УДК 621.396
doi:10.31799/1684-8853-2022-3-20-30

Научные статьи
Articles

# Адаптивное применение моделей машинного обучения на отдельных сегментах выборки в задачах регрессии и классификации

***И. С. Лебедев***[a], *доктор техн. наук, профессор, orcid.org/0000-0001-6753-2181, isl_box@mail.ru*
[a]*Санкт-Петербургский Федеральный исследовательский центр РАН, 14-я линия В.О., 39, Санкт-Петербург, 199178, РФ*

***Введение:*** *достижение заданных качественных показателей в решениях, связанных с машинным обучением, зависит не только от эффективности алгоритмов, но и от свойств данных. Одним из направлений развития моделей классификации и регрессии является уточнение локальных свойств информации.* ***Цель:*** *повышение показателей качества при решении задач классификации и регрессии на основе адаптивного выбора различных моделей машинного обучения на отдельных локальных сегментах выборки данных.* ***Результаты:*** *предложен метод, использующий комбинирование различных моделей и алгоритмов машинного обучения на отдельных подвыборках в задачах регрессии и классификации. Метод основывается на вычислении качественных показателей и выборе лучших моделей на локальных сегментах выборки. Выявление изменений данных и временных последовательностей дает возможность сформировать выборки, где данные имеют различные свойства (например, дисперсия, выборочная доля, размах данных и т. д.). Рассмотрено сегментирование на основе алгоритма поиска точек смены тренда временного ряда и аналитической информации. На примере реальных данных датасета приведены экспериментальные значения функции потерь для предлагаемого метода у различных классификаторов на отдельных сегментах и всей выборке.* ***Практическая значимость:*** *результаты могут быть использованы в задачах классификации и регрессии при разработке моделей и методов машинного обучения. Предложенный метод позволяет повысить показатели качества классификации и регрессии за счет назначения моделей, имеющих лучшие показатели на отдельных сегментах.*

***Ключевые слова*** *— машинное обучение, сегментирование множества данных, временные последовательности, изменяющиеся свойства данных.*

**Для цитирования:** Лебедев И. С. Адаптивное применение моделей машинного обучения на отдельных сегментах выборки в задачах регрессии и классификации. *Информационно-управляющие системы*, 2022, № 3, с. 20–30. doi:10.31799/1684-8853-2022-3-20-30
**For citation:** Lebedev I. S. Adaptive application of machine learning models on separate segments of a data sample in regression and classification problems. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2022, no. 3, pp. 20–30 (In Russian). doi:10.31799/1684-8853-2022-3-20-30

## Введение

Применение технологий искусственного интеллекта в различных областях дает возможность добиваться результатов, сопоставимых с деятельностью человека. Современные подходы, используемые в методах машинного обучения, направлены на автоматическое создание моделей на основе выборок, где обучение происходит непосредственно на данных.

В задачах классификации и регрессии оценивается взаимосвязь между входными и выходными переменными, используются алгоритмы оптимизации, которые минимизируют ошибку аппроксимации. Качество обучения модели зависит от свойств совокупности объектов наблюдений, на которых она обучалась [1, 2]. Основная проблема заключается в том, что ошибка аппроксимации сходится к разумным значениям только с большим объемом данных, которые часто трудно получить, анализировать и интерпретировать. Одновременно с этим возникают ситуации, когда собранные значения показателей системы с течением времени могут менять свои свойства. Изменения распределений, частоты событий, дисбаланс классов приводят к тому, что применение моделей и алгоритмов машинного обучения без учета динамически изменяющихся свойств может существенно влиять на результат, увеличивая ошибки [3].

## Обработка информационных потоков

Большинство задач регрессии и предсказания связаны с анализом информационных потоков. Обработка последовательностей и временных рядов имеет определенные особенности. В простейших случаях поступающие данные образуют размеченную выборку, над которой реализуются различные методы обучения [4, 5]. Такие подходы хорошо отражены в классических работах по искусственному интеллекту, в течение длительного времени они подвергались всесторонней оценке и имеют проработанную технологию внедрения и использования. Однако в случае изменения данных и их свойств возникает необходимость поддерживать заданные качественные показатели

алгоритмов, что может являться трудоемкой задачей.

Огромное количество решений задач классификации, регрессии, предсказания поведения системы в динамике использует представление информационных потоков временной последовательностью [6]. Появление моделей и методологий ARMA, ARIMA и других позволило существенно повысить точность прогнозов временного ряда. Однако их построение требует знаний о природе последовательности. Возникает необходимость ее перенастройки при появлении новых данных. Требуются постоянная оценка и подбор различных параметров для достижения заданной точности.

Современная парадигма машинного обучения состоит в том, что модели учатся непосредственно на данных, автоматически вычисляя и оценивая возникающие в выборках закономерности [7]. Поэтому для достижения качественных показателей приходится особое внимание уделять данным.

В работах [8, 9] акцентируется внимание на ряде проблемных вопросов формирования кортежей признаков, создания паттернов поведения, которые подаются на вход классифицирующих алгоритмов. Такие подходы изначально используют статическое представление информации, что не всегда оправдано, особенно для информации, поступающей от реальных систем. Одновременно с этим необходимо решать вопросы длины последовательности; определять характеристики алгоритмов, на основе которых будет производиться разделение выборки; оценивать влияние изменения распределений, частоты событий, дисбаланса классов [10–12].

При обработке данных имеют место вопросы влияния производительности и скорости алгоритмов анализа на качество результатов. В работе [13] предложена оригинальная каскадная модель, элементами которой являются классифицирующие алгоритмы. Однако в ней возрастает сложность агрегации результатов. В ряде других исследований [14–18] отмечается, что особую важность приобретает выделение наиболее информативных признаков, которые вносят основной вклад в задачах классификации и регрессии. Снижение размерности признакового пространства, например методом главных компонент, подсчетом информативности на основе энтропии, частотными методами и т. д., не всегда возможно, в частности, когда имеется одномерный временной ряд [19, 20]. Сокращение размерности признакового пространства позволяет повысить скорость обработки, но с течением времени в случае возникновения эффекта «дрейфа концепта» свойства признаков могут поменяться, что приведет к устареванию классифицирующей модели [8, 21, 22].

В связи с этим возникает необходимость разработать методы и алгоритмы, ориентированные на повышение качественных показателей моделей в условиях изменения свойств данных.

В статье предлагается решение, направленное на повышение показателей качества обработки выборки данных. Рассматривается задача адаптивного применения моделей на отдельных сегментах.

## Описание предлагаемого метода

Одним из путей повышения качества классификации является использование моделей, которые основаны на уточненной локальной информации [23–25]. В большинстве задач обучающая выборка рассматривается как единое множество. Однако составляющие ее кортежи данных могут быть получены под воздействием различных факторов [26]. Например, появление отдельных управляющих команд вызывает рост количества служебных сообщений в сетевом трафике. Смена сезонов года, увеличение продолжительности дня отражаются на потребляемых мощностях в электросетях. В реальных системах возникают ситуации, когда проявляются воздействия, изменяющие их состояния. Часть таких факторов можно определить заранее, другая часть возникает случайно и не поддается прогнозированию. Однако в любом случае цена ошибки может быть очень высока. Предполагая наличие факторов, под влиянием которых происходит изменение значений целевых переменных, можно идентифицировать кортежи, полученные в момент воздействия.

В связи с этим в ряде случаев возникает определенная возможность осуществить сегментирование выборки с учетом информации о действующих факторах, оказывающих влияние на свойства данных:

$x \in X$ — значения выборки данных $X$;

$\{a_1, ..., a_n\} \in A$ — множество моделей, используемых методами машинного обучения для решения практических задач классификации или регрессии;

$\{v_1, ..., v_k\} \in V$ множество воздействующих факторов, которые изменяют диапазоны значений целевых переменных. Часть из них поддаются аналитике и связаны, например, с цикличностью процессов. Влияние других можно определить исходя из анализа изменения свойств данных.

Формализацию воздействующих факторов можно осуществить с помощью функции принадлежности.

$I(v) : X \to M$, $M = \{1, 2, ..., m\}$ — индикаторная функция, разбивающая выборку данных $X$ на

множество сегментов $X^1$, ..., $X^m$, в которых под влиянием фактора $v \in V$ изменялись диапазоны значений целевых переменных.

Функция разделяет последовательность значений на отдельные сегменты

$$(x_1^1, \ldots, x_{n1}^1) \in X^1, (x_1^2, \ldots, x_{n2}^2) \in X^2, \ldots,$$
$$(x_1^m, \ldots, x_{nm}^m) \in X^m,$$

где $\{X^1, \ldots, X^m\} \in X$ — множество сегментов выборки данных $X$. Временная последовательность делится на $m$ отдельных сегментов. Получается разбиение, где данные можно рассматривать как сегменты временных последовательностей. Каждый сегмент $X^i \in X$ обладает своими свойствами (частотой объектов наблюдений, плотностью вероятности распределения данных и т. д.). В ходе разбиения могут возникать сегменты со сходными свойствами. В целях экономии вычислительных ресурсов определяется функция $f(X^i)$, вычисляющая свойства последовательности $X^i$. В случае если значение лежит внутри диапазона $H_0 \le \frac{f(X^i)}{f(X^j)} \le H_1$, где $H_0$ и $H_1$ — пороговые значения, то возможно формирование подвыборок объединением сегментов $X^i = X^i \cup X^j$.

Каждая модель машинного обучения $a_j(x)$ в зависимости от базовых алгоритмов и свойств данных подвыборки имеет свои качественные показатели, которые можно вычислить в процессе обучения.

$Q(a_j(x), X^i)$ — функционал качества модели $a_j(x)$ для подвыборки $X^i$.

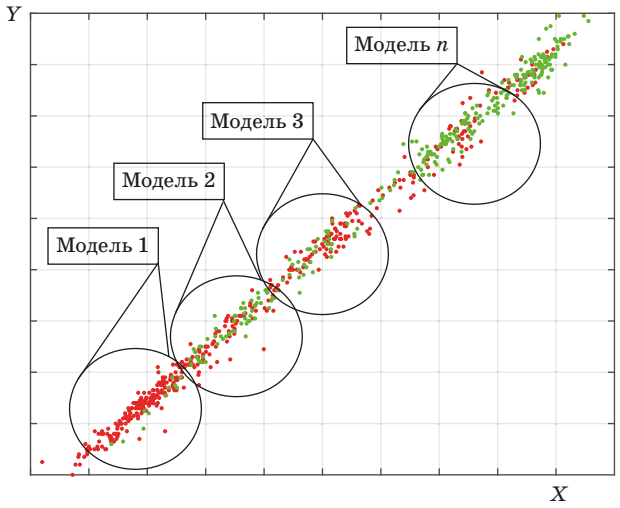
Тогда возникает необходимость выбора модели $a(x)$, обладающей лучшими качественными показателями на подвыборке данных:

$$a(x) = \underset{a_j(x) \in A, X^i \in X}{\arg\max} Q(a_j(x), X^i). \quad (1)$$

В статье рассматривается применение различных моделей машинного обучения на отдельных сегментах выборки данных.

Для каждой модели используются свои словари признаков, которые отличаются друг от друга. В статистических методах могут определяться значения дисперсии, выборочной доли, размаха данных и т. д. При обработке временных рядов, например методом скользящих средних, необходимо вычислять последовательность средних, ширину окна.

Предлагаемый метод иллюстрирует рис. 1. Область данных делится на отдельные сегменты. В зависимости от свойств данных на каждый сегмент назначается своя модель. Выбор модели и ее назначение определяются на основе значений функционала качества.




■ ***Рис. 1.*** Посегментное использование моделей
■ ***Fig. 1.*** Models processing in local segments

Таким образом, в основе предлагаемого метода лежит сегментация выборки, в результате которой необходимо выявить основные свойства входящих в нее данных и исходя из этого назначить наиболее подходящую из заранее предопределенных моделей.

## Метод адаптивного применения моделей на отдельных сегментах выборки

Одним из проблемных вопросов адаптации моделей машинного обучения является отсутствие эффективных методов предобработки информации, направленных на вычисление и анализ свойств, позволяющих в режиме реального времени разделять поступающие последовательности на сегменты. Комплекс таких методов должен не только решать обычные задачи фильтрации, удаления шумов и выбросов, но и предоставлять информацию о свойствах данных для выбора и определения наиболее подходящих моделей. В целях решения обозначенных проблемных вопросов применяются постоянно обучающиеся модели.

Пример последовательности шагов постоянно обучающейся модели показан на рис. 2. Модель является двухуровневой. На первом уровне происходит обработка постоянного информационного потока, на втором действуют процедуры, обеспечивающие реализацию «механизма» обучения. Особенностью представленного решения является сегментирование обучающей выборки.

Для начального запуска процессов необходимо иметь предварительную информацию о значениях $x_1$, ..., $x_n$ информационной последовательности. Они входят в первоначальное обучающее


■ ***Рис. 2.*** Пример последовательности шагов постоянно обучающейся модели

■ ***Fig. 2.*** A sequence steps example of constantly learning model

множество. На верхнем уровне модели выборка анализируется в целях определения отдельных сегментов, где свойства данных различаются. Возможно ее разделение как на основе заданной заранее системы правил, так и с помощью алгоритмов, выполняющих в автоматическом режиме поиск характерных точек, где изменяются свойства поступающих рядов.

В первом случае происходит изучение последовательности. На основе анализа совокупности данных выделяются тренды, периоды, сегменты, кластеры, обладающие отличающимися характеристиками. Эффективность такого подхода к разбиению определяется полнотой знаний о воздействующих факторах, под влиянием которых меняются диапазоны значений целевых переменных, частот событий, распределения вероятностей. В результате получается статичная система, настройка которой при изменении свойств может быть сложной.

Во втором случае разделение объектов наблюдения можно осуществить с помощью моделей, методов, алгоритмов, вычисляющих точки разладки, смену концепции. С помощью алгоритма автоматически определяются границы сегментов. Однако априорная информация о моделях смены концепции или разладки временного ряда может быть ограничена или отсутствовать. Недостатки моделей связаны с необходимостью повысить эффективность процедур детектирования изменения свойств временных последовательностей. Требуется постоянно отслеживать текущие настройки и базовые параметры при появлении новых данных.

Цель сегментирования состоит в том, чтобы обнаружить ситуации трансформации свойств последовательностей данных. Это осуществляется поиском момента θ, где происходит изменение характеристик наблюдаемого процесса:

$$x_t^i = \begin{cases} x_t^i, & 0 < t < \theta_i \\ x_t^{i+1}, & t \geq \theta_i \end{cases}.$$

В результате исходная выборка делится на несколько частей $X^1$, ..., $X^m$. Их свойства анализируются, и если имеется совпадение, где заранее определенные параметры одинаковы, то можно уменьшить количество рассматриваемых сегментов.

Подвыборки $X^1$, ..., $X^m$ поступают на вход моделей $a_1$, $a_2$, ..., $a_m$. Происходит их обучение и анализ достигаемых качественных показателей. На каждом сегменте $X^i$ для каждой модели $a_j(x)$ определяется функционал качества $Q(a_j(x), X^i)$. На основе его значений возможно ранжировать модели $\{a_1, ..., a_n\} \in A$ и осуществлять выбор имеющих наиболее высокие качественные показатели для каждого сегмента. В качестве условия выбора рассматривается выражение (1).

Процедуры сегментирования и определения свойств последовательности данных выполняются при обработке поступающего потока. Анализ свойств сегментов, выявленных при обработке информационного потока, и сопоставление их со свойствами подвыборок, полученных из обучающей выборки, позволяют назначить одну из заранее обученных моделей $\{a_1, ..., a_n\} \in A$ на текущий сегмент.

На последнем этапе выбранная модель $a_j(x)$ используется для решения задач обработки потока. Полученные результаты сравниваются с имеющимися, производится их анализ. Сопоставление

полученных моделью и реальных значений позволяет принять решение о формировании данных для уточнения алгоритма, которые впоследствии добавляются в обучающую выборку.

Таким образом, возможна реализация постоянно обучающейся модели, где процессы обучения и обработки информационных потоков могут осуществляться параллельно. В случае использования сложных моделей классификации или регрессии заранее предобученные модели позволяют уменьшить временные затраты на обучение при изменении свойств данных.

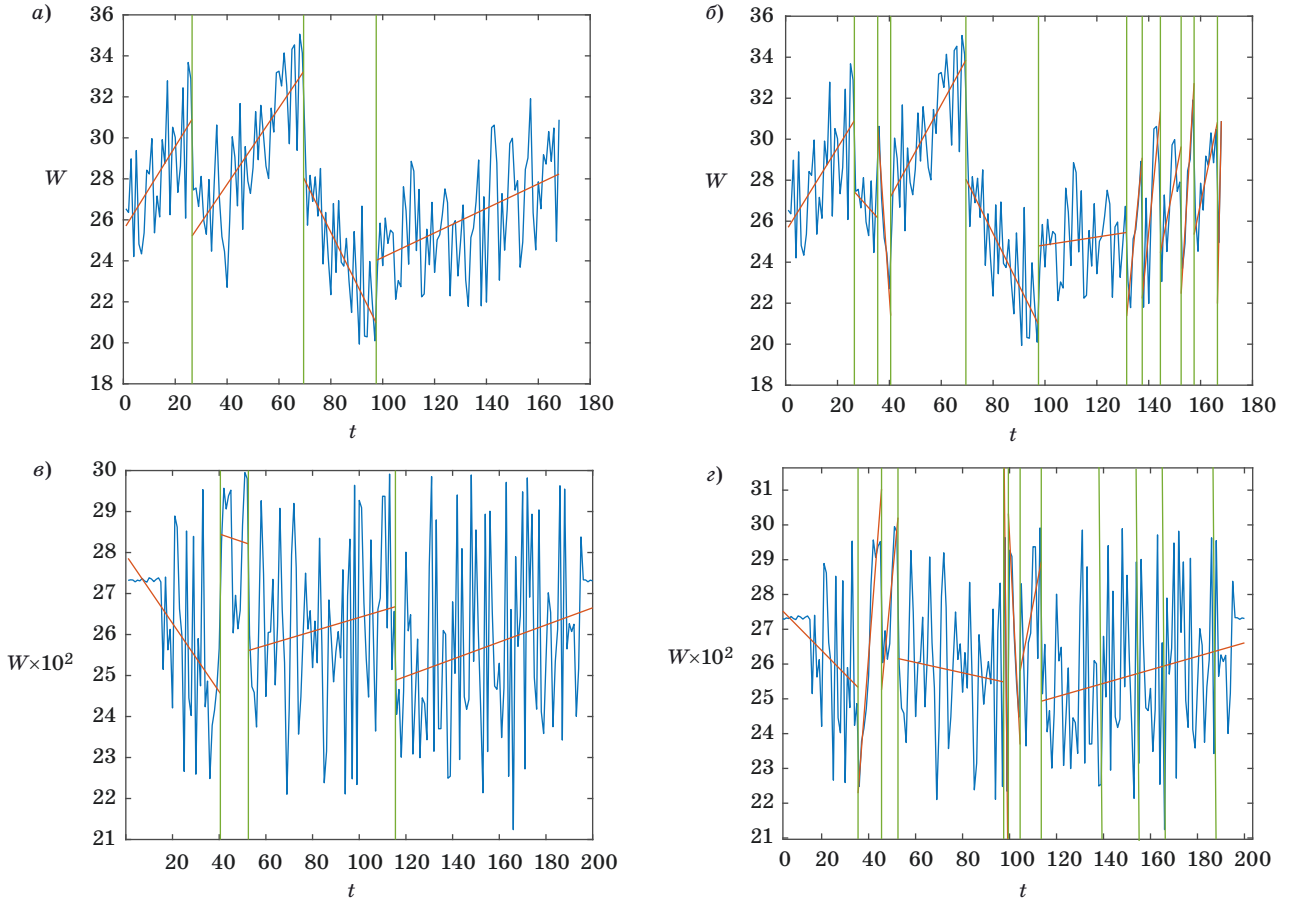
## Эксперимент

Анализ применения предложенного подхода осуществлялся на данных, содержащих информацию о почасовой генерации электроэнергии солнечными и ветровыми электростанциями.

Целью эксперимента являлся анализ влияния размеров и способов получения сегментов данных на достигаемые качественные показатели в задачах регрессии по сравнению с целой выборкой. В первом случае данные датасета были разбиты на четыре части по кварталам и на двенадцать частей по месяцам согласно информации календаря. Во втором случае разбиение производилось с помощью алгоритма поиска точек смены направления тренда [27, 28]. Параметры алгоритма были подобраны таким образом, чтобы осуществить разбиение автоматическим способом также на четыре и двенадцать частей.

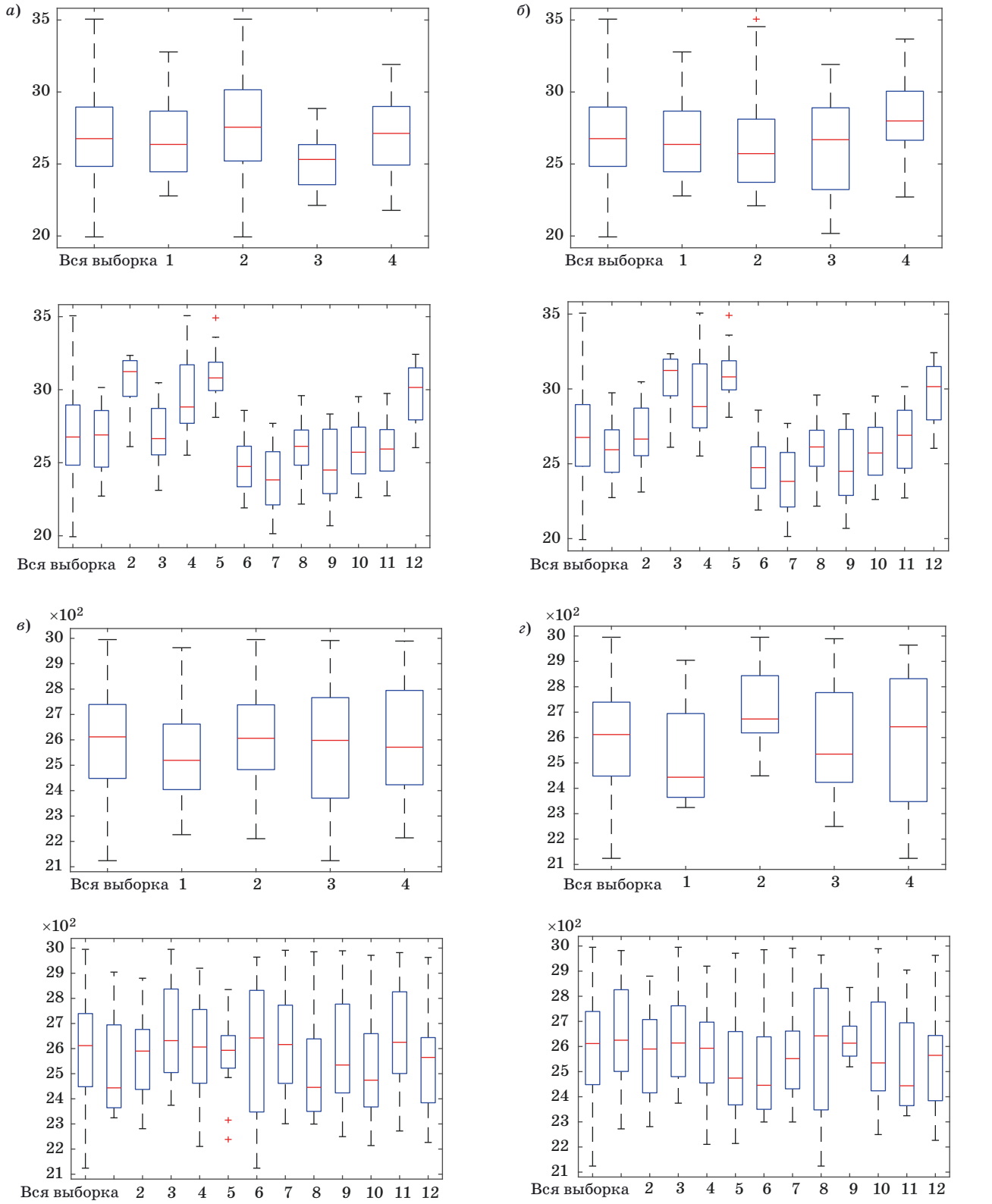
Получены сегменты временных последовательностей значений для солнечной и ветровой генерации энергии (рис. 3, *а–г*), в которых в целях определения свойств проводился статистический анализ данных.

Диаграммы результатов обработки значений целевых значений генерации электроэнергии солнечными батареями и ветряными установками (рис. 4, *а–г*) отражают медианное значение, первый и второй квартили, разброс. На диаграммах виден больший разброс значений для всей выборки в представленных множествах по сравнению с сегментами по месяцам или кварталам и



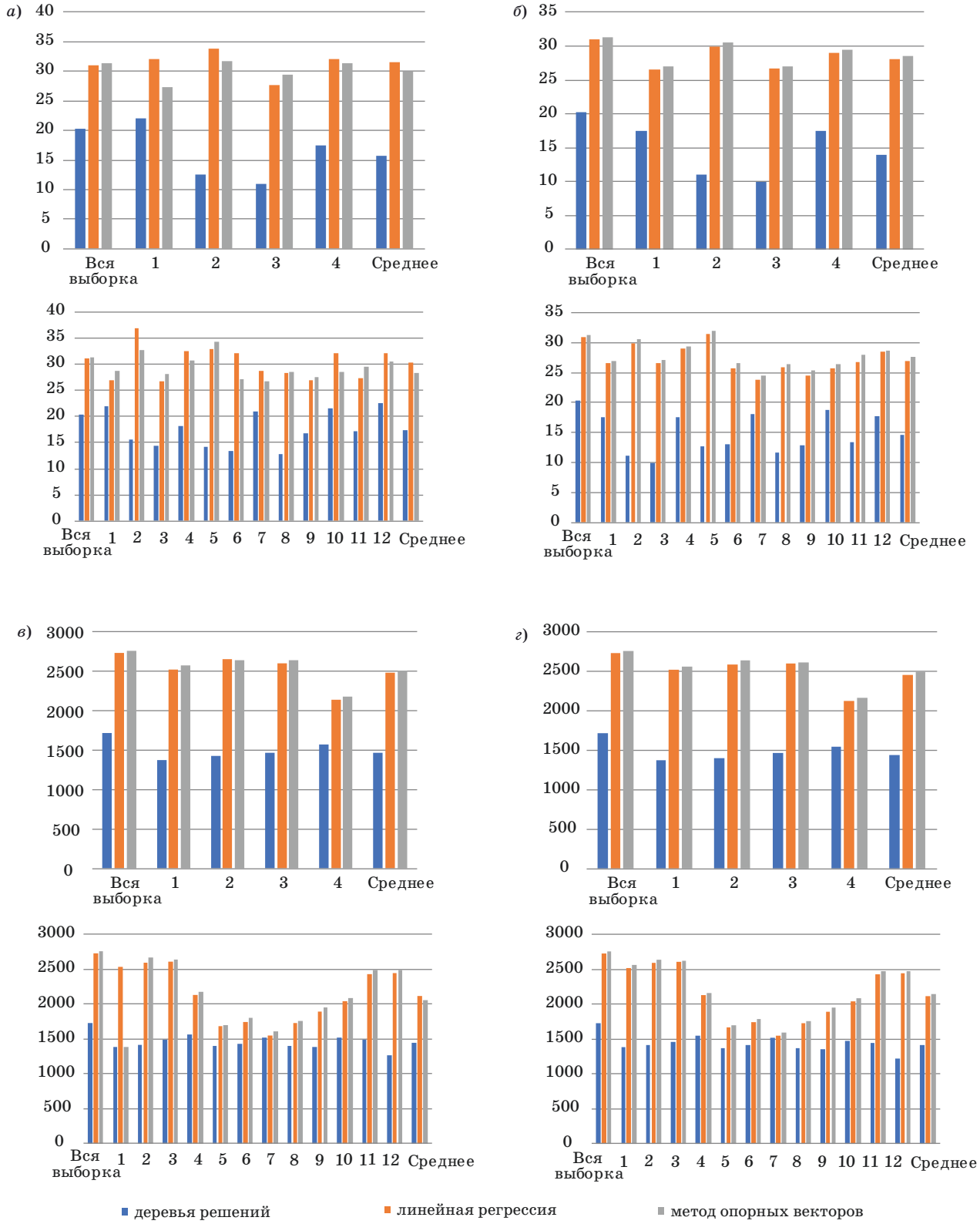

■ ***Рис. 3.*** Сегментирование датасета временной последовательности генерации солнечной (*а*, *б*) и ветровой (*в*, *г*) энергии

■ ***Fig. 3.*** Segmentation dataset of solar energy generation (*а*, *б*) and wind energy (*в*, *г*) time sequences values




■ ***Рис. 4.*** Свойства выборки данных при сегментации: по календарной информации для временной последовательности генерации солнечной (*а*) и ветровой (*в*) энергии; на основе алгоритма для временной последовательности генерации солнечной (*б*) и ветровой (*г*) энергии

■ ***Fig. 4.*** Data sampling properties: segmentation by calendar information for the time sequence of solar energy generation (*а*), wind energy (*в*); segmentation based on the algorithm for the time sequence of solar energy generation (*б*), wind energy (*г*)




■ ***Рис. 5.*** Функции потерь RMSE регрессионных моделей предсказания генерации электроэнергии при сегментации: по календарной информации для временной последовательности генерации солнечной (*а*) и ветровой (*в*) энергии; на основе алгоритма для временной последовательности генерации солнечной (*б*) и ветровой (*г*) энергии

■ ***Fig. 5.*** Loss functions RMSE of regression models in predicting electricity generatio: segmentation by calendar information for the time sequence of solar energy generation (*a*), wind energy (*в*); segmentation based on the algorithm for the time sequence of solar energy generation (*б*), wind energy (*г*)

сегментами, выявленными автоматическим способом. Применительно к рассматриваемому датасету диаграммы демонстрируют, что, несмотря на возможные «выбросы» данных, при сегментировании выборки диапазон между крайними значениями уменьшается по сравнению со всей выборкой в целом (что иллюстрирует визуальное сравнение сегментов с левым элементом «Вся выборка»). Применение сегментирования в ряде случаев уменьшает размах данных, частично борется с выбросами.

Для оценки влияния разделения на сегменты выборки на качество результатов машинного обучения были выбраны различные модели: линейной регрессии, деревьев решений и метод опорных векторов.

Данные представлялись одномерными временными рядами. На практике реальны более сложные модели. Рассматривалась возможность повышения качества за счет адаптивного выбора моделей. Выбор алгоритма определялся низкой вычислительной сложностью. На каждую модель подавались вся выборка полностью и данные из разделенных сегментов.

В качестве меры оценки алгоритма регрессии была выбрана функция потерь RMSE — классическая регрессионная метрика с одним выходом, которая вычисляет абсолютную разницу между прогнозируемыми и фактическими выходными данными:

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

где $y_i = a(x_i)$ — результат предсказания выбранного алгоритма; $\hat{y}_i$ — фактическое значение целевой переменной.

Функции потерь регрессионных моделей для различных сегментов данных согласно выражению (2) показаны на рис. 5, *а–г*. На гистограммах видно, что в результате разделения временной последовательности удается получить значения функции потерь меньше как в среднем, так и на большинстве отдельных сегментов.

При сравнении выбранных моделей лучшие результаты показывает алгоритм деревьев решений. При делении выборки на четыре части этот классификатор имеет значения функции потерь меньше, чем у других алгоритмов, во всех сегментах. Однако при делении на 12 сегментов на отдельных подвыборках его опередили метод опорных векторов и линейная регрессия. Анализируя свойства данных, можно назначать разные модели на отдельные сегменты.

Сегментация данных дает возможность уменьшить функцию потерь для разных областей выборки. Алгоритм поиска точек смены направления тренда позволяет выделить отдельные сегменты с меньшим размахом данных, что определяет более низкие значения функции потерь в среднем в регрессионной задаче.

Выделение сегментов последовательностей информационного потока данных и оценка их свойств позволяют осуществлять поиск и выбор моделей машинного обучения, обладающих лучшими характеристиками. Произведя оценку гистограмм рис. 5, сравнив значения функции потерь классификаторов на отдельных сегментах с левым столбцом «Вся выборка», видим, что на отдельных сегментах алгоритмы имеют лучшие результаты, чем при обработке всей выборки целиком. Результаты показывают, что применение предложенного метода, где каждому сегменту выборки данных назначается модель, имеющая на нем лучшие показатели качества, дает возможность уменьшить значения функции потерь RMSE от 8 до 18 % по сравнению с обработкой выборки целиком.

Предварительное обучение на выборках со сходными свойствами может сократить время на подготовку модели. Анализ результатов, полученных моделью, и реальных значений последовательности возможно использовать для формирования обучающих данных в целях уточнения модели. В дальнейшем осуществимо построение иерархий, когда модель верхнего уровня применяется для назначения наиболее эффективной модели нижнего уровня на отдельный сегмент.

## Заключение

Одним из направлений, связанных с увеличением качественных показателей моделей классификации, является повышение качества данных, поступающих на вход алгоритмов. Для этих целей предложен метод, использующий адаптивное применение моделей машинного обучения на отдельных сегментах выборки. Сегментация, в определенных случаях, позволяет уменьшить разброс данных, выбросов и использовать изменение диапазонов значений переменных для повышения качества моделей.

Применение предложенного метода, основанного на разделении данных и выбора моделей с лучшими качественными показателями, помогает уменьшить значения функции потерь по сравнению с обработкой выборки целиком. Разделение последовательностей дает возможность бороться с выбросами и шумами и формировать компактно локализованные подмножества в пространстве объектов.

Свойства данных, на которых обучаются и тестируются регрессионные модели, влияют на их эффективность. Анализ информации об из-

менении диапазонов значений, балансов событий используется для формирования обучающих выборок в целях локального повышения качественных показателей моделей.

Новизна предлагаемого метода заключается в том, что с помощью правил или алгоритмов выборка разделяется на отдельные сегменты, каждый из которых обладает своими свойствами. Предварительное обучение на них алгоритмов способствует при изменении свойств потоков данных выбирать и назначать модели, обладающие лучшими качественными показателями.

На пути дальнейшего развития метода возможна его адаптация для задач прогнозирования и проактивного управления, направленного на оценку развития ситуации в динамике. Разбиение временных рядов на отдельные сегменты, анализ и сопоставление свойств последовательностей могут служить информацией для определения состояний. В результате можно выявить последовательности сегментов и определить переходы состояний. Обработка последовательностей, выделение сегментов предоставляют информацию, на основе которой можно строить графы и матрицы переходов. Анализ переходов состояний делает возможным в текущий дискретный момент времени определение наиболее вероятных переходов из текущего состояния в последующие.

## Литература


1. **Di Franco G., Santurro M.** Machine learning, artificial neural networks and social research. *Qual Quant*, 2021, no. 5, pp. 1007–1025. https://doi.org/10.1007/s11135-020-01037-y
2. **Bonikowski B., Di Maggio P.** Varieties of American popular nationalism. *American Sociological Review*, 2016, no. 81(5), pp. 949–980.
3. **Sabar N. R., Ayob M., Kendall G., Qu R.** A dynamic multiarmed bandit-gene expression programming hyper-heuristic for combinatorial optimization problems. *IEEE Transactions on Cybernetics*, 2014, no. 45(2), pp. 217–228.
4. **Park J., Kim S.** Machine learning-based activity pattern classification using personal PM2.5 exposure information. *International Journal of Environmental Research and Public Health*, 2020, no. 17(18), pp. 65–73. https://doi.org/10.3390/ijerph17186573
5. **Maletzke A., dos Reis D., Batista G.** Combining instance selection and self-training to improve data stream quantification. *Journal of the Brazilian Computer Society*, 2018, vol. 24, no. 12, pp. 123–141. doi:10.1186/s13173-018-0076-0
6. **Jordan M. I., Mitchell T. M.** Machine learning: Trends, perspectives, and prospects. *Science*, 2015, no. 349(6245), pp. 255–260.
7. **Fanaee T. H., Gama J.** Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2014, no. 2, pp. 113–127. https://doi.org/10.1007/s13748-013-0040-3
8. **Bishop C. M., Nasser M. N.** Pattern recognition and machine learning. *Journal of Electronic Imaging*, 2007, no. 16, pp. 49–69.
9. **Sukhoparov M. E., Semenov V. V., Salakhutdinova K. I., Boitsova E. P., Lebedev I. S.** The state identification of industry 4.0 mechatronic elements based on behavioral patterns. *Internet of Things, Smart Spaces, and Next Generation Networks and Systems: Proc. of 20th Intern. Conf., NEW2AN 2020, and 13th Conf., ruSMART 2020*, Saint-Petersburg, Russia, August 26–28, 2020, Part I, Aug 2020, pp. 126–134. https://doi.org/10.1007/978-3-030-65726-0_12
10. **Oikarinen E., Tiittanen H., Henelius A.** Detecting virtual concept drift of regressors without ground truth values. *Data Mining and Knowledge Discovery*, 2021, vol. 35, iss. 3, pp. 821–859. doi:10.1007/s10618-021-00739-7
11. **Lu J., Liu A., Dong F., Gu F., Gama J., Zhang G.** Learning under concept drift: a review. *IEEE Transactions on Knowledge and Data Engineering*, 2019, no. 31(12), pp. 2346–2363.
12. **Wang L. Y., Park C., Yeon K., Choi H.** Tracking concept drift using a constrained penalized regression combiner. *Computational Statistics & Data Analysis*, 2017, no. 108, pp. 52–69.
13. **Lei P., Todorovic S.** Temporal deformable residual networks for action segmentation in videos. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 6742–6751.
14. **Khan S., Yairi T.** A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 2018, no. 107, pp. 241–265. doi:10.1016/j.ymssp.2017.11.024
15. **Zhou Z.-H., Feng J.** Deep forest. *National Science Review*, 2019, vol. 6, no. 1, pp. 74–86. doi:10.1093/nsr/nwy108
16. **Salehi H., Burgueño R.** Emerging artificial intelligence methods in structural engineering. *Engineering Structures*, 2018, no. 171, pp. 170–189. doi:10.1016/j.engstruct.2018.05.084
17. **Zissis D., Lekkas D.** Addressing cloud computing security issues. *Future Generation Computer Systems*, 2012, vol. 28, no. 3, pp. 583–592.
18. **Liu J., Li Y., Song S., Xing J., Lan C., Zeng W.** Multi-modality multi-task recurrent neural network for online action detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018, no. 29(9), pp. 2667–2682.
19. **Татарникова Т. М., Богданов П. Ю.** Обнаружение атак в сетях интернета вещей методами машинного обучения. *Информационно-управляющие систе-*

*мы*, 2021, № 6, с. 42–52. doi:10.31799/1684-8853-2021-6-42-52
20. **Зегжда Д. П., Калинин М. О., Крундышев В. М., Лаврова Д. С., Москвин Д. А., Павленко Е. Ю.** Применение алгоритмов биоинформатики для обнаружения мутирующих кибератак. *Информатика и автоматизация*, 2021, № 4(20), с. 820–844.
21. **Chao Y. W., Vijayanarasimhan S., Seybold B.** Rethinking the faster r-cnn architecture for temporal action localization. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 1130–1139.
22. **Nguyen P., Liu T., Prasad G.** Weakly supervised action localization by sparse temporal pooling network. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 6752–6761.
23. **Atsuya O., Genki Y.** Computational mechanics enhanced by deep learning. *Computer Methods in Applied Mechanics and Engineering*, 2017, vol. 327, pp. 327–351. https://doi.org/10.1016/j.cma.2017.08.040
24. **Takacs A., Toledano-Ayala M., Dominguez-Gonzalez A., Pastrana-Palma A., Velazquez D. T., Ramos J. M., Rivas-Araiza A. E.** Descriptor generation and optimization for a specific out-door environment. *IEEE Access*, 2020, vol. 8, pp. 2169–2176. doi:10.1109/ACCESS.2020.2975474
25. **Wong J. C., Lian H., Cheong S. A.** Detecting macroeconomic phases in the Dow Jones Industrial Average time series. *Physica A: Statistical Mechanics and its Applications*, 2009, no. 388 (21), pp. 4635–4645.
26. **Лебедев И. С.** Сегментирование множества данных с учетом информации воздействующих факторов. *Информационно-управляющие системы*, 2021, № 3, с. 29–38. doi:10.31799/1684-8853-2021-3-29-38
27. **Killick R., Paul F., Eckley I. A.** Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 2012, vol. 107, no. 500, pp. 1590–1598.
28. **Lavielle M.** Using penalized contrasts for the change-point problem. *Signal Processing*, 2005, vol. 85, pp. 1501–1510.

UDC 621.396
doi:10.31799/1684-8853-2022-3-20-30

**Adaptive application of machine learning models on separate segments of a data sample in regression and classification problems**

I. S. Lebedev[a], Dr. Sc., Tech., Professor, orcid.org/0000-0001-6753-2181, isl_box@mail.ru
[a]St. Petersburg Federal Research Center of the RAS, 39, 14th Line, 199178, Saint-Petersburg, Russian Federation

**Introduction:** Achievement of specified qualitative indicators in machine learning solutions depends not only on the efficiency of algorithms, but also on data properties. One of the lines for the development of classification and regression models is the specification of local properties of data. **Purpose:** To improve the qualitative predictors when solving classification and regression problems based on the adaptive selection of various machine learning models on separate local segments of data sample. **Results:** We propose a method that uses a combination of different models and machine learning algorithms on subsamples in regression and classification problems. The method is based on the calculation of qualitative predictors and the selection of the best models on the local segments of data sample. The finding of transformations of data and time series allows to create sample sets, with the data having different properties (for example, variance, sampling fraction, data range, etc.). We consider the data segmentation based on the change point detection algorithm in time series trends and on analytical information. On the example of the real dataset, we show the experimental values of the loss function for the proposed method with different classifiers on separate segments and on the whole sample. **Practical relevance:** The results can be used in classification and regression problems for the development of machine learning models and methods. The proposed method allows to improve classification and regression qualitative predictors by assigning models that have the best performance on separate segments.

**Keywords** — machine learning, data segmentation, time series, data transformations.

**For citation:** Lebedev I. S. Adaptive application of machine learning models on separate segments of a data sample in regression and classification problems. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2022, no. 3, pp. 20–30 (In Russian). doi:10.31799/1684-8853-2022-3-20-30

**Reference**

1. Di Franco G., Santurro M. Machine learning, artificial neural networks and social research. *Qual Quant*, 2021, no. 5, pp. 1007–1025. https://doi.org/10.1007/s11135-020-01037-y
2. Bonikowski B., Di Maggio P. Varieties of American popular nationalism. *American Sociological Review*, 2016, no. 81(5), pp. 949–980.
3. Sabar N. R., Ayob M., Kendall G., Qu R. A dynamic multi-armed bandit-gene expression programming hyper-heuristic for combinatorial optimization problems. *IEEE Transactions on Cybernetics*, 2014, no. 45(2), pp. 217–228.
4. Park J., Kim S. Machine learning-based activity pattern classification using personal PM2.5 exposure information. *International Journal of Environmental Research and Public Health*, 2020, no. 17(18), pp. 65–73. https://doi.org/10.3390/ijerph17186573
5. Maletzke A., dos Reis D., Batista G. Combining instance selection and self-training to improve data stream quantification. *Journal of the Brazilian Computer Society*, 2018, vol. 24, no. 12, pp. 123–141. doi:10.1186/s13173-018-0076-0
6. Jordan M. I., Mitchell T. M. Machine learning: Trends, perspectives, and prospects. *Science*, 2015, no. 349(6245), pp. 255–260.
7. Fanaee T. H., Gama J. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2014, no. 2, pp. 113–127. https://doi.org/10.1007/s13748-013-0040-3

8. Bishop C. M., Nasser M. N. Pattern recognition and machine learning. *Journal of Electronic Imaging*, 2007, no. 16, pp. 49–69.
9. Sukhoparov M. E., Semenov V. V., Salakhutdinova K. I., Boitsova E. P., Lebedev I. S. The state identification of industry 4.0 mechatronic elements based on behavioral patterns. *Internet of Things, Smart Spaces, and Next Generation Networks and Systems: Proc. of 20th Intern. Conf., NEW2AN 2020, and 13th Conf., ruSMART 2020*, Saint-Petersburg, Russia, August 26–28, 2020, Part I, Aug 2020, pp. 126–134. https://doi.org/10.1007/978-3-030-65726-0_12
10. Oikarinen E., Tiittanen H., Henelius A. Detecting virtual concept drift of regressors without ground truth values. *Data Mining and Knowledge Discovery*, 2021, vol. 35, iss. 3, pp. 821–859. doi:10.1007/s10618-021-00739-7
11. Lu J., Liu A., Dong F., Gu F., Gama J., Zhang G. Learning under concept drift: a review. *IEEE Transactions on Knowledge and Data Engineering*, 2019, no. 31(12), pp. 2346–2363.
12. Wang L. Y., Park C., Yeon K., Choi H. Tracking concept drift using a constrained penalized regression combiner. *Computational Statistics & Data Analysis*, 2017, no. 108, pp. 52–69.
13. Lei P., Todorovic S. Temporal deformable residual networks for action segmentation in videos. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 6742–6751.
14. Khan S., Yairi T. A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 2018, no. 107, pp. 241–265. doi:10.1016/j.ymssp.2017.11.024
15. Zhou Z.-H., Feng J. Deep forest. *National Science Review*, 2019, vol. 6, no. 1, pp. 74–86. doi:10.1093/nsr/nwy108
16. Salehi H., Burgueño R. Emerging artificial intelligence methods in structural engineering. *Engineering Structures*, 2018, no. 171, pp. 170–189. doi:10.1016/j.engstruct.2018.05.084
17. Zissis D., Lekkas D. Addressing cloud computing security issues. *Future Generation Computer Systems*, 2012, vol. 28, no. 3, pp. 583–592.
18. Liu J., Li Y., Song S., Xing J., Lan C., Zeng W. Multi-modality multi-task recurrent neural network for online action detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018, no. 29(9), pp. 2667–2682.
19. Tatarnikova T. M., Bogdanov P. Yu. Intrusion detection in internet of things networks based on machine learning methods. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2021, no. 6, pp. 42–52 (In Russian). doi:10.31799/1684-8853-2021-6-42-52
20. Zegzhda D., Kalinin M., Kundyshev V., Lavrova D., Moskvin D., Pavlenko E. Application of bioinformatics algorithms for polymorphic cyberattacks detection. *Informatics and Automation (SPIIRAS Proc.)*, 2021, no. 4(20), pp. 820–844 (In Russian).
21. Chao Y. W., Vijayanarasimhan S., Seybold B. Rethinking the faster r-cnn architecture for temporal action localization. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 1130–1139.
22. Nguyen P., Liu T., Prasad G. Weakly supervised action localization by sparse temporal pooling network. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 6752–6761.
23. Atsuya O., Genki Y. Computational mechanics enhanced by deep learning. *Computer Methods in Applied Mechanics and Engineering*, 2017, vol. 327, pp. 327–351. https://doi.org/10.1016/j.cma.2017.08.040
24. Takacs A., Toledano-Ayala M., Dominguez-Gonzalez A., Pastrana-Palma A., Velazquez D. T., Ramos J. M., Rivas-Araiza A. E. Descriptor generation and optimization for a specific out-door environment. *IEEE Access*, 2020, vol. 8, pp. 2169–2176. doi:10.1109/ACCESS.2020.2975474
25. Wong J. C., Lian H., Cheong S. A. Detecting macroeconomic phases in the Dow Jones Industrial Average time series. *Physica A: Statistical Mechanics and its Applications*, 2009, no. 388 (21), pp. 4635–4645.
26. Lebedev I. S. Dataset segmentation considering the information about impact factors. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2021, no. 3, pp. 29–38 (In Russian). doi:10.31799/1684-8853-2021-3-29-38
27. Killick R., Paul F., Eckley I. A. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 2012, vol. 107, no. 500, pp. 1590–1598.
28. Lavielle M. Using penalized contrasts for the change-point problem. *Signal Processing*, 2005, vol. 85, pp. 1501–1510.


## УВАЖАЕМЫЕ АВТОРЫ!

Научные базы данных, включая Scopus и Web of Science, обрабатывают данные автоматически. С одной стороны, это ускоряет процесс обработки данных, с другой — различия в транслитерации ФИО, неточные данные о месте работы, области научного знания и т. д. приводят к тому, что в базах оказывается несколько авторских страниц для одного и того же человека. В результате для всех по отдельности считаются индексы цитирования, что снижает рейтинг ученого.

Для идентификации авторов в сетях Thomson Reuters проводит регистрацию с присвоением уникального индекса (ID) для каждого из авторов научных публикаций.

Процедура получения ID бесплатна и очень проста, есть возможность провести регистрацию на 12 языках, включая русский (чтобы выбрать язык, кликните на зеленое поле вверху справа на стартовой странице): https://orcid.org