УДК 004.056 **DOI** 10.15622/ia.25.3.1

И.С. Лебедев

АДАПТИВНОЕ ФОРМИРОВАНИЕ ВЫБОРОК ДАННЫХ ДЛЯ САМООРГАНИЗУЮЩИХСЯ СИСТЕМ КОНТРАСТНОГО ОБУЧЕНИЯ

Лебедев И.С. **Адаптивное формирование выборок данных для самоорганизующихся систем контрастного обучения.**

Аннотация. Применение самообучения и самоорганизующихся моделей для задач классификации и прогнозирования при обработке временных рядов и информационных последовательностей сталкивается с рядом проблемных вопросов организации данных. Повышение показателей качества обработки таких систем вызывает необходимость совершенствования методов выбора объектов наблюдения для обучающих выборок данных. В статье предложен метод формирования и анализа выборок данных на основе определения патчей информационной последовательности, обладающих разными характеристиками длины и сдвига, отличающаяся использованием функционала качества моделей обработки. На основе эксперимента на модельных данных и выборках проведена оценка предлагаемого метода. Получены значения показателя качества accuracy для разных алгоритмов обработки при различных длинах и сдвигах последовательностей патча. Определены свойства полученных патчей с использованием метрик коэффициента силуэта и расстояния между центройдами. Проанализированы ошибки классифицирующих алгоритмов. Выделены доверительные интервалы ошибок. Определено, что на изменение длины и сдвига патча влияет на достигаемые значения accuracy классифицирующих алгоритмов. Предлагаемый метод дает возможность повысить показатель accuracy за счет выбора длины и сдвига при формировании патча и назначения моделей, которые имеют наилучшие показатели. Результаты показывают возможность увеличения на 6-10% для слабых моделей, в то время как для сильных моделей наблюдается улучшение на 1-5% в сценариях с ограниченными данными. Анализ предложенного решения показывает, что варьирование параметрами сдвига и длины последовательности при формирования обучающих выборок данных оказывает влияние на эффективность обработки данных.

Ключевые слова: машинное обучение, адаптивные модели, повышение качества обработки, формирование патчей.

1. Введение. Временные ряды и информационные последовательности играют важную роль в различных областях. Анализ финансовых рынков, состояния информационных систем, климатических изменений, генерации энергии, различных видов деятельности человека обуславливает необходимость разработки эффективных моделей и методов обработки, предполагающих решение задач классификации и прогнозирования.

В настоящее время одно из интенсивно развивающихся направлений обработки временных рядов и информационных последовательностей является применение различных методов самообучения для задач классификации [1 – 3] и прогнозирования [4 – 6].

Объемы накопленной информации настолько велики, что их разметка с помощью экспертов затруднительна.

Подходы к обработке на основе сверточных [7, 8], рекуррентных сетей [8, 9], трансформерных технологий [10 – 12] доказали свою эффективность при решении ряда прикладных задач. Ряд исследований показывает, что такие модели могут захватывать пространственную и временную информацию последовательностей. Они часто превосходят традиционные методы Dynamic Time Warping, Bag of Stochastic Frontier Analysis Symbols и Collective of Transformation-Based Ensembles [13]. Однако формирование обучающих выборок по-прежнему является одним из проблемных вопросов во многих практических приложениях, от которого зависит повышение качественных показателей. Это связано с рядом архитектурных ограничений моделей обработки.

Традиционные методы и модели глубокого обучения направлены на анализ наиболее значимой информации для классификации или прогнозирования. В большинстве случаев в задачах извлечения знаний последовательности данных ограничиваются подряд идущими последовательностями. Например, CNN при обработке временных рядов обычно оптимизируются для захвата шаблонов только в пределах ограниченного размера окна [14, 15]. Если последовательности, содержащие значимые информационные значения, имеют длительные временные интервалы, при обработке может наблюдаться понижение качественных показателей. Расширенные сверточные нейронные сети (Extended Convolutional Neural Networks, E-CNN) частично решают эту проблему за счет скользящих фильтров. Однако они по-прежнему ограничены своей внутренней архитектурой и зависят от количества слоев, которое может быть недостаточно большим для полного захвата долгосрочных зависимостей и приводить к исчезновению градиентов.

Аналогично, эффективность трансформеров [16] в значительной степени определяется множеством факторов, таких как длина последовательности, свойства обрабатываемых данных, особенностей разметки и определения токенов. Помимо этого, при создании на их основе моделей и методов обработки приходится решать ряд вопросов архитектурных ограничений, сложности оптимизации и настройки, снижения высоких вычислительных затрат при ограниченности ресурсов.

Подобными недостатками обладает большинство современных, недавно появившихся моделей, ориентированных на обработку временных рядов и информационных последовательностей. Например, Tiny Time Mixers (TTM) чувствительна к длине последовательностей [17]. В ряде задач это обуславливает необходимость обучения совокупности моделей для разных настроек длины. Предобученная нейросеть Moment при дообучении

небольшими наборами будет работать только для примеров в обучающих данных, но может не показать положительного результата в обобщении [18].

Модель Moirai для прогнозирования временных рядов имеет недостатки, связанные с ограничениями внутренней архитектуры, не позволяющими выявить частоту появления паттернов для «малых данных», что влияет на качество результата [19].

Производительность и внедрение существующих моделей и методов глубокого обучения, применяемых в анализе временных рядов и информационных последовательностей, ограничены доступностью хорошо аннотированных меток. Несмотря на то, что ежедневно появляется значительное количество новых данных, собираемых устройствами, датчиками, сенсорами, процесс маркировки последовательностей значительно отстаёт от их генерации. Ручная маркировка требует наличия экспертов со знанием предметной области и опытом, которые могут обработать только небольшую часть набора. Более того, в некоторых случаях, когда задействовано несколько экспертов, возникающие события в последовательностях объектов наблюдения часто сложно аннотировать из-за разногласий между ними. В совокупности эти проблемы приводят к дефициту размеченных объектов наблюдения, что является существенным препятствием для применения глубокого обучения в этой области.

Таким образом, возникает задача автоматического формирования обучающих выборок, где структура данных и свойства учитывает внутреннюю архитектуру, длину входящих последовательностей, и позволяет увеличить количество эталонных и уменьшить число шумовых примеров.

2. Модели и методы обработки последовательностей. Временные ряды имеют существенное значение для множества сфер деятельности и активно привлекают внимание научного сообщества. Это обуславливает непрерывное развитие и оптимизацию разнообразных моделей и методов глубокого обучения, применяемых в задачах классификации и прогнозирования.

В современной практике анализа и обработки данных широко используются свёрточные нейронные сети, трансформеры, а также предобученные модели. Однако одна из проблем эффективного применения классических моделей и методов машинного обучения состоит в том, что объем данных растет быстрее, чем возможности по разметке. В связи с этим, автоматическое формирование и повышение качества обучающих выборок данных является

направлением, позволяющим улучшить качество обработки информационных последовательностей.

Сегодня процесс создания размеченных данных и структур тесно связан с архитектурой используемых моделей обработки. В частности, эффективность свёрточных нейронных сетей напрямую определяется их способностью выявлять значимые локальные паттерны в данных [20]. Формирование обучающей выборки из содержащих их примеров позволяет достигать значительного качества обработки изображений, аудио и видео последовательностей. Тем не менее одним из проблемных вопросов является определение и выбор паттернов, позволяющих повысить качественные показатели обработки. Для его решения разрабатываются различные способы, чаще всего основанные на свертке. Например, в методах DSN [21] применяются динамические разреженные соединения между нейронами, чтобы определить значащие паттерны на различных уровнях для временных рядов. Другой подход, основанный на свертке, SciNet [22] фиксирует характеристики, разбивая каждую последовательность на подпоследовательности, позволяющие анализировать временную динамику данных.

Второе направление связано с использованием трансформеров. Изначально эти модели были разработаны для обработки естественного языка (NLP), но нашли применение и в обработке временных рядов и информационных последовательностей. Достижение качественного результата таких моделей зависит от выбора поступающих на вход подпоследовательностей, влияющих на эффективность применения «механизма внимания», определяющего наиболее значимые элементы.

Для этого используются разные решения, внутренние механизмы, направленные на определение свойств обрабатываемых последовательностей. Informer [23] улучшает возможности простого трансформера при работе с длинными входными последовательностями за счет анализа саморегуляции с использованием механизма самовнимания ProbSparse. В Autoformer [24] внедрен декомпозиционный подход и механизм автокорреляции, что позволяет более эффективно работать с длинными последовательностями данных и достигать более высокой точности прогнозирования. ContiFormer [25] использует непрерывное временное представление и модифицированный механизм внимания, что позволяет более эффективно работать с данными, собранными в произвольные моменты времени. PatchTST [26] улучшает прогнозирование временных рядов за счет применения подхода к сегментации данных и механизма самообучения, что позволяет более эффективно обрабатывать

и анализировать временные последовательности, сохраняя при этом глобальный контекст.

Описанные выше модели требуют больших объёмов данных для эффективного обучения (от сотен тысяч до миллиардов токенов). Это связано с большим числом параметров, необходимостью улавливать дальние зависимости в последовательностях, потребностью в обобщении на разнообразных примерах. Они имеют ограничения к выборкам. Для их обучения важна чистота и релевантность данных, что обуславливает необходимость фильтрации шума и некорректных примеров, удаление дубликатов, балансировка классов (для задач классификации). Точность трансформеров варьируется в зависимости от задачи, архитектуры модели, объёма и качества данных, а также других факторов. В целом считается, что трансформеры демонстрируют высокие показатели в различных областях от 91% и выше.

Другое направление определяется применением предварительно обученных моделей для захвата сложных временных паттернов. Например, решения на основе модели Moment [18], обученной на «глобальном датасете» временных рядов Time-series Pile, требуется только предварительная настройка на анализируемую выборку. Однако её качество при анализе временного ряда зависит от определения оптимального размера окна и учета временного сдвига последовательности.

Развиваются контрастные методы анализа временных рядов, использующие сопоставление данных для выявления закономерностей. В [27] представлен подход SoftCLT, который использует расстояния между выборками временных рядов. Он направлен на обнаружение корреляции между соседними последовательностями.

Для устранения дефицита данных применяется контрастное обучение [28]. Оно содержит шаги по расширению выборок временных рядов для создания объектов наблюдений разных классов, анализу расширенных выборок и оптимизацию модели обработки на основе полученных значений функций потерь. Самостоятельное контрастное обучение не требует меток образцов на этапе обучения модели [29 – 31]. Развитие контрастного обучения для временных рядов требует решения ряда проблемных вопросов формирования структур, позволяющих осуществлять эффективную обработку данных.

Несмотря на появление сильных моделей, использование в их составе различных механизмов анализа временных рядов, формирование и выбор последовательностей и паттернов, поступающих

на вход в процессе обучения, настройки и обработки по-прежнему остается сложной задачей, от которой зависят качественные показатели.

3. Предлагаемый метод.

3.1. Постановка задачи. В настоящее время применение методов интеллектуального анализа информационных последовательностей имеет ограничение, связанное с качеством разметки данных. Различные системы в ходе функционирования требуют оценки протекающих процессов. В связи с этим, возникает необходимость решения задачи классификации состояния системы, отнесения её текущего состояния к одному из заранее определённых классов на основе её наблюдаемых параметров и характеристик. Для её решения предлагается метод выделения репрезентативных фрагментов (патчей, паттернов, сегментов) из исходных информационных последовательностей и временных рядов, обеспечивающих эффективную адаптацию и самообучение интеллектуальных систем мониторинга состояния при соблюдении заданных показателей качества.

В предлагаемом методе в основе автоматической разметки лежит разделение последовательностей на патчи. Под «патчем» понимается подпоследовательность, которая используется в анализе временных рядов для обработки данных. Размер патча влияет на точность прогнозирования и классификации в моделях на основе машинного обучения. Несмотря на присутствие в трансформерах и моделях на основе нейронных сетей различных встроенных механизмов патчинга, разбиение временного ряда и формирование сегментов до их входа может оказывать положительное влияние на процессы обучения и качественные показатели обработки.

Формирование сегментов способно обеспечивать воздействие на вычисление свойств паттернов и захват «значащей информации» моделями обработки, сокращать длину последовательности, уменьшать вычислительную сложность. В случае наличия периодичности в последовательностях данных патчи дают возможность модели изучать более длинные временные периоды, что потенциально повышает качество прогнозов. Однако выбор размера патча зависит от задачи и особенностей данных.

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.

Формализованная постановка задачи может быть представлена следующим образом.

Пусть X – обучающая последовательность данных, состоящая из объектов наблюдения $x \in X$. Подпоследовательности $\{X^{c_1}, \ldots, X^{c_k}\} \in X$ получены в разных состояниях $\{c_1, \ldots, c_k\} \in C$.Подпоследовательности разделяются на патчи функцией μ . В результате X преобразуется в последовательность $X^{\mu} = \{X_1^{\mu_{c_1}}, \ldots, X_m^{\mu_{c_k}}\}$, состоящую из m патчей, где каждому патчу ставиться в соответствие состояние c_i . В процессе обработки выборки X^{μ} могут применяться процессы аугментации, выполняться сдвиг данных, изменяться длина патча. Пусть φ способ формирования обучающего подмножества X^{μ} . Разделяя последовательность X^{μ} на обучающую и тестовую часть, формируется обучающая выборка X_{φ}^{μ}.

Тогда возникает необходимость сформировать такое обучающее подмножество, которое позволило бы повысить функционал качества модели обработки последовательности a.

$$Q(a(x, X_{\varphi}^{\mu}), X) \to \max_{\mu,\varphi}. \quad (1)$$

Разделение информационной последовательности и временных рядов на патчи, применение различных способов формирования обучающего подмножества в рассматриваемом подходе (1) дает возможность автоматического выполнения процессов аугментации для систем машинного обучения. В результате образуется искусственный обучающий набор данных, где патчи повторяют свойства элементов подпоследовательности временного ряда. Правильно подобранные патчи по длине, информационным свойствам, содержащихся в них последовательностей данных дают возможность увеличить объём тренировочного набора, повысить устойчивости моделей к вариациям входных последовательностей. Дальнейшее применение классических технических приемов, например, кросс-валидации может служить для повышения обобщающей способности алгоритмов. Формирование выборки патчей дает возможность выполнять балансировку неравномерно представленных объектов наблюдения классов. Такой подход направлен на автоматическое создание выборки данных для систем самоконтролируемого обучения. Последовательность сформированных паттернов обучающей выборки с меткой состояния, в котором они были получены, подается на вход модели обработки. Автоматическая нарезка патчей, которые рассматриваются как объекты наблюдения, и проведение операций по выбору и анализу

способов формирования объектов наблюдения, позволяют анализировать результаты обучения моделей на предмет достижения лучших качественных показателей обработки.

Новизна предлагаемого метода заключается в комплексном подходе, объединяющем автоматическое формирование объектов обучения из временных рядов с адаптивным изменением длины и сдвига подпоследовательностей и анализ функции качества моделей для оценки эффективности обработки. В отличие от традиционных подходов, рассматривается обратная задача. Не модель настраивается и оптимизируется на выборке данных, а наоборот, объекты наблюдения формируются из данных «под модель». Достигаемый показатель качества моделей обработки позволяет выбирать обучающие выборки под конкретные алгоритмы, определять объекты наблюдения с информативными признаками для каждой модели и снижать влияние шумовых элементов.

Предлагаемый подход ориентирован на модели, не требующие значительных вычислительных ресурсов, и обладает относительно простой реализацией. В отличие от PatchTST, где длина подпоследовательности задаётся фиксированным параметром, в предлагаемом методе она определяется динамически на основе функционала качества модели. В Segment/Shuffle (S3) сегменты перемешиваются в оптимальном для задачи порядке, что может влиять на сохранение временных зависимостей, искажать значимые паттерны, зависеть от гипперпараметров.

3.2. Метод формирования и анализа выборок данных. В процессе применения самоконтролируемого обучения, когда на вход подаются последовательности разных классов состояний, возникает необходимость определения образцов для обучения. Большое количество моделей обработки используют поступающие последовательности одинаковой длины. Для их формирования применяется разбиение последовательности на сегменты (патчи). Для этого могут использоваться различные эвристики, отражающие особенности предметной области полученных последовательностей. Предлагаемый метод отличается автоматическим формированием и выбором патчей, на основе анализа качественных показателей моделей обработки, что позволяет осуществлять выбор моделей и способов формирования обучающих последовательностей, где достигаются лучшие значения качественных показателей. Это дает возможность определять локальные закономерности групп объектов наблюдения, настроить длину входной последовательности для

модели, обнаруживать значащие паттерны. Предлагаемый метод предполагает ряд шагов.

Шаг 1. На вход подаются исходные информационные последовательности, полученные в разных состояниях системы.

Шаг 2. Для разделения на патчи фиксированной длины определяется шаг разбиения. В результате образуется обучающая выборка, содержащая набор патчей одинакового размера, помеченных соответствующим классом состояния.

Шаг 3. Обучающая выборка подается на вход заранее выбранных моделей. Происходит формирование обучающего множества, определение совпадающих элементов, применение процессов кросс-валидации.

Шаг 4. Определяется модель, достигающая лучших качественных показателей.

Шаг 5. Далее для каждой длины происходит сдвижка на один объект наблюдения. В результате для каждой длины L формируется L выборок патчей. Осуществляется формирование выборки. Выполняется переход на шаг 3.

Шаг 6. Затем длина патча увеличивается для неё повторяются процессы формирования патчей и сдвигов. Осуществляется переход на шаг 2.

Шаг 7. В конце происходит выбор лучших результатов по значению показателей качества и выбор длины и сдвига для формирования патча.

Шаг 8. Определяются лучшие по значения качественных показателей модели.

На рисунке 1 представлена алгоритмическая последовательность шагов предлагаемого метода. Предлагаемый метод может быть использован в самообучающихся моделях. Поступающий на вход информационный поток подвергается разделению. Полученные патчи рассматриваются как отдельные независимые объекты наблюдений. Например, для задач классификации они могут рассматриваться как многомерные вектора, для прогнозирования – последовательностями, характеризующими элементы ряда. Из них формируется обучающая выборка для последующих процессов обучения и верификации.

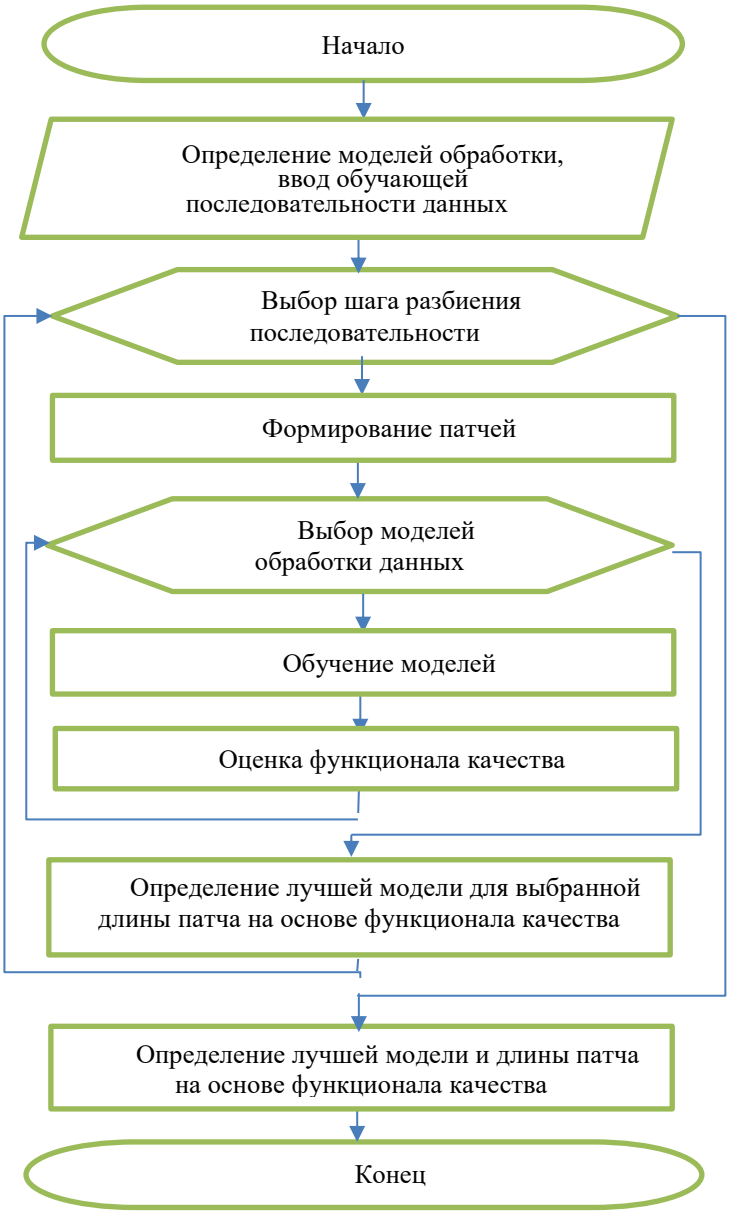


Рис. 1. Алгоритмическая последовательность шагов метода формирования и анализа выборок данных

Формирование патчей производится в автоматическом режиме. При этом возможно рассматривать различные способы расширения. Например, формирования с перекрытием, объединяя в одну выборку все полученные для последовательности одной длины выборки, так и без него. Изменение длины патча зависит от ограниченности вычислительных ресурсов. Для снижения вероятности пропуска значимых свойств информации в последовательности выполняется сдвижка. Эти шаги влияют на объем и количество анализируемых обучающих выборок. Для каждой полученной выборки производится обучение моделей обработки и, с использованием заранее определенного показателя качества осуществляется выбор модели с лучшими значениями.

Применяя многоуровневые модели обработки [32], возможно построение самоорганизующихся самообучающихся систем. Для этого результаты моделей должны сравниваться с реальными значениями объектов наблюдения, полученными от регистрирующих систем и устройств. В случае увеличения ошибок выше заранее определенного порога должно приниматься решение о формировании выборки данных. Над выборкой проводятся действия по формированию патчей, связанных с изменением длины и патчей, происходит очередная настройка и обучение модели.

Предлагаемый метод позволяет реализовывать процессы аугментации. Сегментируя последовательности данных, а затем объединяя выборки и перемешивая объекты наблюдений, полученных в разных условиях, возможно увеличение обучающих примеров.

В отличие от традиционных подходов, обрабатывающих информационные последовательности, этот метод формирует разные сгенерированные подпоследовательностей патчей таким образом, чтобы выбирать патчи, определяя значения функционала качества, свойства которых наилучшим образом соответствуют модели обработки.

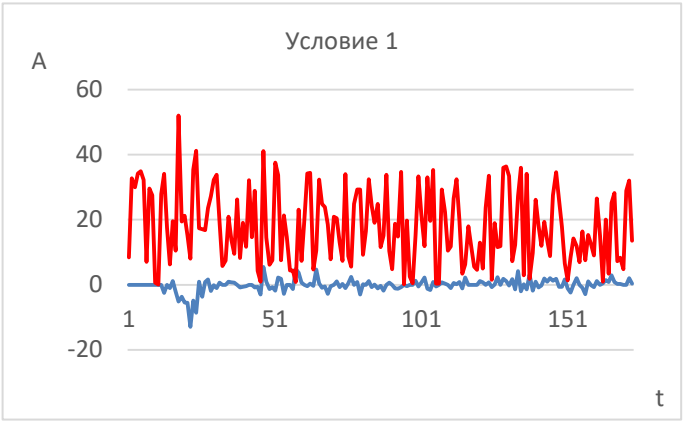
4. Эксперимент

4.1. Параметры эксперимента. В ходе эксперимента рассматривалась возможность создания выборок данных из информационных последовательностей в автоматическом режиме, так же влияние свойств содержащихся в них обучающих примеров на результат обработки различными моделями.

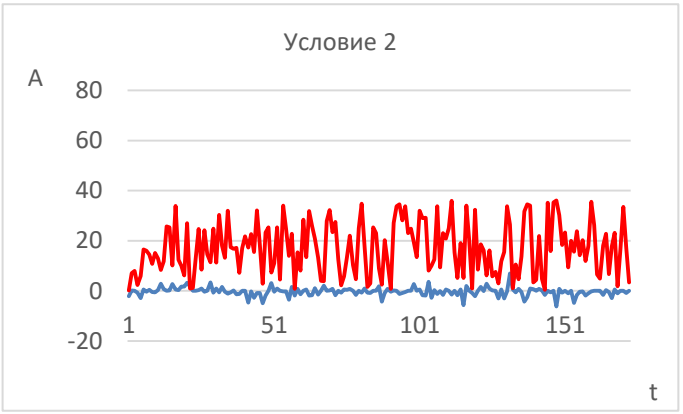
В качестве рассматриваемых данных в разделах 4.1-4.2 были выбраны последовательности датасета [33].

Цель эксперимента состояла в том, чтобы провести анализ влияния на качественные показатели моделей выборок, имеющих для последовательностей разные характеристики.

На вход самоорганизующейся самообучающейся модели подавались ряды, полученные в различных условиях. Их пример приведен на рисунке 2. Амплитуды радиосигналов во временной области двух каналов обозначены красным и синим цветом [34].



а)



б)

Рис. 2. Пример информационных последовательностей

Из поступившей на вход последовательности в автоматическом режиме формировались обучающие выборки с патчами разной длины и сдвига. Для всех выборок анализировались модели обработки на предмет достижения лучших показателей качества.

Состав вектора x определяется размерностью временного ряда и длинной подпоследовательности, выбранной для формирования объектов наблюдения.

Проверка достигаемых показателей качества выполнялась на группе моделей обработки. Их выбор связан с ограничениями на ресурсоемкость. Поэтому в эксперименте приоритет отдавался алгоритмам, имеющим относительно высокую скорость обучения и не требующим больших ресурсов.

На выборках патчей различной длины, при разных сдвигах 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 деревья классификации, допускающих максимум 10 разбиений. Гиперпараметры настройки моделей определены Matlab в [34] и использовались «заданными по умолчанию». Реализация процедур предлагаемого метода представлена в [35]

Формирование патчей информационных последовательностей происходило на основе неперекрывающихся окон. Функция разбиения разделяла последовательность на дискретные блоки. Образовывались обучающие выборки примеров, имеющих различные длины, а для каждой длины выполнялся сдвиг, т.е. для длины n формировалось n обучающих выборок. Оценка предлагаемого метода осуществлялась для задачи бинарной классификации. Формируемые выборки делились в соотношении 90:10 и 70:30 на обучающие и тестовые части. Длина объектов наблюдения, образуемых последовательностями, в выборке изменялась от 2 до 70 отсчетов. Один сдвиг осуществлялся на один отчет. Количество сдвигов равнялось длине.

Каждая полученная выборка подавалась на вход всем моделям. Оценивался заранее заданный показатель качества полноты:

$$accuracy = (TP + TN)/n, \quad (2)$$

где n – количество наблюдений, используемых для оценки модели, TP и TN – верно распознанные положительные и отрицательные объекты наблюдения.

В процессе обучения анализировались кортежи, определяющие из множества моделей A модель обработки $a_i \in A$, обучающую выборку, полученную методом разделения, и метод разделения на патчи $< a_i(x), X^{\mu_j}, \mu_j >$. Каждому кортежу в процессе обучения определялось значение функционала качества (2).

Фреймворк эксперимента представлен на рисунке 3.

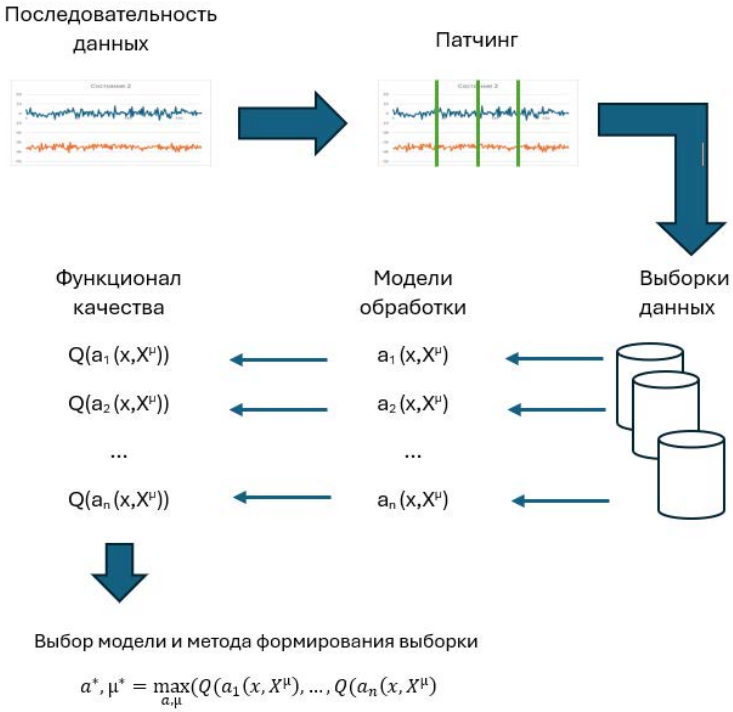


Рис. 3. Фреймворк проведения эксперимента

В результате обработки всех выборок становиться возможным выделить модель, метод формирования патчей и его характеристики, где достигаются лучшие значения выбранного функционала качества.

4.2. Проведение эксперимента. При проведении эксперимента на вход подавалась выборка временных рядов, полученных различных состояниях. На следующем шаге выполнялось разделение последовательности на патчи.

Способ формирования патчей для эксперимента представлен на рисунке 4.

Последовательности, полученные в состояниях 1 и 2, разбивались на патчи длиной *w*. Для каждой длины формировалось *w* сдвигов. Рассматривалось количество w последовательностей для каждой длины.

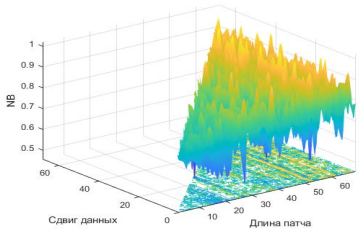
Рис. 4. Формирование патчей (w-ширина окна, 1 и 2 в верхнем индексе – состояние, нижний индекс – сдвиг)

Из полученных патчей составлялась выборка, которая разделялась на обучающую и тестовую части. Множеству патчей ставился в соответствие класс состояния последовательности, из которого оно было получены. Одна часть была помечена классом «состояние 1», а другая – классом «состояние 2». Последовательно изменяя длину шага разбиения, формировались патчи различной длины, получались выборки, содержащие обучающие примеры подпоследовательностей с разным количеством отсчетов и с различным сдвигом.

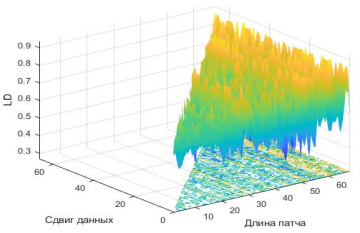
В результате для рассматриваемой задачи классификации в автоматическом режиме формировалась выборка данных, где каждый патч может рассматриваться как многомерный вектор, описывающий объект наблюдения.

Далее каждая обучающая выборка подавалась на вход классифицирующих алгоритмов, которые обучались и тестировались на ней. В результате для каждой модели и каждой выборки получались значения качественных показателей.

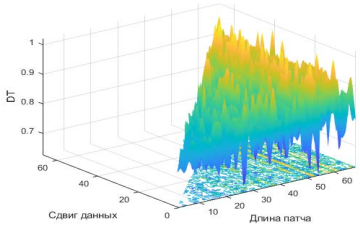
На рисунке 5 в осях «длина патча-сдвиг данных-точность(accuracy)» представлены значения для рассматриваемых моделей при соотношении 90:10 обучающих и тестирующих примеров. Графики показывают наличие областей, где достигается различные значения показателей качества. Для классифицирующих алгоритмов при одних способах формирования патчей значения достигаемых качественных показателей лучше, на других – хуже. Лучшие значения accuracy для рассматриваемых данных получаются на длинах 52 отсчета.



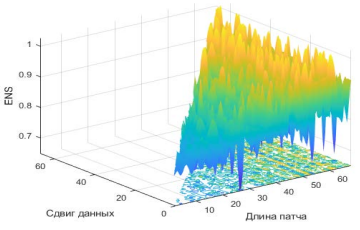
а) байесовский классификатор (NB)



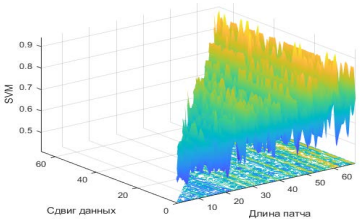
б) линейный дискриминант (LD)



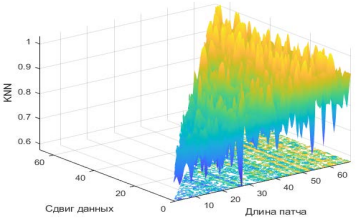
в) метод деревьев решений (DT)



г) ансамбль алгоритмов (ENS)



д) машина опорных векторов (SVM)

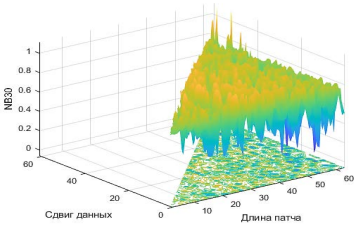


е) метод к-ближайших соседей (KNN)

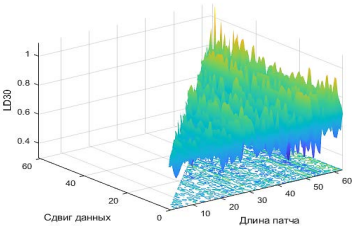
Рис. 5. Точность accuracy моделей при соотношении 90:10 для обучающих и тестирующих примеров

Увеличение размера объекта наблюдения приводит к появлению дополнительных признаков, которые могут содержать «значимую» информацию, улучшающую работу алгоритмов обработки. Дальнейший рост размера объекта наблюдения может привносить в его состав «шумовые» признаки и ухудшать обобщающую способность моделей. Изменение сдвига на фиксированной длине при формировании обучающей выборки оказывает влияние на количество дублирующихся, очень похожих примеров. Увеличение их количества может ухудшать, а уменьшение – улучшать достигаемые качественные показатели.

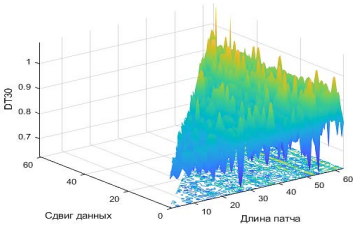
Такие же области получаются при изменении соотношений обучающих и тестовых примеров. На рисунке 6 показаны значения для тех же выборок данных в тех же осях при соотношении 70:30 для обучающих примеров и тестирующих примеров.



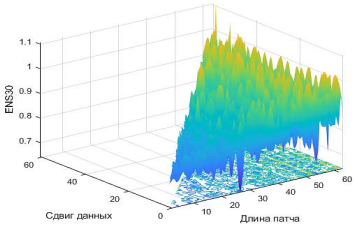
а) байесовский классификатор (NB)



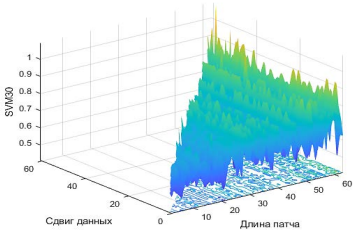
б) линейный дискриминант (LD)



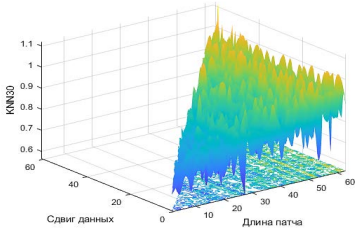
в) метод деревьев решений (DT)



г) ансамбль алгоритмов (ENS)



д) машина опорных векторов (SVM)



е) метод к-ближайших соседей (KNN)

Рис. 6. Точность accuracy моделей при соотношении 70:30 для обучающих и тестирующих примеров

Графики показывают, что для анализируемых данных для соотношений 90:10 и 70:30 области высоких и низких значений показателей качества accuracy прослеживаются и совпадают. Судя по графикам определенные длины и сдвиги оказывают влияние на достигаемые значения качества моделей. На рисунке 6 показана

развёртка значений accuracy по значениям «длина патча-сдвиг» при соотношении 70:30 обучающих и тестирующих примеров.

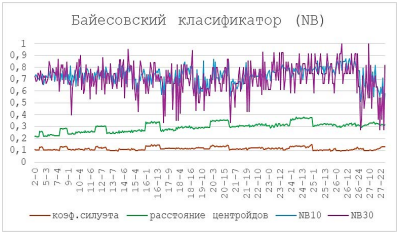
На рисунке 7 для оценки свойств обучающих примеров, полученных для каждой длины и сдвига выборки, были определены значения коэффициента силуэта и расстояния центройдов. Условно объекты наблюдения обучающей выборки разбивались на две группы, описывающие «состояние 1» и «состояние 2». Тогда, чем больше расстояние между центроидами групп обучающих примеров, помеченных разными классами, тем сильнее различаются группы. Это показывает на графиках рост и всплески зеленой линии расстояния центройдов. С другой стороны, для оценки свойств таких групп можно применить «коэффициент силуэта» (красная линия). Он показывает для группы объектов наблюдения насколько точно каждый объект соответствует своей группе и насколько он отличается от других групп.

Представленные в нижней части графиков линии коэффициента силуэта и расстояния центройдов показывают, что получаемые выборки объектов наблюдения в результате сдвига и изменения размера патча неоднородны.

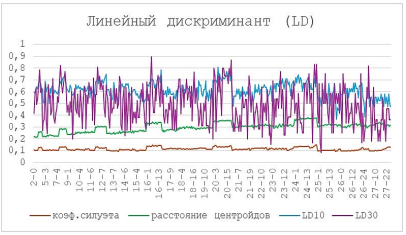
В обучающей выборке объекты наблюдения демонстрируют разнородность свойств. Среди них находятся как эталонные объекты, выполняющие функцию репрезентативных представителей класса, так и неинформативные, находящиеся на периферии границ классов или окруженные другими объектами, принадлежащими разным классам. Большое количество неинформативных, периферийных, шумовых объектов наблюдения негативно отражается на качестве классификации.

При формировании выборки изменения размера патчей и сдвиги внутри них могут влиять на относительное количество появления дублирующих или схожих примеров, затрудняющих процесс обучения модели.

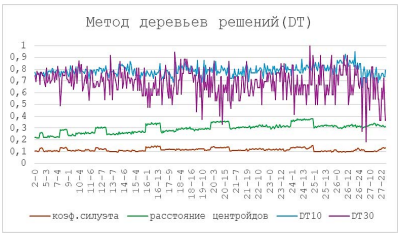
Тем не менее, графики показывают, что для различных соотношений обучающих и тестовых примеров, где наблюдается рост расстояния центройдов и увеличивается коэффициент силуэта прослеживается повышения значений качества accuracy, а при их падении – происходит понижение значений этого показателя качества.



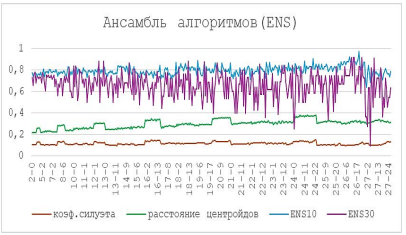
а) байесовский классификатор (NB)



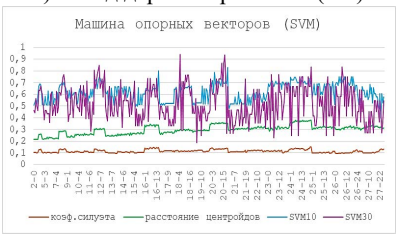
б) линейный дискриминант (LD)



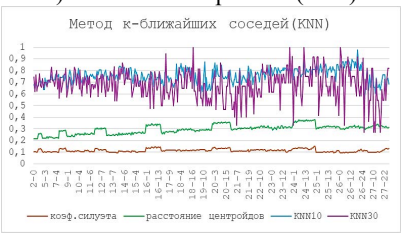
в) метод деревьев решений (DT)



г) ансамбль алгоритмов (ENS)



д) машина опорных векторов (SVM)



е) метод к-ближайших соседей (KNN)

Рис. 7. Значения accuracy (ось ординат) по совмещенной оси абсцисс «длина патча-сдвиг» для соотношений 90:10 и 70:30 обучающей и тестовой выборки

В рассматриваемом эксперименте анализируется бинарная классификация. Оценка качества обработки последовательности ошибки классифицирующих моделей определяются выражением:

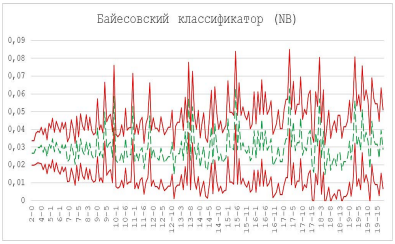
$$\rho = 1 - accuracy. \tag{3}$$

Предполагая гауссовское распределение, доверительный интервал ошибки классификаторов (3) (наивного Байеса, линейного дискриминанта, деревьев решений, ансамблевых классификаторов, KNN с гауссовском ядром) рассчитывается по формуле:

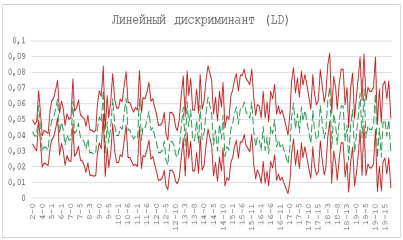
$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}}, \quad (4)$$

где p – ошибки классификации, 1,96 – критическое значение гауссова распределения для вероятности 95%, n – количество наблюдений, используемых для оценки модели.

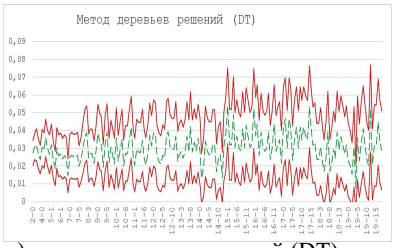
На рисунке 8 представлены изменения доверительных интервалов (4) классификаторов.



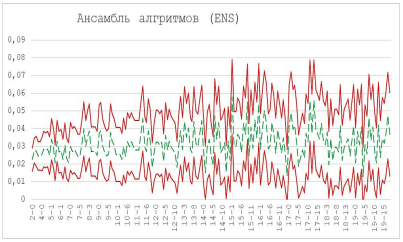
а) байесовский классификатор (NB)



б) линейный дискриминант (LD)

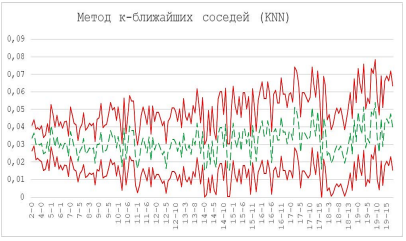


в) метод деревьев решений (DT)



г) ансамбль алгоритмов (ENS)

д) машина опорных векторов (SVM)



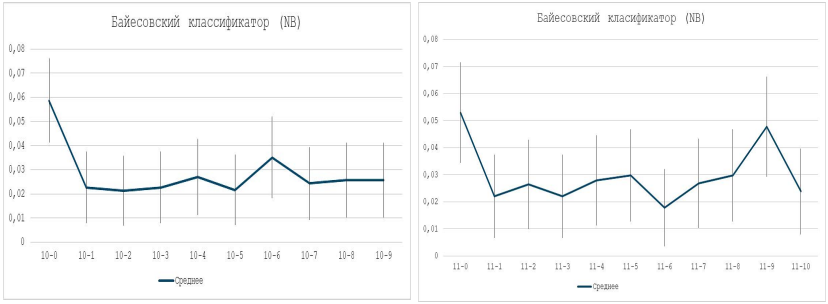
е) метод к-ближайших соседей (KNN)

Рис. 8. Доверительные интервалы классификаторов. Ось ординат – ошибка классификатора (зеленая линия) и его доверительный интервал (красная линия). Ось абсцисс – значения «длина патча-сдвиг»

Графики показывают, что по мере увеличения размера патча заметен рост доверительного интервала ошибки. Это связано

с ограниченностью объема выборки данных. Увеличение размера длины патча приводит к тому, что сокращается количество наблюдений, используемых для оценки модели. При формировании обучающей выборки необходимо оценивать её объем. Небольшой размер может преувеличивать точность модели из-за переобучения или случайных эффектов. Ограничения статистических методов оценки рекомендуют значения $n>=30$ для достижения значимых результатов, что является определенной границей при формировании множества обучающих примеров.

Тем не менее, на отдельных значениях длины патча и сдвига наблюдаются тенденции к улучшению или ухудшению результатов, где диапазоны доверительных интервалов не перекрываются. Рассматривая отдельные длины и сдвиги, образуемые для них, можно видеть неперекрывающиеся доверительные интервалы ошибок классификаторов. Например, на рисунке 9 для наивного Байеса представлены значения, ошибок для разных длин и сдвигов.



а) Ошибка на длине патча 10 б) Ошибка на длине патча 11

Рис.9 Доверительные интервалы на патчах разной длины. Ось ординат – ошибка классификатора и его доверительный. Ось абсцисс – значения «длина патча-сдвиг»

Графики рисунка 9 показывают существование неперекрывающихся доверительных интервалов (4), что позволяет говорить о влиянии длины и сдвига при формировании патча на достигаемые значения показателя качества.

4.3. Проверка на сторонних выборках данных.

В качестве анализируемых могут рассматриваться временные ряды, описывающие любые процессы, где, в зависимости от задачи классификации состояния, необходимо определить объекты из подпоследовательности, обладающими лучшими характеристиками

для достижения показателей качества. Для проверки были выбраны три набора.

Набор данных для обучения и тестирования, собранный компанией Ford в типичных условиях эксплуатации с минимальным шумовым загрязнением. Задача классификации состоит в диагностике наличия или отсутствия определенного симптома в автомобильной подсистеме [36].

Набор данных PowerCons содержит данные об индивидуальном потреблении электроэнергии домохозяйствами за один год, распределенные по двум сезонным классам: теплому (класс 1) и холодному (класс 2) [37].

Набор данных компании Sony для робота, оснащенного акселерометрами для измерения крена, тангажа и рыскания. Данные представлены только по оси X. Задача состоит в определении поверхности, по которой ходит робот (цемент или ковер) [38].

Для наборов данных проводилась нормировка, удаление дубликатов, неправильных значений выбросов, осуществлялось формирование последовательностей, которым ставилась метка состояния/класса для проверки метода.

Проверка ошибки алгоритмов классификации (4) вычислялась при формировании патчей и сдвигов на последовательностях других датасетов [36 – 38].

Определялись подпоследовательности, полученные для различных длин окна и сдвигов, затем выбирались средние значения и диапазоны доверительных интервалов. Размер патча для рассматриваемых датасетов варьировался от 2 до 20 отсчетов временного ряда.

В таблице 1 приведены достигаемые значения ошибок алгоритмов при формировании выборок данных описанным способом.

Значения ошибок алгоритмов, представленных в таблице 1 показывают, что выбор длины патча и сдвига оказывает влияние на результаты обработки последовательностей рассмотренными классифицирующими алгоритмами. Применяя описанный способ формирования обучающих выборок данных возможно повышение качественного показателя. Сравнение значений ошибок классификации позволяет определить, что сегментация выборки в большинстве случаев улучшает показатели качества обработки. Результаты, приведенные в таблице 1, показывают, что выбор размера патча и сдвига данных последовательности, в основном, улучшает показатели на 3-8%.

Применение патчинга со сдвигом изменяет свойства данных. Его влияние на слабые модели проявляется в большей степени, чем

на сильные. Результаты показывают возможность увеличения accuracy на 6-10% для слабых моделей, в то время как для сильных моделей наблюдается улучшение на 1-5% в сценариях с ограниченными данными.

Диапазоны лучших значений показателя (2) классифицирующих алгоритмов в таблице 1 в большинстве выборок для рассматриваемых классификаторов не пересекаются, что говорит о статистических значимых результатах

Таблица 1. Ошибки классифицирующих алгоритмов на выборках [36 – 38]

	Алгоритм	**Средняя ошибка алгоритма на датасете**	**Лучшее значение ошибки алгоритма**		
			Ошибка с доверительным интервалом	**Длина**	**Сдвиг**
Датасет 1	NB	0,11	0,016±0,012	9	1
	LD	0,09	0,013±0,011	9	8
	SVM	0,12	0,029±0,016	9	7
	DT	0,06	0,005±0,004	5	1
	KNN	0,07	0,005±0,004	5	3
	ENS	0,05	0,005±0,004	5	3
Датасет 2	NB	0,1	0,014±0,007	10	0
	LD	0,08	0,012±0,006	10	0
	SVM	0,13	0,016±0,009	10	2
	DT	0,09	0,013±0,003	4	1
	KNN	0,07	0,012±0,003	5	2
	ENS	0,06	0,012±0,003	4	0
Датасет 3	NB	0,16	0,019±0,014	11	4
	LD	0,11	0,013±0,012	9	7
	SVM	0,21	0,028±0,015	10	5
	DT	0,06	0,007±0,002	6	0
	KNN	0,05	0,009±0,007	6	2
	ENS	0,05	0,007±0,003	10	2

Проведённый эксперимент показывает, что варьирование параметрами сдвига и длины последовательности при формировании обучающих выборок оказывает влияние на эффективность обработки данных. При этом формируемые патчи должны адекватно отражать значащие свойства последовательностей, чтобы создать предпосылки для достижения классифицирующими алгоритмами более высоких показателей качества обработки информации.

Предлагаемый метод, использующий автоматическое формирование и выбор патчей на основе анализа достигаемых значений моделями обработки выбранного функционала качества, способствует снижению влияния шумовых компонентов в исходных данных.

Полученные результаты позволяют утверждать, что внедрение данного подхода увеличивает робастность алгоритмов классификации к вариациям входных данных.

4.4 Обсуждение результатов эксперимента. Формирование обучающей выборки приводит к тому, что объекты наблюдения в ней обладают разными свойствами. Часть из них являются эталонными экземплярами, позволяющими характеризовать класс, а другая часть – неинформативные, периферийные, шумовые, которые могут негативно влиять на качество классификации.

Эксперименты показывают, что оптимальное значение количества патчей сильно зависит от таких факторов, как сложность набора данных, длина, сдвиг. В зависимости от размера разделенных подпоследовательностей можно искать сдвиги, где захватываются значимые признаки, благодаря которым модель повышает качество обработки.

Изменение сдвига при формировании обучающей выборки приводит к тому, что могут появляться дублирующиеся или очень похожие примеры, что мешает обучению. Увеличение количества таких объектов приводит к настройке модели на отдельные характеристики данных в ущерб другим. Кроме того, может возникать ситуация, когда неконтролируемый осуществляемый в автоматическом режиме сдвиг формирует большое количество похожих или одинаковых объектов в одном классе, что может приводить к дисбалансу обучающих данных. Варьирование параметром сдвига может приводить как к возникновению эффекта шума, когда появляются значительное количество примеров на границах классов, так и к исчезновению таких объектов.

Изменение длины последовательности также оказывает влияние на качественные показатели моделей обработки. При рассмотрении

объекта наблюдения в виде вектора увеличение его размера приводит к появлению дополнительных признаков, которые могут как улучшать работу алгоритмов классификации путем внесения значимой информации, так и ухудшать обобщающую способность. В случае равномерного распределения объектов в признаковом пространстве добавление признака, может повысить удаленность объектов разных классов.

Значительное увеличение количества характеристик объектов наблюдения обучающей выборки может влиять на результат классификации из-за проклятия размерности. Появление лишних признаков, слабо влияющих на отклик модели снижает показатели качества обработки.

5. Заключение. Несмотря на значительные успехи и перспективы автоматического формирования анализируемых объектов наблюдения для самоорганизующихся систем, их обучение требует решения ряда проблемных вопросов для обработки информационных последовательностей.

Автоматическое определение структур данных и формирование обучающих, тестовых, валидационных выборок является одним из важнейших процессов, который сильно влияет на достигаемые значения показателей качества моделями обработки. Однако этим вопросам уделяется недостаточно внимания. В большинстве исследований формирование выборок данных осуществляется эмпирически. Для повышения качества обработки данных по-прежнему применяются различные эвристики, подходящие под рассматриваемые информационные свойства последовательностей, но не позволяющие производить обобщение в рамках теоретических и практических подходов.

Рассматриваемый в статье метод направлен на автоматическое формирование выборки на основе значений функционала качества. Применение разделения временного ряда на патчи и осуществление сдвигов при формировании объектов наблюдения является относительно универсальным способом и позволяет осуществлять построение данных под модель обработки.

Однако необходимо отметить, что эффективное применение описанного в статье метода зависит от ряда условий. Информационные последовательности обладают различными свойствами, влияющими на качество процессов обработки моделями. В результате такой «нарезки» могут возникать ситуации, когда размер патча не позволит выделить паттерн со значащей для обработки информации, или размер паттерна окажется слишком большим и будет

содержать много шумовых компонент, негативно влияющих на качество обработки.

На достигаемые показатели качества модели влияет размер обучающей выборки, но при автоматическом разбиении на патчи это влияние может быть разным. Изменение размера патча и манипуляции сдвигом при формировании выборки, повышающие различимость объектов наблюдения, принадлежащих разным классам, увеличивающие количество эталонные и уменьшающих число шумовых данных повышает точность обработки. В то же время эти манипуляции могут не только не приводить к существенным изменениям точности, но и ухудшать результаты. Это происходит, когда полученные таким образом обучающие объекты наблюдения оказываются нерепрезентативными или неинформативными. Возможность возникновения подобных ситуаций обуславливает необходимость валидационной поверки моделей обучения.

При формировании обучающей выборки необходимо оценивать её объем. Небольшой размер может преувеличивать точность модели из-за переобучения или случайных эффектов.

Тем не менее автоматические процессы разделение информационных последовательностей, формирование обучающих примеров, и их оценка на основе функционала качества в большинстве случает дают возможность повысить качество обработки.

Литература

1. Jarantow S.W., Pisors E.D., Chiu M.L. Introduction to the Use of Linear and Nonlinear Regression Analysis in Quantitative Biological Assays // Current Protocols. 2023. vol. 3. 801 p. DOI: 10.1002/cpz1.801.
2. He Y., Zhang X., Kong X., Yao L., Song Z. Causality-driven sequence segmentation assisted soft sensing for multiphase industrial processes // Neurocomputing. 2025. vol. 631 p. 129612. DOI: 10.1016/j.neucom.2025.129612.
3. Ци Д., Буре В.М. Исследование методов прогнозирования временных рядов для предсказания качества воздуха: объяснительный сравнительный анализ // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2024. Т. 20. №2. С. 206–219. DOI: 10.21638/spbu10.2024.206.
4. Man T., Osipov V.Yu., Zhukova N., Subbotin A.N., Ignatov D.I. Neural networks for intelligent multilevel control of artificial and natural objects based on data fusion: A survey // Information Fusion. 2024. vol. 110. 102427 p. DOI: 10.1016/j.inffus.2024.102427.
5. Belitser E., Ghosal S. Bayesian uncertainty quantification and structure detection for multiple change points models // Bernoulli. 2025. vol. 31. no. 2. pp. 1181–1205. DOI: 10.3150/24-BEJ1766.
6. 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.

7. Liu M., Zeng A., Chen M., Xu Z., Lai Q., Ma L., Xu Q. Scinet: Time Series Modeling and Forecasting With Sample Convolution and Interaction // Advances in Neural Information Processing Systems. 2022. vol. 35. pp. 1–13.
8. 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.
9. Chen D., Chen L., Zhang Y., Wen B., Yang C. A Multiscale Interactive Recurrent Network for Time-Series Forecasting // IEEE Transactions on Cybernetics. 2021. vol. 52. no. 9. pp. 8793–8803. DOI: 10.1109/TCYB.2021.3055951.
10. 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.
11. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is All you Need // Advances in Neural Information Processing Systems. 2017. vol. 30. pp. 1–11.
12. Wang J., Jiang J., Jiang W., Li C., Zhao W.X. LibCity: An open library for traffic prediction // Proceedings of the 29th International Conference on Advances in Geographic Information Systems, SIGSPATIAL'21. 2021. pp. 145–148. DOI: 10.1145/3474717.3483923.
13. Grover S., Jalali A., Etemad A. Segment, Shuffle, and Stitch: A Simple Layer for Improving Time Series Representations // Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS). 2024. pp. 4878–4905. DOI: 10.52202/079017-0158.
14. Salehi A., Balasubramanian M. DDCNet: Deep dilated convolutional neural network for dense prediction // Neurocomputing. 2023. vol. 523. pp. 116–129. DOI: 10.1016/j.neucom.2022.12.024.
15. Chen Y., Lu X., Xie Q. Collaborative networks of transformers and convolutional neural networks are powerful and versatile learners for accurate 3D medical image segmentation // Computers in Biology and Medicine. 2023. vol. 164 p. 107228. DOI: 10.1016/j.compbiomed.2023.107228.
16. Woo G., Liu C., Kumar A., Xiong C., Savarese S., Sahoo D. Unified Training of Universal Time Series Forecasting Transformers // Proceedings of the 41st International Conference on Machine Learning (ICML). 2024. pp. 1–25.
17. Ekambaram V., Ati A., Nguyen N., Sinthong P., Kalagnanam J. TSMixer: Lightweight MLP-Mixer Model for Multivariate Time Series Forecasting // Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23). 2023. pp. 459–469. DOI: 10.1145/3580305.3599533.
18. Goswami M., Szafer K., Choudhry A. et al. Moment: A Family of Open Time-Series Foundation Models // Proceedings of the 41st International Conference on Machine Learning. vol. 235. 2024. pp. 16115–16152.
19. Vishwas B.V.K., Macharla S.R. MOIRAI: A Time Series LLM for Universal Forecasting // Time Series Forecasting Using Generative AI. 2025. pp. 183–194. DOI: 10.1007/979-8-8688-1276-7_7.
20. Yang H., Yang D. CSwin-PNet: A CNN-Swin Transformer combined pyramid network for breast lesion segmentation in ultrasound images // Expert Systems with Applications. 2023. vol. 213. part b. 119024 p. DOI: 10.1016/j.eswa.2022.119024.
21. Xiao Q., Wu B., Zhang Yu., Liu S., Pechenizkiy M., Mocanu E., Mocanu D.C. Dynamic Sparse Network for Time Series Classification: Learning What to "see" // Advances in Neural Information Processing Systems. 2022. vol. 35. pp. 1–14.
22. Liu M., Zeng A., Chen M., Xu Z., Lai Q., Ma L., Xu Q. Scinet: Time Series Modeling and Forecasting with Sample Convolution and Interaction // Advances in Neural Information Processing Systems. 2022. vol. 35. pp. 5816–5828.

23. Zhou H., Zhang S., Peng J., Zhang S., Li J., Xiong H., Zhang W. Informer: Beyond efficient transformer for long sequence time-series forecasting // AAAI Conference on Artificial Intelligence. 2021. vol. 35. no. 12. pp. 11106–11115.
24. Wu H., Xu J., Wang J., Long M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting // Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 22419–22430.
25. Chen Y., Ren K., Wang Y., Fang Y., Sun W., Li D. Contiformer: Continuous-time transformer for irregular time series modeling // Advances in Neural Information Processing Systems. 2024. vol. 36. pp. 47143–47175.
26. Zeng A., Chen M., Zhang L., Xu Q. Are transformers effective for time series forecasting? // Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37. no. 9. pp. 11121–11128.
27. Lee S., Park T., Lee K. Soft contrastive learning for time series // The International Conference on Learning Representations. 2024. vol. 37. pp. 11121–11128.
28. Tallman E., West M. Bayesian predictive decision synthesis // Journal of the Royal Statistical Society. Series B: Statistical Methodology. 2024. vol. 86. no. 2. pp. 340–363.
29. 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.
30. Peng Y., Hu X., Hao X., Liu P., Deng Y., Li Z. Spider-Net: High-resolution multi-scale attention network with full-attention decoder for tumor segmentation in kidney, liver and pancreas // Biomedical Signal Processing and Control. 2024. vol. 93. p. 106163. DOI: 10.1016/j.bspc.2024.106163.
31. Lebedev I., Sukhoparov M., Semenov V., Khasanov D. Adaptive Segmentation of Information Sequences for Machine Learning Modular Regression Models // Emerging Science Journal. 2025. vol. 9. no. 5. pp. 2420–2438.
32. Shi X., Song X., Deng M., Zhang D., Li X., Chen B. UNet and Swin Transformer Fusion Network for Lesion Segmentation in Biological Kidney Imaging // International Journal of Pattern Recognition and Artificial Intelligence. 2025. vol. 39. no. 12. 2550021 p. DOI: 10.1142/s0218001425500211.
33. Kaggle. E-Commerce Data. URL: https://www.kaggle.com/datasets/carrie1/ecommerce-data (дата обращения: 01.12.2025).
34. Kaggle. Hourly energy demand generation and weather. URL: https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather/data (дата обращения: 01.12.2025).
35. Kaggle. Pima Indians Diabetes Database. URL: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (дата обращения: 01.12.2025).

Лебедев Илья Сергеевич — д-р техн. наук, профессор, главный научный сотрудник, лаборатория интеллектуальных систем, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН). Область научных интересов: методы машинного обучения, представление и обработка слабоструктурированных данных, применение методов искусственного интеллекта в системах информационной безопасности. Число научных публикаций — 200. isl_box@mail.ru; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-3311.

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

DOI 10.15622/ia.25.3.1

I. LEBEDEV

ADAPTIVE DATA SAMPLES GENERATION FOR SELF-ORGANIZING CONTRAST-BASED LEARNING SYSTEMS

Lebedev I. **Adaptive Data Samples Generation for Self-Organizing Contrast-Based Learning Systems.**

Abstract. The application of contrast-based learning and self-organizing models for classification and forecasting tasks in processing time series and information sequences faces a number of data organization challenges. Improving the processing quality of such systems necessitates improving the methods for generating observation objects and training data samples. This article proposes a method for generating and analyzing data samples based on selecting information sequence patches with different length and shift characteristics, distinguished by the use of the quality functional of processing models. An experiment on simulated data and samples is used to evaluate the proposed method. Accuracy values are obtained for different processing algorithms and with different patch lengths and shifts. The properties of the resulting data patches are determined using the silhouette coefficient and inter-centroid distance metrics. The errors of the classification algorithms are analyzed, and confidence intervals for the errors are determined. It is determined that at a statistical significance level of $p = 0.05$, changing the patch length and shift affects the achieved accuracy values of the classification algorithms. The proposed method improves accuracy by selecting the length and shift during patch generation and assigning the models with the best performance. Results show a potential improvement of 6–10% for weak models, while strong models achieve an improvement of 1–5% in data-constrained scenarios. Analysis of the proposed solution shows that varying the shift and sequence length parameters when generating training data sets impacts data processing efficiency.

Keywords: machine learning, adaptive models, processing quality improvement, patch generation.

References

1. Jarantow S.W., Pisors E.D., Chiu M.L. Introduction to the Use of Linear and Nonlinear Regression Analysis in Quantitative Biological Assays. Current Protocols. 2023. vol. 3. 801 p. DOI: 10.1002/cpz1.801.
2. He Y., Zhang X., Kong X., Yao L., Song Z. Causality-driven sequence segmentation assisted soft sensing for multiphase industrial processes. Neurocomputing. 2025. vol. 631 p. 129612. DOI: 10.1016/j.neucom.2025.129612.
3. Qi D., Bure V.M. [Research on Time Series Forecasting Methods for Air Quality Prediction: An Explanatory Comparative Analysis]. Vestnik Sankt-Peterburgskogo universiteta. Prikladnaya matematika. Informatika. Protsessy upravleniya – Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes. 2024. vol. 20. no. 2. pp. 206–219. DOI: 10.21638/spbu10.2024.206. (In Russ.).
4. Man T., Osipov V.Yu., Zhukova N., Subbotin A.N., Ignatov D.I. Neural networks for intelligent multilevel control of artificial and natural objects based on data fusion: A survey. Information Fusion. 2024. vol. 110. 102427 p. DOI: 10.1016/j.inffus.2024.102427.
5. Belitser E., Ghosal S. Bayesian uncertainty quantification and structure detection for multiple change points models. Bernoulli. 2025. vol. 31. no. 2. pp. 1181–1205. DOI: 10.3150/24-BEJ1766.

6. 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.
7. Liu M., Zeng A., Chen M., Xu Z., Lai Q., Ma L., Xu Q. Scinet: Time Series Modeling and Forecasting With Sample Convolution and Interaction. Advances in Neural Information Processing Systems. 2022. vol. 35. pp. 1–13.
8. 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.
9. Chen D., Chen L., Zhang Y., Wen B., Yang C. A Multiscale Interactive Recurrent Network for Time-Series Forecasting. IEEE Transactions on Cybernetics. 2021. vol. 52. no. 9. pp. 8793–8803. DOI: 10.1109/TCYB.2021.3055951.
10. Lebedev I.S. [Adaptive application of machine learning models on individual sample segments in regression and classification tasks]. Informatsionno-upravliaiushchie sistemy – Information and Control Systems. 2022. no. 3. pp. 20–30. DOI: 10.31799/1684-8853-2022-3-20-30. (In Russ.).
11. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is All you Need. Advances in Neural Information Processing Systems. 2017. vol. 30. pp. 1–11.
12. Wang J., Jiang J., Jiang W., Li C., Zhao W.X. LibCity: An open library for traffic prediction. Proceedings of the 29th International Conference on Advances in Geographic Information Systems, SIGSPATIAL'21. 2021. pp. 145–148. DOI: 10.1145/3474717.3483923.
13. Grover S., Jalali A., Etemad A. Segment, Shuffle, and Stitch: A Simple Layer for Improving Time Series Representations. Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS). 2024. pp. 4878–4905. DOI: 10.52202/079017-0158.
14. Salehi A., Balasubramanian M. DDCNet: Deep dilated convolutional neural network for dense prediction. Neurocomputing. 2023. vol. 523. pp. 116–129. DOI: 10.1016/j.neucom.2022.12.024.
15. Chen Y., Lu X., Xie Q. Collaborative networks of transformers and convolutional neural networks are powerful and versatile learners for accurate 3D medical image segmentation. Computers in Biology and Medicine. 2023. vol. 164 p. 107228. DOI: 10.1016/j.compbiomed.2023.107228.
16. Woo G., Liu C., Kumar A., Xiong C., Savarese S., Sahoo D. Unified Training of Universal Time Series Forecasting Transformers. Proceedings of the 41st International Conference on Machine Learning (ICML). 2024. pp. 1–25.
17. Ekambaram V., Ati A., Nguyen N., Sinthong P., Kalagnanam J. TSMixer: Lightweight MLP-Mixer Model for Multivariate Time Series Forecasting. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23). 2023. pp. 459–469. DOI: 10.1145/3580305.3599533.
18. Goswami M., Szafer K., Choudhry A. et al. Moment: A Family of Open Time-Series Foundation Models. Proceedings of the 41st International Conference on Machine Learning. vol. 235. 2024. pp. 16115–16152.
19. Vishwas B.V.K., Macharla S.R. MOIRAI: A Time Series LLM for Universal Forecasting. Time Series Forecasting Using Generative AI. 2025. pp. 183–194. DOI: 10.1007/979-8-8688-1276-7_7.
20. Yang H., Yang D. CSwin-PNet: A CNN-Swin Transformer combined pyramid network for breast lesion segmentation in ultrasound images. Expert Systems with Applications. 2023. vol. 213. part b. 119024 p. DOI: 10.1016/j.eswa.2022.119024.

21. Xiao Q., Wu B., Zhang Yu., Liu S., Pechenizkiy M., Mocanu E., Mocanu D.C. Dynamic Sparse Network for Time Series Classification: Learning What to "see". Advances in Neural Information Processing Systems. 2022. vol. 35. pp. 1–14.
22. Liu M., Zeng A., Chen M., Xu Z., Lai Q., Ma L., Xu Q. Scinet: Time Series Modeling and Forecasting with Sample Convolution and Interaction. Advances in Neural Information Processing Systems. 2022. vol. 35. pp. 5816–5828.
23. Zhou H., Zhang S., Peng J., Zhang S., Li J., Xiong H., Zhang W. Informer: Beyond efficient transformer for long sequence time-series forecasting. AAAI Conference on Artificial Intelligence. 2021. vol. 35. no. 12. pp. 11106–11115.
24. Wu H., Xu J., Wang J., Long M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 22419–22430.
25. Chen Y., Ren K., Wang Y., Fang Y., Sun W., Li D. Contiformer: Continuous-time transformer for irregular time series modeling. Advances in Neural Information Processing Systems. 2024. vol. 36. pp. 47143–47175.
26. Zeng A., Chen M., Zhang L., Xu Q. Are transformers effective for time series forecasting?. Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37. no. 9. pp. 11121–11128.
27. Lee S., Park T., Lee K. Soft contrastive learning for time series. The International Conference on Learning Representations. 2024. vol. 37. pp. 11121–11128.
28. Tallman E., West M. Bayesian predictive decision synthesis. Journal of the Royal Statistical Society. Series B: Statistical Methodology. 2024. vol. 86. no. 2. pp. 340–363.
29. 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.
30. Peng Y., Hu X., Hao X., Liu P., Deng Y., Li Z. Spider-Net: High-resolution multi-scale attention network with full-attention decoder for tumor segmentation in kidney, liver and pancreas. Biomedical Signal Processing and Control. 2024. vol. 93. p. 106163. DOI: 10.1016/j.bspc.2024.106163.
31. Lebedev I., Sukhoparov M., Semenov V., Khasanov D. Adaptive Segmentation of Information Sequences for Machine Learning Modular Regression Models. Emerging Science Journal. 2025. vol. 9. no. 5. pp. 2420–2438.
32. Shi X., Song X., Deng M., Zhang D., Li X., Chen B. UNet and Swin Transformer Fusion Network for Lesion Segmentation in Biological Kidney Imaging. International Journal of Pattern Recognition and Artificial Intelligence. 2025. vol. 39. no. 12. 2550021 p. DOI: 10.1142/s0218001425500211.
33. Kaggle. E-Commerce Data. Available at: https://www.kaggle.com/datasets/carrie1/ecommerce-data (accessed 01.12.2025).
34. Kaggle. Hourly energy demand generation and weather. Available at: https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather/data (accessed 01.12.2025).
35. Kaggle. Pima Indians Diabetes Database. Available at: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (accessed 01.12.2025).

Lebedev Ilya — Ph.D., Dr. Sci., Professor, Chief scientific officer, Laboratory of Intelligent Systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: machine learning methods, representation and processing of poorly structured data, application of artificial intelligence methods in information security systems. The number of publications — 200. isl_box@mail.ru; 39, 14th line V.O., 199178, Saint-Petersburg, Russia; office phone: +7(812)328-3311.

Acknowledgements. The study was funded by the Russian Science Foundation grant No. 25-21-00269, https://rscf.ru/project/25-21-00269/.