

МЕТОД АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ ОБУЧАЮЩЕГО НАБОРА ДАННЫХ ВРЕМЕННЫХ РЯДОВ СЕТЕВОГО ТРАФИКА В ЗАДАЧЕ КЛАССИФИКАЦИИ СОСТОЯНИЯ СИСТЕМЫ

DOI: 10.36724/2072-8735-2026-20-7-36-44

Manuscript received 08 April 2026;

Accepted 10 June 2026

Лебедев Илья Сергеевич,Санкт-Петербургский Федеральный исследовательский центр
Российской академии наук, г. Санкт-Петербург, Россия,
isl_box@mail.ruИсследование выполнено за счет гранта Российского
научного фонда № 25-21-00269, <https://rscf.ru/project/25-21-00269/>**Сухопаров Михаил Евгеньевич,**Государственный университет морского и речного флота имени
адмирала С.О. Макарова, Санкт-Петербург, Россия,
mail@sukhoparovm.ru**Ключевые слова:** сетевой трафик, временные ряды,
классификация состояний, автоматическое
формирование объектов, машинное обучение

В статье представлен метод автоматического формирования объектов наблюдения временных рядов сетевого трафика для классификации состояний информационно-телекоммуникационных систем. Рост объемов данных сетевого трафика, неоднозначность проявления различных событий функционирования информационных систем, высокая сложность и трудоемкость процессов разметки обуславливает необходимость разработки адаптивных методов обработки данных. В отличие от классических подходов, в предлагаемом методе происходит не настройка и оптимизация модели на выборке данных, а, наоборот, объекты наблюдения формируются из последовательности "под модель". Предлагаемый метод автоматически генерирует обучающие выборки путём автоматического разбиения исходных последовательностей на патчи (подпоследовательности) с варьируемыми длиной и сдвигом. Из них формируются выборки данных. На каждой итерации, на полученных выборках, происходит обучение, оценка достигаемых показателей качества и выбор параметров патчей и моделей, которые достигают лучших результатов. Проведён эксперимент на синтетическом наборе данных для бинарной классификации состояния системы. Показано, что применение сдвига в последовательности улучшает обобщающую способность за счёт аугментации данных, а оптимальный размер патча балансирует информативность и уровень шума. Предложенный метод автоматизирует подготовку данных и позволяет выбирать оптимальные параметры формирования патчей, повышает различимость классов, может использоваться как метод аугментации данных для улучшения обучения моделей, даёт возможность применять менее ресурсоёмкие модели для обработки сложных структур данных. Метод позволяет повысить качество классификации в среднем на 2-8%, снизить вычислительные затраты, повысить интерпретируемость и эффективно обрабатывать сложные структуры данных.

Информация об авторах:**Лебедев Илья Сергеевич**, д.т.н., профессор, главный научный сотрудник, лаборатория интеллектуальных систем, СПИИРАН, СПб ФИЦ РАН, г. Санкт-Петербург, Россия. ORCID 0000-0001-6753-2181**Сухопаров Михаил Евгеньевич**, д-р техн. наук, доцент, профессор кафедры радиосвязи на морском флоте, ГУМРФ имени адмирала С.О. Макарова, г. Санкт-Петербург, Россия. ORCID 0000-0003-1798-8257**Для цитирования:**

Лебедев И.С., Сухопаров М.Е. Метод автоматического формирования и выделения объектов наблюдения во временных рядах сетевого трафика в задаче классификации состояния системы // Т-Сотм: Телекоммуникации и транспорт. 2026. Том 20. №7. С. 36-44.

For citation:

Lebedev I.S., Sukhoparov M.E. "A method for automatic formation and selection of observation objects in network traffic time series in sys-tem state classification problem", T-Comm, 2026, vol. 20, no.7, pp. 36-44. (in Russian)

Введение

Повсеместное использование технологий сбора и хранения данных сетевого трафика способствует накоплению большого объема временных рядов и информационных последовательностей. Их разметка является сложным, трудоемким процессом, в котором выделение отдельных событий может трактоваться по-разному [1, 2]. В связи с этим, особую актуальность приобретает задача разработки эффективных моделей для прогнозирования поведения систем, состояния сетевых устройств и каналов передачи данных.

Возникающие колебания интенсивности сетевого трафика, динамически изменяющиеся характеристики входных данных, проявление "дрейфа концепта" обуславливают необходимость развития адаптивных самонастраивающихся методов обработки данных [3, 4].

При этом применение таких подходов требует автоматической обработки данных, выявления информативных признаков, что приводит к решениям, опирающимся, например, на методы AutoML. Они направлены на оптимизацию гиперпараметров моделей, динамических настроек обрабатывающих алгоритмов, поиск оптимизированных конфигураций [5, 6]. Однако высокие ресурсные требования, сложность отладки "черного ящика", риск переобучения и ложной оптимизации, зависимость от качества данных может являться существенным ограничением данной технологии [7, 8].

В связи с этим актуальной становится задача предварительного формирования и оценки входной выборки, с учётом выявленных свойств данных и последующего построения моделей обработки, демонстрирующих наилучшие качественные показатели применительно к характеристикам каждой выделенной подпоследовательности [9].

Цель работы – разработать метод автоматического формирования и выделения объектов наблюдения в данных временных рядов сетевого трафика для классификации состояний информационно-телекоммуникационных систем, позволяющий снизить трудоёмкость подготовки данных и повысить качество классификации.

Автоматическое формирование обучающих выборок путём разбиения исходных последовательностей сетевого трафика на патчи с варьируемыми длиной и сдвигом, а также последующий выбор оптимальных параметров патчей и моделей на основе оценки функционала качества могут позволить повысить различимость классов состояний системы, улучшить обобщающую способность моделей за счёт аугментации данных, снизить вычислительные затраты за счёт выбора менее ресурсоёмких моделей, подходящих для конкретных характеристик патчей, повысить интерпретируемость результатов классификации.

Научная новизна работы заключается в том, что предлагается решение обратной задачи к обработке временных рядов сетевого трафика, где вместо настройки модели под данные осуществляется формирование объектов наблюдения и автоматическая оптимизация их параметров «под модель». Кроме того, формирование патчей со сдвигом может рассматриваться как метод аугментации данных, улучшающий обобщающую способность моделей.

Практическая значимость состоит в том, что предлагаемый метод дает возможность автоматизации трудоёмкого процесса разметки данных сетевого трафика, позволяет выбирать оптимальные параметры формирования патчей для

конкретных моделей обработки и позволяет применять менее ресурсоёмкие модели для обработки сложных структур данных без потери качества классификации.

1 Обзор существующих методов

Анализ временных рядов сетевого трафика является одной из основных задач для мониторинга и классификации состояний информационно-телекоммуникационных систем. Ее решение базируется на применении методов машинного обучения таким образом, чтобы оптимизировать модель под конкретные свойства последовательности трафика. Процесс оптимизации предполагает два направления: улучшение качества входных данных и построение эффективной модели обработки [10, 11].

Для повышения качества данных сетевого трафика используют методы формирования пространства признаков, кластеризацию, поиск точек разладки временных рядов и обнаружение «дрейфа концепта». Выявление точек изменения свойств предикторов и целевых переменных критично для задач анализа и обработки трафика. Здесь применяют байесовский анализ, метод максимального правдоподобия, квантильную регрессию и другие подходы. Но из-за сложности аппроксимирующих функций такие методы требуют больших вычислительных ресурсов [12-14].

При работе с временными рядами сетевого трафика важно выделять сегменты с устойчивыми характеристиками и отслеживать закономерности их изменений. Сегментация, по сути, заключается в разработке детектора, фиксирующего моменты смены свойств ряда.

Во многих исследованиях процессы разбиения выборки в задачах машинного обучения рассматриваются как вспомогательные и анализируются преимущественно в контексте решения узкоспециализированных задач. Выбор методов и их эффективность в значительной степени определяются типом и характеристиками обрабатываемых данных [15].

При работе с информационными последовательностями сетевого трафика отмечаются несколько ключевых направлений. Классические методы кластеризации используются для обработки временных рядов, что улучшает качество данных, обладающих определёнными свойствами. В условиях высокой размерности, где имеется более одной точки изменения, применяется сегментация данных с использованием методов динамического программирования, бинарной сегментации и байесовских подходов [16, 17]. Байесовские методы для регрессионных моделей с точками изменения, как показано в работах, имеют высокую вычислительную сложность и требуют много итераций. Другие подходы используют статистические пороги для сегментации и учитывают выбросы и шумы в данных.

Эффективность применения различных подходов глубокого обучения зависит от вычислительных ресурсов, выделения и анализа сегментов сетевого трафика [18, 19]. Применение методов расширенного теста Дики-Фуллера для проверки стационарности, оценки максимального правдоподобия, предлагают различные способы выявления точек изменений. Однако у всех этих методов есть общие недостатки: сложность обучения, высокая ресурсоёмкость и риск некорректного выбора моделей, что может ухудшить прогноз [20, 21].

Анализ существующих методов выявляет ряд проблемных вопросов задач классификации состояний информационно-телекоммуникационных систем. Большинство методов требуют значительных вычислительных ресурсов, огромного объема качественно размеченных наборов данных. Однако динамически меняющиеся характеристики сетевого трафика, неоднозначность трактовки событий, появления новых типов аномалий не всегда, проявление внешних воздействий позволяют добиться приемлемого качества. Кроме того, некачественная разметка может приводить к возникновению риска переобучения.

В связи с этим возникает задача оценки входной выборки, её сегментации с учётом свойств данных, а также формирования моделей обработки, обеспечивающих наилучшие качественные показатели для выявленных характеристик анализируемой последовательности.

Таким образом, необходим адаптивный метод, который автоматически формирует обучающие выборки с учётом свойств данных, минимизирует требования к разметке, обеспечивает интерпретируемость результатов, а также обеспечивает аугментацию данных для улучшения обобщающей способности. Предлагаемый в работе метод направлен на решение этих задач за счёт автоматического формирования объектов наблюдения с оптимизацией их параметров под конкретную модель обработки.

2 Предлагаемый метод

Функционирование различных систем сопряжено с необходимостью определения их текущего состояния. Анализ временных рядов сетевого трафика широко применяется в системах мониторинга информационно-телекоммуникационных сетей, где решается задача классификация подпоследовательностей. Она состоит в отнесении отдельных элементов временного ряда к определённым состояниям, таким как «рабочее», «предотказное» или «неработоспособное», выявления аномалий в сетевом трафике путём анализа характеристик пакетов, интервалов между их поступлениями, направления передачи последовательности и других показателей.

Однако накопленный значительный объем исследуемых последовательностей, необходимость анализа большого количества характеристик, наличие неявных закономерностей обуславливает необходимость создания и применения сложных моделей обработки. Кроме того, возникает ряд проблемных вопросов, связанных с трудоемкостью процессов разметки данных.

Решение задачи классификации состояния на основе автоматического выделения репрезентативных фрагментов из информационных последовательностей временных рядов сетевого трафика.

В рассматриваемом случае автоматическая разметка данных основывается на разделении последовательностей на патчи, содержащие различное количество отсчетов. Формирование патчей разной длины может «захватить» значимую информацию, сократить длину последовательности и уменьшить вычислительную сложность. В случаях периодичности в данных выбор оптимального размера подпоследовательности позволяет моделям изучать более длинные временные интервалы, что может повысить качество прогнозов. Однако

выбор размера патча и его свойств должен зависеть от конкретной задачи и характеристик данных.

Классификация состояния системы является одной из важных задач при идентификации неисправностей техники, мониторинге состояния киберфизических систем, обнаружения аномалий в трафике. В результате возникает задача, в которой необходимо определить, к какому состоянию относится элемент временного ряда на основе анализа данных, полученных через определённые промежутки времени. Формальная постановка будет выглядеть следующим образом.

Имеется множество состояний $\{c_1, \dots, c_n\} \in C$ информационной системы. В каждом состоянии c_j получен временной ряд сетевого трафика $X^{(c_j)} = \{x_{t_i}^{(c_j)}\}_{i=1}^T$.

Он разбивается на патчи $p_i = \{x_{t_i}^{(c_j)}, x_{t_{i+1}}^{(c_j)}, \dots, x_{t_{i+w-1}}^{(c_j)}\}$, и, в зависимости в каком состоянии был получен, ему ставится соответствующая метка c_j .

Получается обучающее множество патчей $P^\theta = \{p_i^{(c_j)}\}_{i=1}^N$, характеризующееся параметрами $\theta = (w, s)$ длины и сдвига. Меняя эти параметры в автоматическом режиме, можно формировать различные обучающие выборки для моделей обучения.

Для обработки обучающей выборки определяется ряд моделей обработки $\{a_1, \dots, a_K\} \in A$.

Необходимо определить множество P и параметры его формирования (длина окна, сдвиг), назначить модель обработки $a_k \in A$ имеющей лучшее значение функционала качества

$$Q(a_k(p_i^{(c_j)}, P^\theta), C) \rightarrow \max_{a_k \in A, \theta \in \Theta} \quad (1)$$

Таким образом, выбор параметров на основе выражения (1) формулируется как задача дискретной оптимизации:

$$(\theta^*, a^*) = \arg \max_{a_k \in A, \theta \in \Theta} Q(a_k(p_i^{(c_j)}, P^\theta), C) \quad (2)$$

Особенностью предлагаемого метода, описанного выражениями (1) и (2), является автоматический выбор и формирование объектов наблюдения из данных для повышения качества моделей 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». Критерием оценки являются значения функционала качества.

Обработка данных с использованием предлагаемого подхода проходит несколько этапов. Сначала выбираются базовые алгоритмы обработки. Из последовательностей создаются объекты наблюдения и формируются множества обучающих выборок данных. Причем на каждой итерации изменяются свойства создаваемых объектов, например, длина и сдвиг. Затем проводится анализ значений выбранного показателя качества заранее определенных моделей на всех выборках данных. После этого выбирается метод формирования объектов наблюдения и назначается "лучшая" модель.

На рисунке 1 представлена Алгоритмическая последовательность шагов метода формирования и анализа выборок данных.

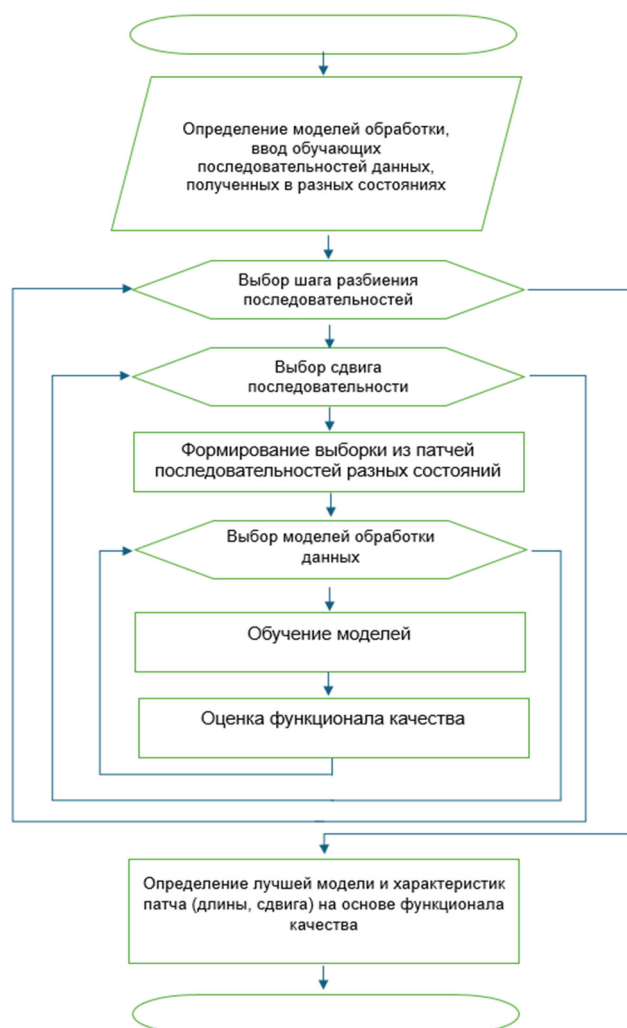


Рис. 1. Алгоритмическая последовательность шагов метода формирования и анализа выборок данных

Предлагаемый метод включает следующую последовательность действий. Вначале задается пул моделей обработки. Определяются исходные информационные последовательности, полученные в различных состояниях системы. Затем происходит циклический перебор длины и сдвига для формирования патчей в результате которого автоматически изменяются свойства объектов. Меняется длина разбиения и сдвиг одной последовательности относительно другой. Образуется обучающая выборка, содержащая набор патчей одинакового размера, каждый из которых помечен классом, соответствующим состоянию системы.

Выборка поочередно подается на вход отобранных моделей. Осуществляется процесс обучения. Для каждой модели и выборки, содержащей патчи, характеризующиеся длиной и сдвигом, определяется значение выбранного функционала качества. Далее происходит повторение процессов формирования выборок для других значений сдвига и шага разбиения, обучения моделей, вычисления значений функционала качества.

По окончании перебора определяются параметры длины и сдвига патча для формирования выборки и модель, достигающая лучших значений выбранного показателя качества. Псевдокод предлагаемого метода представлен на рисунке 2.

Вход:

временные ряды $X^{(c)}$;
пул моделей $\{a_1, \dots, a_k\} \in A$;
диапазоны параметров w, s
метрика качества Q .

Выход:

оптимальные параметры w^*, s^* ,
лучшая модель a_k^* , обученная на выборке P^{s*} .

Начало

```

for w in [1, w_max]:
    for s in [1, w]:
        Ps ← сформировать_патчи(X(c), w, s)
        for ak in A:
            ak.обучить(Ps_train)
            Q_val ← оценить_качество(ak, Ps_test)
            сохранить(w, s, ak, Q_val)
(w*, s*, ak*) ← argmax(Q_val)
ak*.обучить(Ps*_full)
Конец
  
```

Рис. 2. Псевдокод предлагаемого метода

Предложенный подход направлен на реализацию самоорганизующихся самообучающихся систем машинного обучения. На вход системы поступают помеченные последовательности временных рядов сетевого трафика, из которых в автоматическом режиме создаются объекты наблюдения с разными свойствами и формируются выборки данных. Далее решается задача обучения с учителем, оцениваются достигаемые значения функционала качества на каждой выборке и определяются свойства объектов наблюдения и модели, где достигнуты лучшие значения показателей качества.

Таким образом, предлагаемый метод автоматизирует формирование и выбор патчей на основе анализа качественных показателей моделей, что позволяет выбирать модели и способы формирования обучающих последовательностей для достижения лучших результатов.

Эффективность применения метода, описанного в статье, напрямую зависит от соблюдения ряда условий. Качество обработки информационных последовательностей моделями определяется их свойствами. В ходе формирования объектов наблюдения необходимо выбирать размер патча таким образом, чтобы в нем содержались информативные признаки. Для успешного применения предлагаемого метода анализируемые последовательности должны обладать различными паттернами или свойствами. При этом формируемые векторы объектов наблюдения обязаны включать признаки, коррелирующие с целевой переменной, описывающей состояние. Хотя большинство моделей обработки могут обнаруживать такую информацию на основе фактических значений, тем не менее для повышения качества в ряде случаев при обработке патчей временных рядов целесообразно добавлять статистическую информацию [22]. Требуемыми характеристиками для нее является простота вычислений, нечувствительность к преобразованиям и результирующая производительность.

3 Эксперимент

В рамках эксперимента исследовалась возможность автоматического формирования выборок данных из информационных последовательностей для оценки состояния. Рассматриваемые значения иллюстрирующих обработку данных предлагаемым методом получены синтетическим методом.

Основная цель заключалась в анализе влияния свойств объектов наблюдения и получаемых из них выборок данных на качественные показатели моделей. Самоорганизующаяся самообучающаяся модель получала на вход временные ряды сетевого трафика, полученные в различных состояниях. Пример таких рядов представлен на рисунке 3. Из них изменением длины и сдвига формировались объекты наблюдения, из которых формировались выборки данных. Затем они подавались на вход моделям обучения и происходила оценка достигаемых показателей качества.

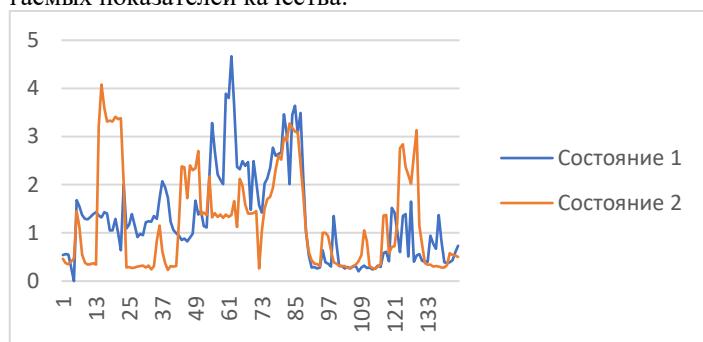


Рис. 3. Последовательности различных состояний

В ходе эксперимента при оценке достигаемых показателей качества моделей учитывались ресурсные ограничения. Проводилось обучение и оценка качественных показателей моделей рекуррентной нейронной сети с долговременной кратковременной памятью LSTM (архитектура: 2 слоя LSTM (50 нейронов), полносвязный слой (10 нейронов), выходной слой (2 нейрона); функция активации: ReLU; оптимизатор: Adam (learning rate=0,001); epochs: 50; batch size: 32), рекуррентной нейронной сети RNN (архитектура: 1 слой RNN (30 нейронов), полносвязный слой (10 нейронов), выходной слой (2 нейрона); функция активации: tanh; оптимизатор: SGD (learning rate=0,01); epochs: 50; batch size: 32) и метода kNN-DTW (k=5; метрика расстояния: DTW с ограничением окна (radius=10)).

В качестве оцениваемых значений определялись точность (accuracy), точность (precision), полнота, F-мера:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$precision = \frac{TP}{TP + FP} \quad (4)$$

$$recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (6)$$

Оценка эффективности предлагаемого метода проводилась при решении задачи бинарной классификации, где сетевой трафик рассматривался в двух состояниях «нормальное» и «аномальное». Сформированные выборки разделялись на обучающую и тестовую части в соотношения 90:10. Длина объектов наблюдения, представленных подпоследовательностями, варьировалась в диапазоне от 2 до 17 отсчетов, а количество сдвигов для каждой длины соответствовало самой этой длине.

На рисунке 4 показано значение accuracy (3) для выбранных классифицирующих моделей обработки данных.

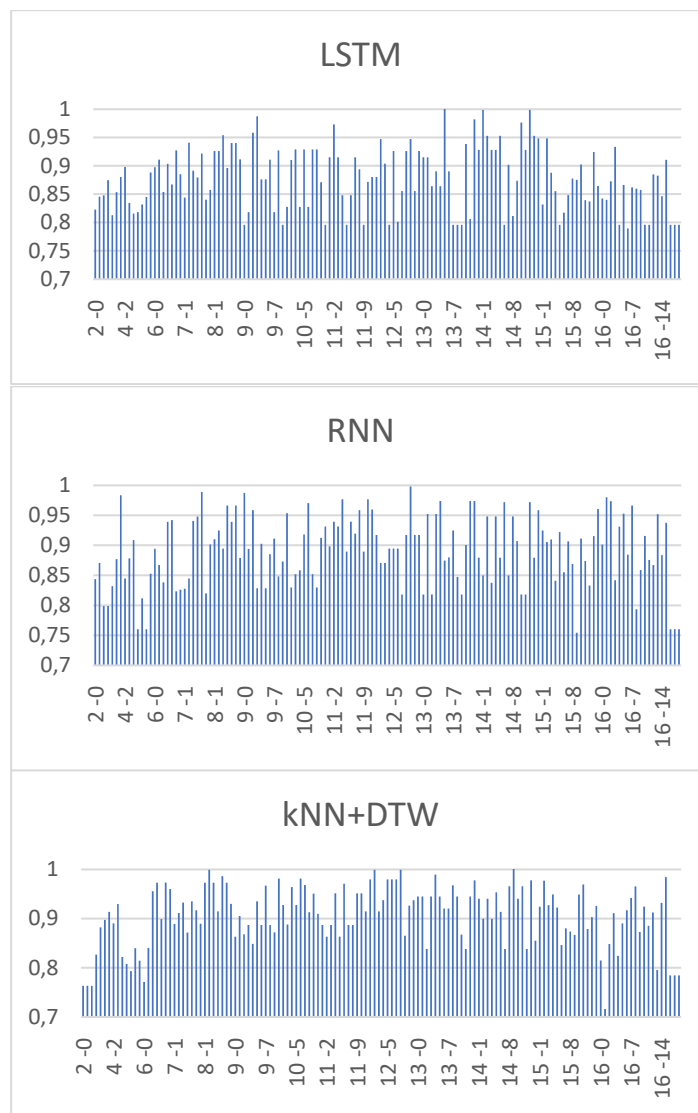


Рис. 4. Значения accuracy (ось ординат) по совмещенной оси «длина патча-сдвиг» (ось абсцисс)

Графики рисунка 3 показывают влияние значений длины и сдвига при формировании объектов наблюдения временных рядов сетевого трафика на достигаемые показатели accuracy. На одной и той же длине при разном сдвиге могут получаться разные значения точности.

На рисунке 5 приведены лучшие значения (2) accuracy (3) для выбранных моделей, указаны длина и сдвиг патча на которых они достигнуты.

Изменение сдвига и длины окна меняет состав признаков внутри формируемого объекта наблюдения, что влияет на работу алгоритма. В окна временных рядов сетевого трафика могут попадать разные аномальные значения (выбросы), которых не было при меньших размерах, моменты смены состояния системы, участки с разными шумовыми компонентами, паттерны, характерные для других состояний, неполные циклы событий, которые не укладываются целиком в окно.

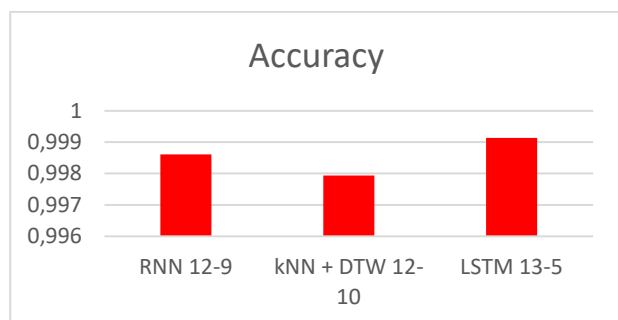


Рис. 5. Лучшие значения точности accuracy RNN, LSTM, kNN+DTW

Может проявляться проблема несовпадения границ окна и границ реальных событий, когда аномалии только частично попадают в окно, из-за чего их признаки искажаются и теряются часть важных характеристик. Изменяются характеристики последовательностей, искажаются корреляции между признаками, меняется соотношение «сигнал/шум», ухудшая видимость значимых паттернов временных рядов сетевого трафика.

В таблице 1 приведены лучшие значения показателей качества (3)-(6) на обработанном датасете для выбранных моделей.

Таблица 1

Лучшие значения показателей качества (3)-(6) на обработанном датасете

Модель	LSTM	RNN	kNN + DTW
Precision	0,857242	0,849506	0,858265
длина-сдвиг	14-5	12-9	12-2
Recall	0,811715	0,802622	0,814916
длина-сдвиг	14-5	12-9	12-3
F1-мера	0,833858	0,825399	0,83602895
длина-сдвиг	14-5	12-9	12-3
Accuracy	0,998607	0,997939	0,999135
длина-сдвиг	13-5	12-9	12-10

Различные значения показателей качества при изменении параметров длины и сдвига могут получаться ввиду чувствительности к локальному распределению данных, где небольшое смещение изменяет набор ближайших соседей и, соответственно, результат классификации. Нейросетевые модели способны по-разному интерпретировать частично перекрывающиеся паттерны и незначительный сдвиг способен изменить входные данные для слоев нейронов. Для модели kNN + DTW может происходить потеря информативных паттернов и включение шума.

В результате автоматического формирования объектов наблюдения временных рядов сетевого трафика изменением длины и сдвига образуются области классов, имеющих разные свойства. Компактность образуемых объектами разных классов областей их разделенность оказывает влияние на достигаемые показатели качества различными функциями. На рисунке 6 приведены средний коэффициент силуэта и индекс Дэвиса-Болдуина при изменении параметра «длина-сдвиг».



Рис. 6. Значения среднего коэффициента силуэта и индекс Дэвиса-Болдуина (ось ординат) по совмещенной оси «длина патча-сдвиг» (ось абсцисс)

Графики рисунка 6 позволяют оценить получаемые на различных длинах объектов наблюдения области классов. Лучшие значения разделения достигаются при длине объектов наблюдения 12 и 14 отчетов, где значения среднего коэффициента силуэта показывают хорошее разделение. Так же высокая компактность кластеров, когда объекты внутри кластера близки к его центру, и хорошая разделимость прослеживается на этих же длинах 12 и 14 в индексе Дэвиса-Болдуина.

Длина последовательности временного ряда сетевого трафика оказывает влияние на обобщающую способность моделей обработки данных. Длина патча может оказывать существенное влияние на достигаемые значения показателей качества обучения модели. Слишком короткие патчи не позволяют выявить информативные признаки и увеличивают количество трудноинтерпретируемых объектов наблюдения, находящихся на границе классов. Длинные патчи содержат большое количество слабо влияющих на разделение признаков, что увеличивает размерность пространства и вычислительные затраты.

Формирование объектов наблюдения сдвигом подпоследовательности можно рассматривать как метод аугментации данных. Он увеличивает число обучающих примеров, улучшает обобщающую способность модели и учит её распознавать признаки независимо от их позиции в последовательности. При перекрытии фрагментов, когда размер сдвига меньше длины патча, модель анализирует один и тот же признак в разных контекстах, что помогает улавливать локальные зависимости и игнорировать фоновый шум.

При формировании выборок из временных рядов сетевого трафика часть объектов может оказаться неинформативной или шумовой, что снижает качество классификации. Оптимальное определение длины патча и размера сдвига позволяет выделить значимые признаки. Однако изменение сдвига может привести к появлению схожих примеров, относящихся к разным классам, создавая дисбаланс в данных и затрудняя обучение модели. Значительное количество признаков может ухудшить результаты классификации из-за снижения качества обработки при наличии лишних, слабо влияющих на модель признаков.

Таким образом, в процессе подготовки обучающей выборки важно учитывать баланс между количеством и качеством признаков для успешной классификации.

4 Выводы

Формирование признакового пространства, выявление структур данных является важным этапом создания моделей машинного обучения.

Предлагаемое решение автоматической генерации объектов наблюдения временных рядов сетевого трафика позволяет создавать выборки и оценивать их эффективность используя анализ значений функционала качества. В отличие от классических подходов, в рассматриваемом методе происходит не настройка и оптимизация модели на выборке данных, а, наоборот, объекты наблюдения формируются из последовательности «под модель». Оценка достигаемого значения функционала качества модели дает возможность определить их свойства.

Достижимое качество моделей машинного обучения определяется не только архитектурой используемой модели, но и характеристиками входных данных. Существенное влияние на итоговые результаты также оказывают параметры объектов наблюдения. Их оптимальная настройка повышает различимость классов, увеличивает долю эталонных данных и снижает уровень шума. Кроме того, предложенный подход может использоваться в качестве метода аугментации.

Однако успешное применение описанного в статье метода обусловлено соблюдением ряда условий. Свойства информационных последовательностей оказывают существенное влияние на качество их обработки моделями. В процессе формирования объектов наблюдения возможны ситуации, когда размер патча оказывается недостаточным для выделения паттерна, содержащего значимую информацию, либо, напротив, чрезмерно большим. В результате паттерн включает избыточное количество шумовых компонент, что негативно сказывается на качестве обработки.

Предложенный метод имеет ряд существенных ограничений. Для его эффективного применения анализируемые последовательности должны иметь паттерны или свойства, позволяющие их различать. Формируемые вектора объектов наблюдения должны содержать признаки, зависящие от целевой переменной, описывающей состояние. В противном случае может возникать эффект «случайности», не позволяющий создать адекватную модель, что потребует дополнительного анализа на поиск дополнительных влияющих факторов, формирования и вычисления новых признаков подпоследовательностей.

В случае соблюдения ограничений предложенное решение позволяет повысить качество обработки данных, и использовать менее ресурсоемкие модели при обработке сложных структур.

Литература

1. Wang C., Li X., Zhou T., Cai Z. Unsupervised Time Series Segmentation: A Survey on Recent // *Advances Computers, Materials & Continua*, 2024, vol. 80, no.2, pp. 2657-2673. doi:10.32604/cmc.2024.054061.
2. Golzari Oskouei A., Samadi N., Khezri S., Najafi Moghaddam A., Babaei H., Hamini K., Fath Nojavan S., Bouyer A., Arasteh B. Feature-weighted fuzzy clustering methods: An experimental review // *Neurocomputing*, 2025, vol. 619, art. 129176. doi:10.1016/j.neucom.2024.129176.
3. Schober P., Vetter T. R. Segmented Regression in an Interrupted Time Series Study Design // *Anesthesia and Analgesia*, 2021, vol.132, no. 3, pp. 696–697. doi:10.1213/ANE.0000000000000526
4. 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, art.129612. doi:10.1016/j.neucom.2025.129612.
5. Леохин Ю.Л., Дымкова С.С., Фатхулин Т.Д. Методы машинного обучения в прикладных задачах прогнозирования динамично изменяющихся данных // *T-Comm: Телекоммуникации и транспорт*. 2025. Т. 19, № 8. С. 49-63.
6. Sun C., Wu W., Zhang Z., Li Z., Ji B., Wang C. Time series clustering of dynamical systems via deterministic learning // *International Journal of Machine Learning and Cybernetics*, 2024, vol. 15, pp. 2761-2779.
7. Naveed Anjum, Zohaib Latif, Hongsong Chen. Security and privacy of industrial big data: Motivation, opportunities and challenges // *Journal of Network and Computer Applications*, 2025, vol. 237, art. 04130, <https://doi.org/10.1016/j.jnca.2025.104130>.
8. Лебедев И.С. Адаптивное применение моделей машинного обучения на отдельных сегментах выборки в задачах регрессии и классификации // *Информационно-управляющие системы*. 2022. № 3(118). С. 20-30.
9. 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.
10. Pourafzal A., Fereidunian A., Safarihamid K. Chaotic Time Series Recognition: A Deep Learning Model Inspired by Complex Systems Characteristics // *International Journal of Engineering, Transactions A: Basics* 2023, vol. 36, pp. 1-9.
11. Леохин Ю.Л., Дымкова С.С., Фатхулин Т.Д., Мяличева А. А. Методы прогнозирования дефектов в программных продуктах на основе ретроспективной и текущей информации // *T-Comm: Телекоммуникации и транспорт*. 2026. Том 20. №1. С. 52-59.
12. 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
13. 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.
14. Salehi A., Balasubramanian M. Ddcnet: Deep dilated convolutional neural network for dense prediction // *Neurocomputing*, 2023, vol. 523, pp. 116-129.
15. 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.
16. Woo G., Liu C., Kumar A., Xiong C., Savarese S., Sahoo D. Unified training of universal time series forecasting transformers //

International Conference on Machine Learning (ICML), 2024, pp. 1023-1035.

17. Zhiwu Huang, Xinlong Zhao. A novel multi-scale competitive network for fault diagnosis in rotating machinery // *Engineering Applications of Artificial Intelligence*, 2024, vol. 128, p. 107441.

18. Etemad A., Grover S., Jalali A. Segment, Shuffle, and Stitch: A Simple Layer for Improving Time-Series Representations. // *Advances in Neural Information Processing Systems*, 2024, vol. 37, pp. 4878-4905. doi:10.52202/079017-0158.

19. 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.

20. 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.

21. 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. https://doi.org/10.1007/979-8-8688-1276-7_7

22. Nanopoulos A., Alcock R., Manolopoulos Y. Feature-based classification of time-series data. // *International Journal of Computer Research*, 2001, vol. 10, pp. 49-61.

A METHOD FOR AUTOMATICALLY CONSTRUCTING A TRAINING DATASET OF NETWORK TRAFFIC TIME SERIES IN A SYSTEM STATE CLASSIFICATION PROBLEM

Ilya S. Lebedev, St. Petersburg Federal Research Center of the Russian Academy of Sciences, St. Petersburg, Russia, isl_box@mail.ru

Mikhail E. Sukhoparov, Admiral Makarov State University of Maritime and Inland Shipping, St. Petersburg, Russia,

mail@sukhoparovm.ru

This research is supported by RSF (grant № 25-21-00269, <https://rscf.ru/project/25-21-00269/>)

Abstract

This article presents a method for automatically generating network traffic time series observation objects for classifying the states of information and telecommunications systems. Growing network traffic data volumes, the ambiguity of various information system events, and the high complexity and labor intensity of labeling processes necessitate the development of adaptive data processing methods. Unlike classical approaches, the proposed method does not involve tuning and optimizing a model on a data sample; instead, observation objects are formed from a sequence "under the model." The proposed method automatically generates training samples by automatically splitting the original sequences into patches (subsequences) with variable length and shift. Data samples are formed from these patches. At each iteration, training, evaluation of achieved quality indicators, and selection of patch parameters and models that achieve the best results occur on the resulting samples. An experiment was conducted on a synthetic dataset for binary classification of the system state. It was shown that the use of a shift in the sequence improves generalization ability due to data augmentation, and the optimal patch size balances information content and noise level. The proposed method automates data preparation and enables the selection of optimal patch generation parameters, improves class distinguishability, can be used as a data augmentation method to improve model training, and enables the use of less resource-intensive models for processing complex data structures. The method improves classification quality by an average of 2-8%, reduces computational costs, enhances interpretability, and efficiently processes complex data structures.

Keywords: network traffic, time series, state classification, automatic object generation, machine learning

References

- [1] C. Wang, X. Li, T. Zhou, and Z. Cai, "Unsupervised Time Series Segmentation: A Survey on Recent Advances," *Computers, Materials & Continua*, vol. 80, no. 2, pp. 2657-2673, 2024. doi: 10.32604/cmc.2024.054061.
- [2] A. Golzari Oskouei, N. Samadi, S. Khezri, A. Najafi Moghaddam, H. Babaei, K. Hamini, S. Fath Nojavan, A. Bouyer, and B. Arasteh, "Feature-weighted fuzzy clustering methods: An experimental review," *Neurocomputing*, vol. 619, art. 129176, 2025. doi: 10.1016/j.neucom.2024.129176.
- [3] P. Schober and T. R. Vetter, "Segmented Regression in an Interrupted Time Series Study Design," *Anesthesia and Analgesia*, vol. 132, no. 3, pp. 696-697, 2021. doi: 10.1213/ANE.0000000000000526.
- [4] Y. He, X. Zhang, X. Kong, L. Yao, and Z. Song, "Causality-driven sequence segmentation assisted soft sensing for multiphase industrial processes," *Neurocomputing*, vol. 631, art. 129612, 2025. doi: 10.1016/j.neucom.2025.129612.
- [5] Yu. L. Lyokhin, S. S. Dymkova, and T. D. Fatkhulin, "Machine learning methods in applied forecasting tasks for dynamically changing data," *T-Comm*, vol. 19, no. 8, pp. 49-63, 2025.
- [6] C. Sun, W. Wu, Z. Zhang, Z. Li, B. Ji, and C. Wang, "Time series clustering of dynamical systems via deterministic learning," *International Journal of*

Machine Learning and Cybernetics, vol. 15, pp. 2761-2779, 2024.

[7] N. Anjum, Z. Latif, and H. Chen, "Security and privacy of industrial big data: Motivation, opportunities and challenges," *Journal of Network and Computer Applications*, vol. 237, art. 04130, 2025. <https://doi.org/10.1016/j.jnca.2025.104130>.

[8] I. S. Lebedev, "Adaptive application of machine learning models on individual sample segments in regression and classification tasks," *Information and Control Systems*, no. 3(118), pp. 20-30, 2022.

[9] I. S. Lebedev and M. E. Sukhoparov, "Adaptive Learning and Integrated Use of Information Flow Forecasting Methods," *Emerging Science Journal*, vol. 7, no. 3, pp. 704-723, 2023.

[10] A. Pourafzal, A. Fereidunian, and K. Safarihamid, "Chaotic Time Series Recognition: A Deep Learning Model Inspired by Complex Systems Characteristics," *International Journal of Engineering, Transactions A: Basics*, vol. 36, pp. 1-9, 2023.

[11] Yu. L. Lyokhin, S. S. Dymkova, T. D. Fatkhulin, and A. A. Mialicheva, "Defect prediction methods in software products based on retrospective and current information," *T-Comm*, vol. 20, no. 1, pp. 52-59, 2026.

[12] E. Tallman and M. West, "Bayesian predictive decision synthesis," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 86, no. 2, pp. 340-363, 2024.

[13] I. Lebedev, M. Sukhoparov, V. Semenov, and D. Khasanov, "Adaptive Segmentation of Information Sequences for Machine Learning Modular Regression Models," *Emerging Science Journal*, vol. 9, no. 5, pp. 2420-2438, 2025.

[14] A. Salehi and M. Balasubramanian, "DDCNet: Deep dilated convolutional neural network for dense prediction," *Neurocomputing*, vol. 523, pp. 116-129, 2023.

[15] M. Goswami, K. Szafer, A. Choudhry et al., "MOMENT: A family of open time-series foundation models," in *Proceedings of the 41st International Conference on Machine Learning*, vol. 235, pp. 16115-16152, 2024.

[16] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo, "Unified training of universal time series forecasting transformers," in *International Conference on Machine Learning (ICML)*, pp. 1023-1035, 2024.

[17] Z. Huang and X. Zhao, "A novel multi-scale competitive network for fault diagnosis in rotating machinery," *Engineering Applications of Artificial Intelligence*, vol. 128, p. 107441, 2024.

[18] A. Etemad, S. Grover, and A. Jalali, "Segment, Shuffle, and Stitch: A Simple Layer for Improving Time-Series Representations," in *Advances in Neural Information Processing Systems*, vol. 37, pp. 4878-4905, 2024. doi: 10.52202/079017-0158.

[19] Y. Chen, X. Lu, and Q. Xie, "Collaborative networks of transformers and convolutional neural networks are powerful and versatile learners for accurate 3D medical image segmentation," *Computers in Biology and Medicine*, vol. 164, p. 107228, 2023.

[20] V. Ekambaram, A. Ati, N. Nguyen, P. Sinthong, and J. Kalagnanam, "TSMixer: Lightweight MLP-Mixer model for multivariate time series forecasting," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, pp. 459-469, 2023.

[21] B. V. K. Vishwas and S. R. Macharla, "MOIRAI: A Time Series LLM for Universal Forecasting," in *Time Series Forecasting Using Generative AI*, pp. 183-194, 2025. https://doi.org/10.1007/979-8-8688-1276-7_7.

[22] Nanopoulos, A., R. Alcock, and Y. Manolopoulos, "Feature-based classification of time-series data," *International Journal of Computer Research*, vol. 10, pp.49-61, 2001.

Information about authors:

Ilya S. Lebedev, Ph.D., Dr. Sci., Professor, Chief Scientific Officer, Laboratory of Intelligent Systems, SPIIRAN, SPC RAS, St. Petersburg, Russia. ORCID 0000-0001-6753-2181

Mikhail E. Sukhoparov, Ph.D., Dr. Sci., Assoc. Prof., Professor, Maritime Radio Communications Department, Admiral Makarov SU MIS, St. Petersburg, Russia. ORCID 0000-0003-1798-8257