УДК 621.396
doi:10.31799/1684-8853-2025-3-25-36
EDN: SSGKZU

Научные статьи
Articles

Обработка информационных последовательностей с использованием адаптивного анализа сегментов при оценке состояния систем

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: *формирование и разметка выборок является трудоемкой процедурой, играющей важную роль в процессах обучения с учителем и настройки большого количества моделей машинного обучения. При решении задач с применением методов искусственного интеллекта достаточно часто возникает необходимость снизить затраты на маркировку данных.* ***Цель:*** *повысить качество обработки информационных последовательностей за счет формирования, анализа и определения сегментов последовательностей данных, на которых заранее заданными алгоритмами машинного обучения достигаются лучшие показатели качества.* ***Результаты:*** *предложен метод формирования сегментов информационной последовательности на основе анализа показателей качества моделей обработки, отличающийся от известных, осуществляющих настройку моделей машинного обучения на обрабатываемые данные, разделением последовательности на сегменты и выбором способа сегментирования таким образом, чтобы свойства полученных в сегменте данных как можно лучше соответствовали модели обработки. В отличие от классического подхода, когда модель настраивается на данные, в предлагаемом методе сегментированием последовательности данные настраиваются под модель.* ***Практическая значимость:*** *результаты могут быть использованы в моделях и методах, решающих задачи классификации и прогнозирования. Предложенный метод позволяет частично преодолевать ряд проблемных вопросов, связанных с маркировкой выборок данных. В результате становится возможным обучать модели, используя выборки, которые имеют частичные или неточные метки, а также снизить затраты на процесс разметки. Дальнейшее развитие предложенного решения возможно на основе ансамблевых методов.*

Ключевые слова *— машинное обучение, адаптивные модели, повышение качества обработки, адаптивная обработка сегментов последовательностей.*

Для цитирования: Лебедев И. С. Обработка информационных последовательностей с использованием адаптивного анализа сегментов при оценке состояния систем. *Информационно-управляющие системы*, 2025, № 3, с. 25–36. doi:10.31799/1684-8853-2025-3-25-36, EDN: SSGKZU

For citation: Lebedev I. S. Sequential information processing using adaptive pattern analysis in assessing the state of systems. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2025, no. 3, pp. 25–36 (In Russian). doi:10.31799/1684-8853-2025-3-25-36, EDN: SSGKZU

Введение

Традиционные подходы машинного обучения «с учителем» нуждаются в большом количестве размеченных данных. Однако из-за различных факторов, таких как шумы, большая размерность признакового пространства, невозможность однозначной классификации состояния, разметка выборок больших объемов может являться трудоемкой задачей. В связи с этим возникает необходимость в разработке методов обучения на малом количестве данных.

Часто выборки данных трудно интерпретируются для обучения моделей, что приводит к развитию методов слабоконтролируемого обучения, выявляющего значимые шаблоны, паттерны, неявные закономерности данных без участия человека [1]. Слабоконтролируемое обучение направлено на снижение затрат на маркировку данных. Его использование в ряде практических задач может помочь уменьшить потребность в больших объемах помеченных данных в условиях необходимости достижения относительно высоких качественных показателей обработки, где процессы разметки затруднены [2].

Формирование пространства признаков нередко имеет значение в обработке данных. Оно играет важную роль в повышении качества результатов, снижении вычислительной сложности.

Методы и модели обработки последовательностей

Повышение показателей качества обработки данных является одной из фундаментальных задач методов машинного обучения. Основное направление решения связано с формированием эффективных моделей обработки данных [3–5]. Достигаемые значения показателей качества обработки базовых алгоритмов, таких как, например, наивный байесовский классификатор,

линейный дискриминант, деревья решений, зависят от свойств обрабатываемых выборок. В случае изменения распределений, частоты событий, трендов простые модели могут терять свою адекватность.

Для нивелирования подобных недостатков и повышения качества обработки информационных последовательностей применяются подходы, направленные на формирование ансамблей моделей и алгоритмов, сочетающих несколько методов машинного обучения [6, 7] (табл. 1 [8]).

Помимо приведенных в таблице ансамблевых методов, при формировании сложных многоуровневых моделей могут использоваться системы выборочного, взвешенного голосования; системы принятия решений, основанные на правилах, априорных знаниях о данных; каскады простых моделей и глубоких нейронных сетей [9–11]. Основными их недостатками являются сложность обучения, агрегации результатов, ресурсоемкость и увеличение времени работы алгоритмов. Неправильно подобранные модели и способы агрегации их результатов могут приводить к ухудшению общего прогноза.

Другое направление повышения качественных показателей обработки связано с формированием пространства признаков. Его обоснование осуществляется с помощью качественных показателей. Это могут быть различные статистические метрики, например измерения расстояния, меры согласованности, меры корреляции, меры на основе теории информации, либо функции потерь и функционалы качества в методах машинного обучения [12–15]. В небольших наборах данных вычисление ряда статистических метрик может быть затруднено или непрактично.

Среди методов, формирующих пространство признаков для задач обработки данных, можно выделить решения, разделяющие данные на основе кластеризации, динамического программирования, бинарной сегментации, байесовские методы [16–24]. Таких методов довольно много.

Основные подходы к сегментации последовательности, разделяющей ее на подпоследовательности, и их характеристики представлены в табл. 2.

На основе описанных подходов реализуются различные методы разделения последовательностей.

Применение методов зависит от вида и свойств обрабатываемой информации. Целью таких подходов является повышение выбранных показателей качества моделей обучения за счет определения дополнительной информации.

■ ***Таблица 1.*** Характеристики основных ансамблевых методов
■ ***Table 1.*** Characteristics of the main ensemble methods

Основные ансамблевые методы	Предпосылки к применению	Возможные ограничения
Bagging	Уменьшает влияние ошибок одной модели, повышая общую точность ансамбля; снижает чувствительность при трансформации свойств данных; позволяет осуществлять параллельное обучение моделей обработки	В случае «неправильно» подобранных моделей может ухудшить результат; при обучении сложных моделей требует значительных вычислительных ресурсов
Boosting	Устойчив к переобучению, применяется при работе с несбалансированными выборками данных	Имеет «склонность» к переобучению, обладает существенной вычислительной сложностью; при решении практических задач получаются сложные композиции, которые тяжело настраиваются
AdaBoost	Обладает устойчивостью к зашумленным и несбалансированным данным	Чувствителен к выбросам
Gradient Boosting	Более других моделей устойчив к ошибкам и пропускам данных	Крайне чувствителен к выбросам и при их наличии тратит огромное количество ресурсов
Stacking	Обладает устойчивостью к переобучению, позволяет создавать многоуровневые модели	Характеризуется значительным ростом вычислительной сложности при увеличении количества моделей
Hybrid Ensemble	Обладает относительно высокой точностью и универсальностью при обработке выборок с различными свойствами	Зависим от качества и количества данных; имеется сложность настройки моделей

■ ***Таблица 2.*** Характеристики основных подходов разделения последовательности
■ ***Table 2.*** Major approaches characteristics of sequence separation

Метод разделения последовательности	Характеристика
Скользящие окна	Сегмент увеличивается до тех пор, пока не превысит некоторую границу, далее процесс повторяется со следующей точкой данных, не включенной в новый аппроксимированный сегмент
Сверху вниз	Информационная последовательность рекурсивно разделяется до момента выполнения заранее заданных критериев остановки
Снизу вверх	Начиная с максимально возможного приближения, сегменты объединяются, пока не будут выполнены заранее заданные критерии остановки
Линейная интерполяция	По заранее определенным характеристикам формируются аппроксимирующие линии для последовательности и на их основе определяются сегменты
Неконтролируемый подход, основанный на глубоком обучении	Применяются методы машинного обучения для автоматического извлечения знаний из последовательностей и определения сегментов
Сегментация на основе шаблонов	Формируется «словарь» шаблонов, на основе которого происходит определение сегментов информационной последовательности
Сегментация на основе пороговых значений	Обрабатываются заранее заданные характеристики информационной последовательности, отслеживаются пороговые значения для определения границ сегментов
Сегментация на основе периодичности	Определяется периодичность в последовательности для определения границ сегментов

Характеристики предлагаемого метода

В результате применения разных методов разделения последовательности получаются сегменты, обладающие информацией с разными свойствами. Свойства сегмента зависят от количества объектов наблюдения, их распределений, трендов, периодичности. Достигаемые показатели качества модели обработки обусловлены свойствами информации на сегменте. Для одного способа разбиения последовательности на сегменты лучшие результаты покажут одни модели, при выборе другого метода разбиения или изменении его параметров — другие.

В классических подходах машинного обучения происходит «настройка» моделей обучения на свойства обучающей выборки.

В представляемом решении рассматривается обратная задача формирования сегментов последовательностей данных таким образом, чтобы их свойства соответствовали модели обработки.

Предлагается метод формирования информационных подпоследовательностей на основе анализа показателей качества моделей обработки, отличающийся от известных, осуществляющих настройку моделей машинного обучения на обрабатываемые данные, разделением последовательности на сегменты и выбором способа сегментирования так, чтобы свойства полученных в сегменте данных лучшим образом соответствовали модели обработки.

В отличие от классического подхода, когда модель настраивается на данные, в предлагаемом методе анализом и выбором сегментов последовательности данные формируются под модель.

Определение свойств информации последовательности на первоначальных этапах анализа может быть затруднительным, иметь высокую вычислительную сложность, а в случае явления «дрейфа концепта» являться малоэффективным. Поэтому для повышения показателей качества обработки необходимо определять не просто модель обработки, а модель совместно с методом разделения последовательности.

Назначение моделей на сегменты, полученные разными методами разделения последовательности, происходит на основе выбранного показателя качества. Выбор показателя качества зависит от решаемой задачи. В более сложных моделях обработки возможно применять алгоритмы адаптивного взвешивания с учетом вклада различных признаков данных в сегментах, например на основе значений Шепли.

В случае нахождения адекватной модели для свойств информации сегмента становится возможным повысить функционал качества обработки, а в перспективе — уменьшить количество примеров для обучения.

В результате применения метода вначале происходит выбор способа разделения последовательности с одновременным использованием заранее определенного показателя качества модели обработки, а уже как следующий шаг — формирование агрегированной модели, состоящей из алгоритмов, достигающих лучших качественных показателей на сегментах.

Постановка задачи обработки информационных потоков

Поступающие на вход моделей обработки выборки могут представлять собой данные с разнообразными структурами, имеющими неявные закономерности в распределениях, дисбаланс классов и частот появления событий, изменения диапазонов значений переменных под воздействием неопределенных факторов.

Формальную постановку можно представить следующим образом. X — выборка данных, состоящая из последовательностей, полученных в различных состояниях: $\{X^1, \ldots, X^l, \ldots, X^L\} \in X$, где X^l — последовательность значений, полученных в состоянии l.

Последовательность данных может быть разделена разными способами. Функции разбиения последовательности $\{\mu_1, \ldots, \mu_k\} \in \mu$ формируют подпоследовательности на основе различных алгоритмов и методов. Это могут быть решения, основанные на анализе трендов, сезонности, частотных составляющих, методах автокорреляции, скользящих окон и т. д. В зависимости от свойств данных и решаемых задач можно считать, что формирование сегментов последовательностей возможно $k = 1, \ldots, K$ способами.

Функция разбиения $\mu_k : X^l \to \{X_1^{lk}, \ldots, X_m^{lk}\}$ разделяет последовательность $X^l \in X$ в состоянии l на подпоследовательности.

В результате применения функций разбиения меняются состав и свойства подпоследовательностей $X_i^{lk} \in X^l \in X$, количество содержащихся в них объектов наблюдения. А это приводит к тому, что различным способам разделения последовательности будут соответствовать свои модели, достигающие лучших качественных показателей, из заранее предопределенного множества $\{a_1, \ldots, a_N\} \in a$.

Способы $\{\mu_1, \ldots, \mu_k\} \in \mu$ разбиения последовательности будут приводить к разным качественным показателям обработки алгоритмов. Кроме того, если выбирать модели $\{a_1, \ldots, a_n, \ldots, a_N\} \in a$, то результат обработки последовательности X, разделенной на подпоследовательности, зависит и от модели a_n, и от способа μ_k формирования сегмента.

Таким образом, для последовательности X необходимо осуществить выбор способа формирования подпоследовательности μ_k и модели a_n, где функционал качества стремится к максимальному значению:

$$Q(a_n, X, \mu_k) \to \max_{k,n}. \quad (1)$$

Возникает задача обработки информационной последовательности, в которой предлагаемый метод использует оценку показателей качества моделей обработки на сегментах, обладающих разными свойствами данных. Происходит выбор способа разделения и полученных сегментов, на которых модель может приобрести лучшие показатели качества.

Реализация предлагаемого метода

Реализация метода предполагает выполнение ряда шагов по формированию подпоследовательностей и их обработке. На рис. 1 представлен алгоритм действий для выбора способа разделения информационной последовательности и модели обработки, показывающей лучшие значения показателя качества.

Первоначально подготавливается обучающая выборка $\{X^1, \ldots, X^l, \ldots, X^L\} \in X$ последовательности данных для различных состояний. Определяются и формируются модели обработки $\{a_1, \ldots, a_n, \ldots, a_N\} \in a$ и задаются методы разделения $\{\mu_1, \ldots, \mu_k\} \in \mu$. Затем для каждой модели a_n и для каждого способа μ_k выполняются обучение и оценка функционала качества $Q(a_n, X, \mu_k)$. Полученные результаты для каждой модели a_n и для каждого способа μ_k сохраняются. После выполнения циклов перебора моделей и подпоследовательностей осуществляется выбор модели обработки и способа формирования сегментов, на которых получены лучшие значения показателя качества.

Решение задачи, определяемой выражением (1), предполагает большое количество повторяющихся рутинных операций формирования последовательностей данных для обучения, анализа результатов на различных этапах реализации и выбора модели обработки.

Формируемые подпоследовательности могут обладать различными характеристиками, на которых лучшие показатели качества могут демонстрировать разные модели.

Определение лучшего значения функционала качества в простейшем случае может решаться методом прямого перебора. В таком алгоритме максимальное число рассматриваемых подпоследовательностей последовательности равно M^2. Условно накладные расходы на усредненное время обучения и обработки для модели a_n на подпоследовательности можно оценить величиной p,

■ ***Рис. 1.*** Последовательность шагов метода
■ ***Fig. 1.*** The method sequence steps

общая вычислительная сложность алгоритма составляет $O(pNM^2)$. Количество моделей обработки N — обычно относительно небольшие величины. На рост сложности влияет число рассматриваемых подпоследовательностей M, что является существенным ограничением предлагаемого метода.

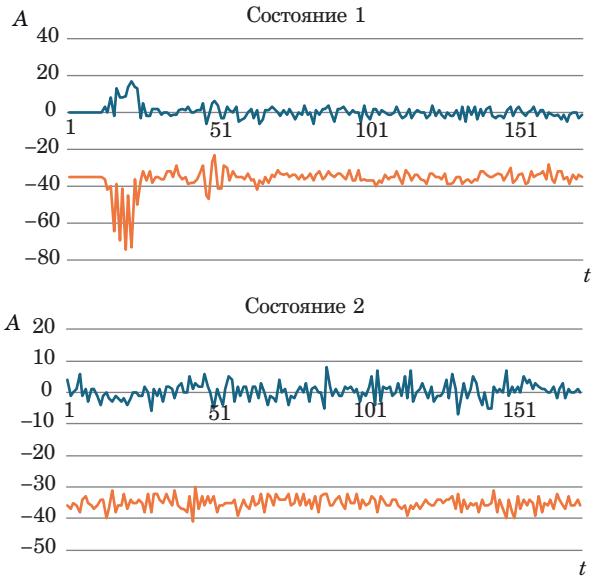
Однако в целях оптимизации и ускорения алгоритмов обработки можно применить ряд подходов, направленных на анализ и объединение «сходных» по свойствам подпоследовательностей, отбрасывание способов разбиения выборки при достижении значений качественных показателей обучения моделей заданного порога, распараллеливание процессов обучения для моделей.

Предлагаемый метод может быть применен в системах многоуровневой обработки [8, 25] для настройки моделей, использован в самообучающихся моделях. Реализация процессов самообучения и самонастройки предполагает ряд шагов. Поступающий на вход информационный поток подвергается обработке. Производится анализ и вычисление свойств объектов наблюдения. Выбирается заранее определенная модель, имеющая лучшие значения функционала качества. Результаты сравниваются с реальными значениями объектов наблюдения, полученными от регистрирующих систем и устройств. В случае увеличения ошибок выше заранее определенного порога принимается решение о формировании выборки данных. Над выборкой проводятся «манипуляции» по разделению на подпоследовательности, происходит настройка, обучение и назначение моделей на сегмент.

Экспериментальная оценка предлагаемого метода

Сравнение качественных показателей выполнялось для модельных данных и различных выборок [26, 27]. Пример анализируемой в эксперименте информационной последовательности представлен на рис. 2.

В приводимых в статье экспериментальных примерах рассматривается один из простых случаев, когда при обработке информационной последовательности формируются подпоследовательности путем выбора оптимальной длины и сдвига, а затем выполняются обучение моделей обработки и



■ ***Рис. 2.*** Пример информационной последовательности
■ ***Fig. 2.*** Information sequences example

анализ достигаемых ими качественных показателей. Выбор способа формирования подпоследовательности и модели происходит путем выявления лучших значений функционала качества.

Способ формирования и обработки подпоследовательностей для эксперимента представлен на рис. 3.

Для оценки качественных показателей обработки информационных последовательностей с применением описанного метода в качестве базовых моделей в зависимости от задачи обработки можно использовать алгоритмы любой сложности. В приводимом эксперименте при их выборе приоритет отдавался моделям, имеющим высокую скорость обучения. Сравнивались результаты наивного байесовского классификатора (NB), линейного дискриминанта (LD), машины опорных векторов (SVM), метода K-ближайших соседей (KNN), деревьев решений (DT) и ансамбля алгоритмов (ENS), включающего в себя все перечисленные методы. Цель эксперимента состояла в том, чтобы проанализировать качественные показатели моделей при формировании подпоследовательностей с различными свойствами, имеющих разные характеристики размера окна и сдвига. Оценка предлагаемого метода осуществлялась для задачи классификации.

Показатели качества обработки данных определялись метриками точности (Precision), полноты (Recall) и F-меры:

$$Pr = \frac{TP}{TP + FP} \times 100\%; \quad (2)$$

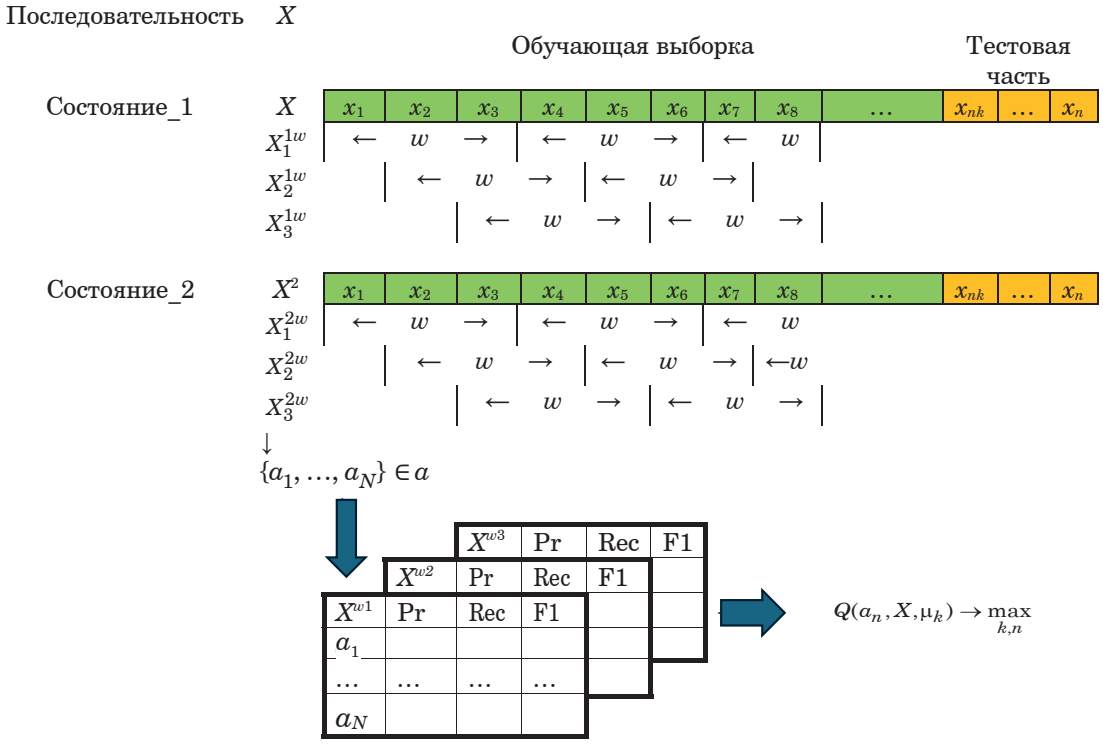
$$Rec = \frac{TP}{TP + FP} \times 100\%; \quad (3)$$

$$F1 = \frac{2 \times Pr \times Rec}{Pr + Rec} \times 100\%. \quad (4)$$

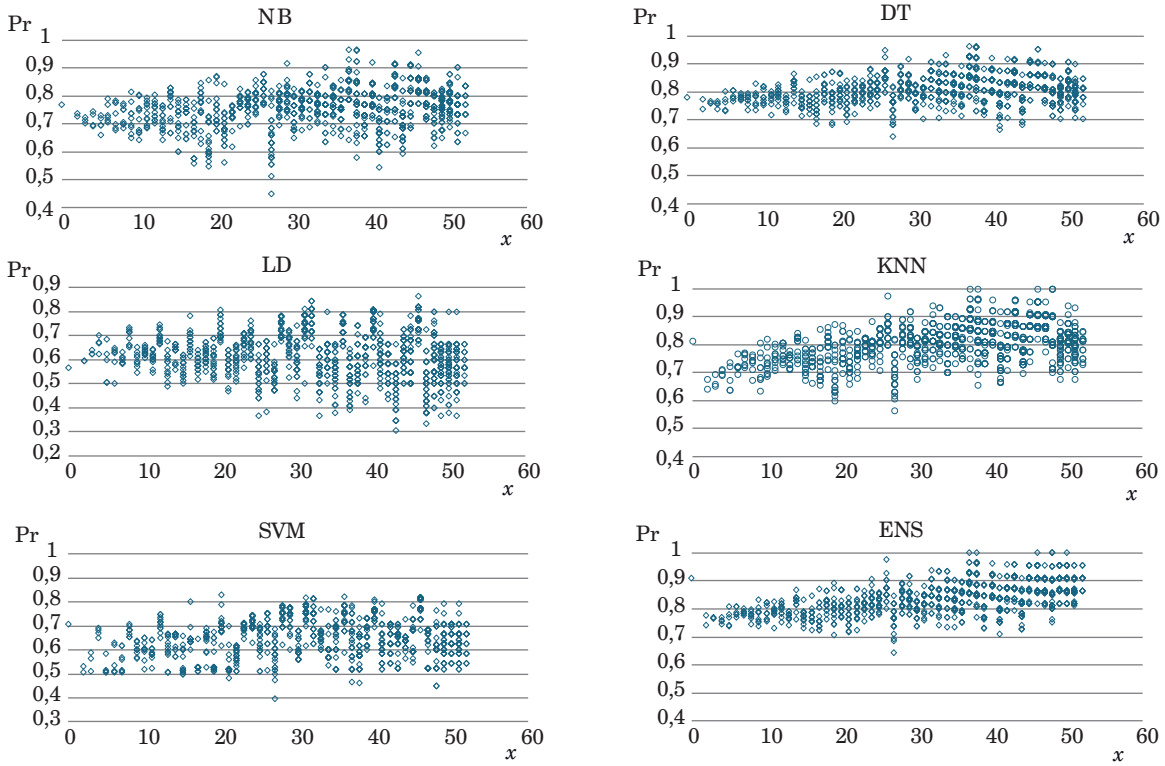
В эксперименте выборка последовательности каждого состояния разделялась в соотношении 70:30 на обучающую и тестовую, размер окна варьировался от двух до 50 отсчетов. На выбранных размерах окна выполнялись сдвиги, количество сдвигов соответствовало длине окна, т. е. для окна длиной два отсчета получалось два варианта разбиения, три отсчета — три варианта и т. д.

Диаграммы результатов зависимости, достигаемой алгоритмами точности (2), от длины подпоследовательностей, полученных в разных состояниях, и сдвига представлены на рис. 4. Каждый вариант разбиения длины окна дает свои значения точности, что отражает разброс по вертикали. Для фиксированной длины окна при различных сдвигах можно увидеть, что полученные значения точности (2) могут существенно расходиться.

Достигаемые качественные показатели зависят от длины окна. Если размер подпоследовательности слишком мал, то значимые для алгоритма характеристики могут быть не обнару-



■ ***Рис. 3.*** Формирование и обработка подпоследовательностей
■ ***Fig. 3.*** Subsequences formation and processing



■ ***Рис. 4.*** Результаты метрики точности Precision классифицирующих алгоритмов для сегментов с разными длинами x и сдвигами

■ ***Fig. 4.*** The results of the classifying algorithms Precision metric for segments with different lengths x and shifts

жены. Если размер слишком большой, то характеристики могут быть пропущены. Уменьшение или увеличение длины подпоследовательности целесообразно в определенных границах. После выхода за границы существенного прироста качественных показателей моделей обработки данных не происходит.

Зависимости на рис. 4 показывают, что на рассматриваемых данных для всех классифицирующих алгоритмов лучших результатов можно достичь при определенных сдвигах на длине окна от 25 до 48 объектов наблюдения. Затем происходит стабилизация результатов.

Оценка показателей качества всех вариантов длин и сдвигов дает возможность выбрать подходящий способ формирования сегментов.

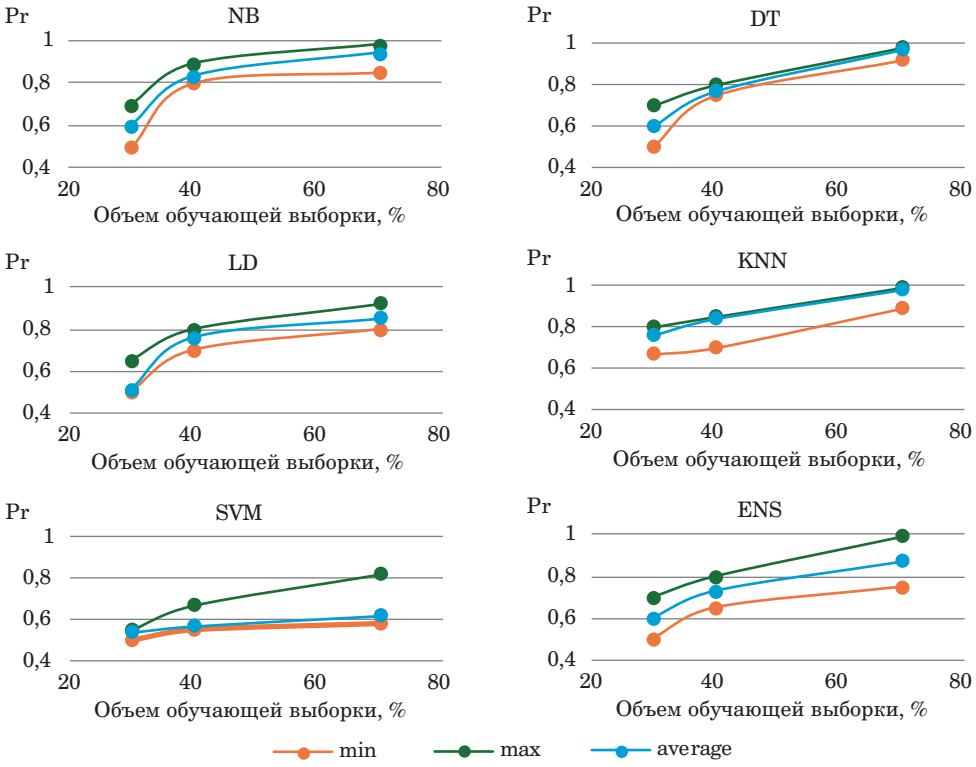
Кроме того, графики рис. 4 показывают, что различные алгоритмы обработки имеют разную чувствительность к изменению данных, связанных со сдвигом. Классифицирующие алгоритмы SVM, LD, NB демонстрируют большие разбросы значений качественных показателей в зависимости от способов формирования данных внутри окна определенной длины. Алгоритмы DT, KNN и ансамбль ENS имеют относительно небольшие разбросы по сравнению с первыми.

Разделение на подпоследовательности выборок данных может повысить качественные показатели обработки модели. Свойства получаемых разными способами подпоследовательностей могут различаться, в связи с чем при их обработке алгоритмы могут достигать разных значений показателей качества.

Далее в эксперименте произведена оценка влияния соотношения объема обучающей и тестовой частей выборки на достигаемые качественные показатели для разных способов формирования подпоследовательностей. Были выделены длины окон и сдвиги, определяющие максимальное, среднее и минимальное значение точности. Значения точности (2) в зависимости от процентного соотношения «промаркированной» информации (обучающей и тестовой выборки) приведены на рис. 5.

Результаты демонстрируют влияние свойств информации сформированных подпоследовательностей на достигаемые показатели классифицирующих алгоритмов. На экспериментальных данных для всех моделей при разных соотношениях обучающей и тестовой выборки можно выделить подпоследовательности, на которых будет наблюдаться более высокий качественный показатель по сравнению с другими.

Для длин окон и сдвигов, на которых получены максимальное, среднее и минимальное значение точности (2), увеличение обучающей



■ ***Рис. 5.*** Значения точности Precision в зависимости от объема обучающей выборки для подпоследовательностей при разных сдвигах

■ ***Fig. 5.*** Precision accuracy values depending on the size of the training sample for the subsequences at different shifts

выборки приводит к улучшению результата показателя качества. Однако изначально «неудачно» выбранная подпоследовательность приводит к ухудшению точности на 10 % в среднем для выбранных данных и моделей классификации.

Далее проверка достигаемых качественных показателей полноты (2), точности (3), F-меры (4) выполнялась на последовательностях других датасетов [25, 26]. Определялись подпоследовательности, полученные для различных длин окна и сдвигов, затем, аналогично графикам рис. 4, выбирались максимальные (max), средние (avr) и минимальные (min) значения показателя точности (2). Размер окна варьировался от двух до 50 объектов наблюдения.

Значения качественных показателей (2)–(4) приведены в табл. 3. Значения ранжировались относительно показателя (2). Представленные результаты показывают, что предложенный способ формирования и выбора подпоследовательностей оказывает влияние на результаты обработки последовательностей рассмотренными классифицирующими алгоритмами. Применяя описанный способ формирования подпоследовательностей, можно повысить качественный показатель относительно «средних» показателей на 10 %, а относительно «худших» — до 30 %.

Эксперимент показывает важность формирования обучающих примеров для повышения качества обработки данных. Необходимо анализировать возможность формирования подпоследовательностей, имеющих свойства данных, на которых классифицирующие алгоритмы могут достигать более высоких качественных показателей обработки. Применение метода автоматического формирования подпоследовательностей и последующий анализ показателей качества использующих их моделей дает возможность локализовывать группы объектов наблюдения и нивелировать влияние различных эффектов, связанных с шумовыми данными и выбросами.

Однако основным ограничением предложенного метода, как и большого количества методов машинного обучения, по-прежнему остается «проблема малых данных». При применении предлагаемого решения необходимо, чтобы данные повторяли свойства генеральной совокупности для формирования адекватной модели обработки.

Оценка потенциальных качественных показателей при обучении моделей на подпоследовательностях дает возможность определить алгоритм обработки, обладающий лучшими показателями функционала качества.

■ ***Таблица 3.*** Результаты классифицирующих алгоритмов
■ ***Table 3.*** Results of classifying algorithms

Модель	Значение	Drone RF dataset			PJM Hourly Energy Consumption Data			Electricity Load Diagrams 2011–2014		
		Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1
NB	min	54,7	73,0	62,5	35,0	74,4	47,6	50,7	23,6	32,2
	avr	75,8	76,2	76,0	65,9	74,4	69,9	68,1	39,2	49,8
	max	79,1	77,1	78,1	75,0	68,5	71,6	70,0	51,4	59,3
LD	min	63,7	75,9	69,3	58,9	67,7	63,0	70,5	53,6	60,9
	avr	77,4	76,1	76,7	80,0	68,8	74,0	68,4	54,7	60,8
	max	82,2	77,7	79,9	81,0	70,1	75,2	72,6	55,4	62,8
DT	min	61,5	70,4	65,6	39,1	67,3	49,5	45,5	26,9	33,8
	avr	70,9	71,9	71,4	64,0	41,6	50,4	56,9	37,7	45,4
	max	80,7	73,1	76,7	78,6	70,4	74,3	70,5	54,7	61,6
KNN	min	82,8	74,0	78,2	75,6	67,5	71,3	68,5	56,6	62,0
	avr	84,9	76,6	80,5	80,8	70,9	75,5	70,4	60,7	65,2
	max	91,4	78,6	84,5	82,5	72,0	76,9	72,1	59,7	65,3
SVM	min	49,7	68,1	57,5	46,6	61,5	53,0	71,1	22,3	34,0
	avr	77,1	81,7	79,3	67,7	76,3	71,7	53,9	43,7	48,3
	max	79,6	87,1	83,2	74,4	81,4	77,7	61,5	55,8	58,5
ENS	min	83,3	65,8	73,5	80,8	60,8	69,4	45,6	29,7	36,0
	avr	90,5	83,0	86,6	84,7	77,8	81,1	61,1	34,5	44,1
	max	95,4	80,2	87,2	92,3	72,0	80,9	73,1	57,2	64,2

Заключение

Анализ данных и выявление свойств выборок на предварительных этапах разработки моделей машинного обучения может оказывать существенное влияние на результат.

В статье предложен метод формирования подпоследовательностей, использующий анализ показателей качества моделей обработки. В отличие от известных методов, осуществляющих настройку моделей машинного обучения на обрабатываемые данные, в предлагаемом решении выбор способа разделения последовательности на подпоследовательности осуществляется с таким, чтобы свойства полученных в подпоследовательности данных лучшим образом соответствовали модели обработки. Происходит не только настройка модели на выборку данных, а формируются и выбираются подпоследовательности со свойствами данных под модель обработки.

Основное преимущество предлагаемого метода состоит в возможности выбора подпоследовательностей на основе анализа функционала качества модели, что дает возможность определить способы разделения на сегменты, позволяющие повысить показатели качества обработки.

В рассмотренном методе предлагается осуществлять формирование подпоследовательностей из выборки данных, а далее тестировать на них алгоритмы машинного обучения и модели обработки. Критерием выбора является достижение лучшего показателя качества при переборе всех имеющихся вариантов, что дает возможность выбрать лучшую модель.

Предложенный метод позволяет частично преодолевать такие проблемы, как недостаточное количество меток в «малых выборках данных». В результате его применения становится возможным обучать модель с использованием данных, которые маркированы частично или неточно, а также снижать затраты на процесс маркировки.

Предложенный метод в зависимости от решаемой задачи обработки данных методами машинного обучения может применяться как отдельно, так и совместно с другими методами.

Финансовая поддержка

Исследование выполнено за счет гранта Российского научного фонда № 25-21-00269, https://rscf.ru/project/25-21-00269/.

Литература

1. **Safonova A., Ghazaryan G., Stiller S., Main-Knorn M., Nendel C., Ryo M.** Ten deep learning techniques to address small data problems with remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 2023, vol. 125, Article 103569. doi:10.1016/j.jag.2023.103569
2. **Lauriola I., Lavelli A., Aiolli F.** An introduction to deep learning in Natural Language Processing: Models, techniques, and tools. *Neurocomputing*, 2022, vol. 470, pp. 443–456. doi:10.1016/j.neucom.2021.05.103
3. **AL-Gburi A., Nazri M., Yaakub M., Alyasseri Z.** Multi-objective unsupervised feature selection and cluster based on symbiotic organism search. *Algorithms*, 2024, vol. 17, no. 8, Article 355. doi:10.3390/a17080355
4. **Bennaceur H., Almutairy M., Alhussain N.** Genetic algorithm combined with the K-Means algorithm: A hybrid technique for unsupervised feature selection. *Intelligent Automation and Soft Computing*, 2023, vol. 37, pp. 2687–2706.
5. **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.
6. **Rinaldo A., Wang D., Wen Q., Willett R.** Localizing changes in highdimensional regression models. *The 24th International Conference on Artificial Intelligence and Statistics*, April 13–15, 2021, pp. 2089–2097.
7. **Huang J., Chen P., Lu L., Deng Y., Zou Q.** WCDForest: A weighted cascade deep forest model toward the classification tasks. *Applied Intelligence*, 2023, vol. 53, no. 23, pp. 1–14.
8. **Ammar M., Rania K.** A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University — Computer and Information Sciences*, 2023, vol. 35, iss. 2, pp. 757–774. doi:10.1016/j.jksuci.2023.01.014
9. **Lebedev I. S., Sukhoparov M. E.** Improving the quality indicators of multilevel data sampling processing models based on unsupervised clustering. *Emerging Science Journal*, 2024, vol. 8, no. 1, pp. 355–371. doi:10.28991/ESJ-2024-08-01-025
10. **Rahul P., Robert D. Nowak.** Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 2021, vol. 22, pp. 1–40.
11. **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. 1908–1930. https://doi.org/10.3390/f13111908
12. **Li Y., Guo X., Lin W., Zhong M., Li Q., Liu Z., Zhong W., Zhu Z.** Learning dynamic user interest sequence in knowledge graphs for click-through rate prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2023, vol. 35, no. 1, pp. 647–657. doi:10.1109/TKDE.2021.3073717
13. **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. doi:10.1109/TKDE.2021.3138962
14. **Qiao Q., Yunusa-Kaltungo A., Rodger E.** Feature selection strategy for machine learning methods in building energy consumption prediction. *Energy Reports*, 2022, vol. 8, pp. 13621–13654. https://doi.org/10.1016/j.egyr.2022.10.125
15. **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
16. **Wang P., Xue B., Liang J., Zhang M.** Feature clustering-Assisted feature selection with differential evolution. *Pattern Recognition*, 2023, vol. 140, Article 109523. doi:10.1016/j.patcog.2023.109523
17. **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, т. 20, № 4, с. 820–844. doi:10.15622/ia.20.4.3, EDN: YNFARR
18. **Jin H., Yin G., Yuan B., Jiang F.** Bayesian hierarchical model for change point detection in multivariate sequences. *Technometrics*, 2022, vol. 64, no. 2, pp. 177–186.
19. **Nevendra M., Singh P**. Software defect prediction using deep learning. *Acta Polytechnica Hungarica*, 2021, vol. 18, no. 10, pp. 173–189.
20. **Tallman E., West M**. Bayesian predictive decision synthesis. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 2024, vol. 86, iss. 2, pp. 340–363. https://doi.org/10.1093/jrsssb/qkad109
21. **Xiao Z., Xing H., Zhao B., Qu R.** Deep contrastive representation learning with self-distillation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024, vol. 8, no. 1, pp. 3–15. doi:10.1109/TETCI.2023.3304948
22. **Xiao Z., Xing H., Qu R., Feng L., Luo S., Dai P., Zhao B., Dai Y.** Densely knowledge-aware network for multivariate time series classification. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, vol. 54, no. 4, pp. 2192–2204. doi:10.1109/TSMC.2023.3342640
23. **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, Article 1954. doi:10.3390/s21061954
24. **Лебедев И. С.** Адаптивное построение регрессионных моделей на основе анализа функционала качества обработки сегментов последовательно-

сти. *Информатика и автоматизация*, 2025, т. 24, № 2, с. 363–394. doi:10.15622/ia.24.2

25. **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. doi:10.28991/ESJ-2023-07-03-03
26. **Allahham M., Al-Sa'd M. F., Al-Ali A., Amr M., Khattab T., Erbad A.** DroneRF dataset: A dataset of drones for RF-based detection, classification and identification. *Data in Brief*, 2019, vol. 26, Article 104313. doi:10.1016/j.dib.2019.104313
27. **Trindade A.** Electricity Load Diagrams 2011–2014 [Dataset]. *UCI Machine Learning Repository*, 2015. doi:10.24432/C58C86. https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014 (дата обращения: 05.08.2024).

UDC 621.396
doi:10.31799/1684-8853-2025-3-25-36
EDN: SSGKZU

Sequential information processing using adaptive pattern analysis in assessing the state of systems

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: The formation and labeling of samples is a labor-intensive procedure that plays an important role in the processes of training and tuning many machine learning models. While solving problems using artificial intelligence methods, there is a need to reduce the data labeling complexity. **Purpose:** To improve the quality of information sequence processing using the formation, analysis and determination of sequential segments, with the best quality indicators achieved by predetermined machine learning algorithms. **Results:** We propose a method for forming information sequence segments based on the analysis of processing model quality indicators. It differs from the existing methods that tune machine learning models to process data, by dividing the sequence into segments and choosing a segmentation method so that the properties of the data obtained in the segment could best match the processing model. In contrast to the classical approach where the model is tuned to the data, in the proposed method the data is adjusted to the model by segmenting the sequence. **Practical relevance:** The results can be used for the models and methods which solve the problems of information sequence classification and forecasting. The proposed method makes it possible to partially overcome several problematic issues by labeling small data samples. As a result, it becomes possible to train models using the data that is partially or inaccurately labeled, and to reduce the labeling process complexity. The further development of the proposed solution is possible based on ensemble methods.

Keywords — machine learning, adaptive models, quality improvement, adaptive sequence processing.

Financial support
This research is supported by Russian Science Foundation, grant No. 25-21-00269, https://rscf.ru/project/25-21-00269/.

For citation: Lebedev I. S. Sequential information processing using adaptive pattern analysis in assessing the state of systems. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2025, no. 3, pp. 25–36 (In Russian). doi:10.31799/1684-8853-2025-3-25-36, EDN: SSGKZU

Reference

1. Safonova A., Ghazaryan G., Stiller S., Main-Knorn M., Nendel C., Ryo M. Ten deep learning techniques to address small data problems with remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 2023, vol. 125, Article 103569. doi:10.1016/j.jag.2023.103569
2. Lauriola I., Lavelli A., Aiolli F. An introduction to deep learning in Natural Language Processing: Models, techniques, and tools. *Neurocomputing*, 2022, vol. 470, pp. 443–456. doi:10.1016/j.neucom.2021.05.103
3. AL-Gburi A., Nazri M., Yaakub M., Alyasseri Z. Multi-objective unsupervised feature selection and cluster based on symbiotic organism search. *Algorithms*, 2024, vol. 17, no. 8, Article 355. doi:10.3390/a17080355
4. Bennaceur H., Almutairy M., Alhussain N. Genetic algorithm combined with the K-Means algorithm: A hybrid technique for unsupervised feature selection. *Intelligent Automation and Soft Computing*, 2023, vol. 37, pp. 2687–2706.
5. 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.
6. Rinaldo A., Wang D., Wen Q., Willett R. Localizing changes in highdimensional regression models. *The 24th International Conference on Artificial Intelligence and Statistics*, April 13–15, 2021, pp. 2089–2097.
7. Huang J., Chen P., Lu L., Deng Y., Zou Q. WCDForest: A weighted cascade deep forest model toward the classification tasks. *Applied Intelligence*, 2023, vol. 53, no. 23, pp. 1–14.
8. Ammar M., Rania K. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University – Computer and Information Sciences*, 2023, vol. 35, iss. 2, pp. 757–774. doi:10.1016/j.jksuci.2023.01.014
9. Lebedev I. S., Sukhoparov M. E. Improving the quality indicators of multilevel data sampling processing models based on unsupervised clustering. *Emerging Science Journal*, 2024, vol. 8, no. 1, pp. 355–371. doi:10.28991/ESJ-2024-08-01-025
10. Rahul P., Robert D. Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 2021, vol. 22, pp. 1–40.
11. 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. 1908–1930. https://doi.org/10.3390/f13111908
12. Li Y., Guo X., Lin W., Zhong M., Li Q., Liu Z., Zhong W., Zhu Z. Learning dynamic user interest sequence in knowledge graphs for click-through rate prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2023, vol. 35, no. 1, pp. 647–657. doi:10.1109/TKDE.2021.3073717
13. 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. doi:10.1109/TKDE.2021.3138962
14. Qiao Q., Yunusa-Kaltungo A., Rodger E. Feature selection strategy for machine learning methods in building energy

consumption prediction. *Energy Reports*, 2022, vol. 8, pp. 13621–13654. https://doi.org/10.1016/j.egyr.2022.10.125
15. 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
16. Wang P., Xue B., Liang J., Zhang M. Feature clustering-Assisted feature selection with differential evolution. *Pattern Recognition*, 2023, vol. 140, Article 109523. doi:10.1016/j.patcog.2023.109523
17. Zegzhda D., Kalinin M., Krundyshev V., Lavrova D., Moskvin D., Pavlenko E. Application of bioinformatics algorithms for polymorphic cyberattacks detection. *Informatics and Automation* (SPIIRAS Proc.), 2021, vol. 20, no. 4, pp. 820–844 (In Russian). doi:10.15622/ia.20.4.3, EDN: YNFARR
18. Jin H., Yin G., Yuan B., Jiang F. Bayesian hierarchical model for change point detection in multivariate sequences. *Technometrics*, 2022, vol. 64, no. 2, pp. 177–186.
19. Nevendra M., Singh P. Software defect prediction using deep learning. *Acta Polytechnica Hungarica*, 2021, vol. 18, no. 10, pp. 173–189.
20. Tallman E., West M. Bayesian predictive decision synthesis. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 2024, vol. 86, iss. 2, pp. 340–363. https://doi.org/10.1093/jrsssb/qkad109
21. Xiao Z., Xing H., Zhao B., Qu R. Deep contrastive representation learning with self-distillation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024, vol. 8, no. 1, pp. 3–15. doi:10.1109/TETCI.2023.3304948
22. Xiao Z., Xing H., Qu R., Feng L., Luo S., Dai P., Zhao B., Dai Y. Densely knowledge-aware network for multivariate time series classification. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, vol. 54, no. 4, pp. 2192–2204. doi:10.1109/TSMC.2023.3342640
23. 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, Article 1954. doi:10.3390/s21061954
24. Lebedev I. S. Adaptive regression model construction based on the functional quality analysis of the sequence segment processing. *Informatics and Automation*, 2025, vol. 24, no. 2, pp. 363–394 (In Russian). doi:10.15622/ia.24.2
25. 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.
26. Allahham M., Al-Sa'd M. F., Al-Ali A., Amr M., Khattab T., Erbad A. DroneRF dataset: A dataset of drones for RF-based detection, classification and identification. *Data in Brief*, 2019, vol. 26, Article 104313. doi:10.1016/j.dib.2019.104313
27. Trindade A. Electricity Load Diagrams 2011-2014 [Dataset]. *UCI Machine Learning Repository*, 2015. doi:10.24432/C58C86. Available at: https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014 (accessed 5 August 2024).