications. *Applied Sciences*, 2022, vol. 12, no. 17, pp. 8654. https://doi.org/10.3390/app12178654

5. Bellman R. On the approximation of curves by line segments using dynamic programming // Communications of the ACM. 1961. V. 4. N 6. P. 284–301. https://doi.org/10.1145/366573.366611
6. Page E. A test for a change in a parameter occurring at an unknown point // Biometrika. 1955. V. 42. N 3/4. P. 523–527. https://doi.org/10.2307/2333401
7. Fisher W.D. On grouping for maximum homogeneity // Journal of the American Statistical Association. 1958. V. 53. N 284. P. 789–798. https://doi.org/10.1080/01621459.1958.10501479
8. Melnyk I., Banerjee A. A spectral algorithm for inference in hidden semi-Markov models // Journal of Machine Learning Research. 2017. V. 18. N 35. P. 1–39.
9. Bardwell L., Fearnhead P. Bayesian detection of abnormal segments in multiple time series // Bayesian Analysis. 2017. V. 12. N 1. P. 193–218. https://doi.org/10.1214/16-ba998
10. Chung F.-L., Fu T.-C., Ng V., Luk R.W.P. An evolutionary approach to pattern-based time series segmentation // IEEE Transactions on Evolutionary Computation. 2004. V. 8. N 5. P. 471–489. https://doi.org/10.1109/tevc.2004.832863
11. Levchenko O., Kolev B., Yagoubi D.E., Akbarinia R., Masseglia F., Palpanas T., Shasha D., Valduriez P. BestNeighbor: efficient evaluation of kNN queries on large time series databases // Knowledge and Information Systems. 2020. V. 63. N 2. P. 349–378. https://doi.org/10.1007/s10115-020-01518-4
12. Nikolaou A., Gutiérrez P.A., Durán A., Dicaire I., Fernández-Navarro F., Hervás-Martínez C. Detection of early warning signals in paleoclimate data using a genetic time series segmentation algorithm // Climate Dynamics. 2015. V. 44. N 7. P. 1919–1933. https://doi.org/10.1007/s00382-014-2405-0
13. Liu N., Zhao J. Streaming data classification based on hierarchical concept drift and online ensemble // IEEE Access. 2023. V. 11. P. 126040–126051. https://doi.org/10.1109/access.2023.3327637
14. Zhong G., Shu T., Huang G., Yan X. Multi-view spectral clustering by simultaneous consensus graph learning and discretization // Knowledge-Based Systems. 2022. V. 235. P. 107632. https://doi.org/10.1016/j.knosys.2021.107632
15. Liakos P., Papakonstantinopoulou K., Kotidis Y. Chimp: efficient lossless floating point compression for time series databases // Proceedings of the VLDB Endowment. 2022. V. 15. N 11. P. 3058–3070. https://doi.org/10.14778/3551793.3551852
16. Лебедев И.С. Сегментирование множества данных с учетом информации воздействующих факторов // Информационно-управляющие системы. 2021. № 3(112). С. 29–38. https://doi.org/10.31799/1684-8853-2021-3-29-38
17. Мальцев Г.Н., Якимов В.Л. Подход к формированию обобщенных параметров технического состояния сложных технических систем с использованием нейросетевых структур // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 4. С. 828–835. https://doi.org/10.17586/2226-1494-2023-23-4-828-835
18. Shili H. Clustering in big data analytics: a systematic review and comparative analysis (review article) // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 5. С. 967–979. https://doi.org/10.17586/2226-1494-2023-23-5-967-979
19. Lebedev I.S., Sukhoparov M.E. Adaptive Learning and integrated use of information flow forecasting methods // Emerging Science Journal. 2023. V. 7. N 3. P. 704–723. https://doi.org/10.28991/esj-2023-07-03-03
20. Tong W., Wang Y., Liu D. An adaptive clustering algorithm based on local-density peaks for imbalanced data without parameters // IEEE Transactions on Knowledge and Data Engineering. 2023. V. 35. N 4. P. 3419–3432. https://doi.org/10.1109/tkde.2021.3138962
21. Silva R.P., Zarpelão B.B., Cano A., Junior S.B. Time series segmentation based on stationarity analysis to improve new samples prediction // Sensors. 2021. V. 21. N 21. P. 7333. https://doi.org/10.3390/s21217333
22. Barzegar V., Laflamme S., Hu C., Dodson J. Multi-time resolution ensemble LSTMs for enhanced feature extraction in high-rate time series // Sensors. 2021. V. 21. N 6. P. 1954. https://doi.org/10.3390/s21061954
23. Huang W., Ding N. Privacy-preserving support vector machines with flexible deployment and error correction // Lecture Notes in Computer Science. 2021. V. 13107. P. 242–262. https://doi.org/10.1007/978-3-030-93206-0_15

5. Bellman R. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 1961, vol. 4, no. 6, pp. 284–301. https://doi.org/10.1145/366573.366611
6. Page E. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 1955, vol. 42, no. 3/4, pp. 523–527. https://doi.org/10.2307/2333401
7. Fisher W.D. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 1958, vol. 53, no. 284, pp. 789–798. https://doi.org/10.1080/01621459.1958.10501479
8. Melnyk I., Banerjee A. A spectral algorithm for inference in hidden semi-Markov models. *Journal of Machine Learning Research*, 2017, vol. 18, no. 35, pp. 1–39.
9. Bardwell L., Fearnhead P. Bayesian detection of abnormal segments in multiple time series. *Bayesian Analysis*, 2017, vol. 12, no. 1, pp. 193–218. https://doi.org/10.1214/16-ba998
10. Chung F.-L., Fu T.-C., Ng V., Luk R.W.P. An evolutionary approach to pattern-based time series segmentation. *IEEE Transactions on Evolutionary Computation*, 2004, vol. 8, no. 5, pp. 471–489. https://doi.org/10.1109/tevc.2004.832863
11. Levchenko O., Kolev B., Yagoubi D.E., Akbarinia R., Masseglia F., Palpanas T., Shasha D., Valduriez P. BestNeighbor: efficient evaluation of kNN queries on large time series databases. *Knowledge and Information Systems*, 2020, vol. 63, no. 2, pp. 349–378. https://doi.org/10.1007/s10115-020-01518-4
12. Nikolaou A., Gutiérrez P.A., Durán A., Dicaire I., Fernández-Navarro F., Hervás-Martínez C. Detection of early warning signals in paleoclimate data using a genetic time series segmentation algorithm. *Climate Dynamics*, 2015, vol. 44, no. 7, pp. 1919–1933. https://doi.org/10.1007/s00382-014-2405-0
13. Liu N., Zhao J. Streaming data classification based on hierarchical concept drift and online ensemble. *IEEE Access*, 2023, vol. 11, pp. 126040–126051. https://doi.org/10.1109/access.2023.3327637
14. Zhong G., Shu T., Huang G., Yan X. Multi-view spectral clustering by simultaneous consensus graph learning and discretization. *Knowledge-Based Systems*, 2022, vol. 235, pp. 107632. https://doi.org/10.1016/j.knosys.2021.107632
15. Liakos P., Papakonstantinopoulou K., Kotidis Y. Chimp: efficient lossless floating point compression for time series databases. *Proceedings of the VLDB Endowment*, 2022, vol. 15, no. 11, pp. 3058–3070. https://doi.org/10.14778/3551793.3551852
16. Lebedev I. Dataset segmentation considering the information about impact factors. *Information and Control Systems*, 2021, no. 3(112), pp. 29–38. (in Russian). https://doi.org/10.31799/1684-8853-2021-3-29-38
17. Maltsev G.N., Yakimov V.L. Approach to the generalized parameters formation of the complex technical systems technical condition using neural network structures. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 4, pp. 828–835. (in Russian). https://doi.org/10.17586/2226-1494-2023-23-4-828-835
18. Shili H. Clustering in big data analytics: a systematic review and comparative analysis (review article). *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 5, pp. 967–979. https://doi.org/10.17586/2226-1494-2023-23-5-967-979
19. Lebedev I.S., Sukhoparov M.E. Adaptive Learning and integrated use of information flow forecasting methods. *Emerging Science Journal*, 2023, vol. 7, no. 3, pp. 704–723. https://doi.org/10.28991/esj-2023-07-03-03
20. Tong W., Wang Y., Liu D. An adaptive clustering algorithm based on local-density peaks for imbalanced data without parameters. *IEEE Transactions on Knowledge and Data Engineering*, 2023, vol. 35, no. 4, pp. 3419–3432. https://doi.org/10.1109/tkde.2021.3138962
21. Silva R.P., Zarpelão B.B., Cano A., Junior S.B. Time series segmentation based on stationarity analysis to improve new samples prediction. *Sensors*, 2021, vol. 21, no. 21, pp. 7333. https://doi.org/10.3390/s21217333
22. Barzegar V., Laflamme S., Hu C., Dodson J. Multi-time resolution ensemble LSTMs for enhanced feature extraction in high-rate time series. *Sensors*, 2021, vol. 21, no. 6, pp. 1954. https://doi.org/10.3390/s21061954
23. Huang W., Ding N. Privacy-preserving support vector machines with flexible deployment and error correction. *Lecture Notes in Computer Science*, 2021, vol. 13107, pp. 242–262. https://doi.org/10.1007/978-3-030-93206-0_15

24. Zhang X., Wang M. Weighted random forest algorithm based on Bayesian algorithm // Journal of Physics: Conference Series. 2021. V. 1924. P. 012006. https://doi.org/10.1088/1742-6596/1924/1/012006
25. Di Franco G., Santurro M. Machine learning, artificial neural networks and social research // Quality & Quantity. 2021. V. 55. N 3. P. 1007–1025. https://doi.org/10.1007/s11135-020-01037-y
26. Si S., Zhao J., Cai Z., Dui H. Recent advances in system reliability optimization driven by importance measures // Frontiers of Engineering Management. 2020. V. 7. N 3. P. 335–358. https://doi.org/10.1007/s42524-020-0112-6
27. Xu S., Song Y., Hao X. A comparative study of shallow machine learning models and deep learning models for landslide susceptibility assessment based on imbalanced data // Forests. 2022. V. 13. N 11. P. 1908. https://doi.org/10.3390/f13111908

24. Zhang X., Wang M. Weighted random forest algorithm based on Bayesian algorithm. *Journal of Physics: Conference Series*, 2021, vol. 1924, pp. 012006. https://doi.org/10.1088/1742-6596/1924/1/012006
25. Di Franco G., Santurro M. Machine learning, artificial neural networks and social research. *Quality & Quantity*, 2021, vol. 55, no. 3, pp. 1007–1025. https://doi.org/10.1007/s11135-020-01037-y
26. Si S., Zhao J., Cai Z., Dui H. Recent advances in system reliability optimization driven by importance measures. *Frontiers of Engineering Management*, 2020, vol. 7, no. 3, pp. 335–358. https://doi.org/10.1007/s42524-020-0112-6
27. Xu S., Song Y., Hao X. A comparative study of shallow machine learning models and deep learning models for landslide susceptibility assessment based on imbalanced data. *Forests*, 2022, vol. 13, no. 11, pp. 1908. https://doi.org/10.3390/f13111908

Авторы

Тихонов Даниил Дмитриевич — аспирант, инженер-программист, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, https://orcid.org/0009-0008-0128-4144, tikhonovdanil@gmail.com

Лебедев Илья Сергеевич — доктор технических наук, профессор, заведующий лабораторией, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 56321781100, https://orcid.org/0000-0001-6753-2181, isl_box@mail.ru

Authors

Daniil D. Tikhonov — PhD Student, Programmer, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, https://orcid.org/0009-0008-0128-4144, tikhonovdanil@gmail.com

Ilya S. Lebedev — D.Sc., Professor, Head of Laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, SC 56321781100, https://orcid.org/0000-0001-6753-2181, isl_box@mail.ru

Статья поступила в редакцию 16.02.2024
Одобрена после рецензирования 19.04.2024
Принята к печати 19.05.2024

Received 16.02.2024
Approved after reviewing 19.04.2024
Accepted 19.05.2024