DOI 10.37882/2223-2966.2026.01.15

РАЗРАБОТКА АЛГОРИТМА ПРЕДПОДГОТОВКИ ДАННЫХ ДЛЯ ПОДСИСТЕМЫ ДЕТЕКТИРОВАНИЯ АНОМАЛИЙ

DEVELOPMENT OF AN ALGORITHM FOR DATA PRE-PREPARATION FOR THE ANOMALY DETECTION SUBSYSTEM

V. Eliseeva
N. Kosyura

Summary. The aim of the study is to develop an algorithm for data pre-preparation for the anomaly detection subsystem, aimed at increasing the stability and accuracy of intelligent analysis in conditions of noisy and heterogeneous input streams. An algorithm proposed that combines the stages of purification, adaptive normalization, feature selection, and nonlinear data transformation into a single processing sequence. The algorithm considers the specifics of streaming and high-dimensional data typical for cybersecurity tasks, industrial monitoring, and network telemetry analysis. The analysis of the principles of the algorithm structure carried out and the choice of the pre-training methods used is justified. It shown that the proposed approach makes it possible to reduce the impact of noise and emissions, increase the information content of the feature space, and create conditions for more stable operation of anomaly detection models. The results obtained can used in the design of analytical subsystems and data processing pipelines as part of modern monitoring and security systems.

Keywords: data pre-preparation, data processing algorithm, anomaly detection, data purification, normalization, recognition selection, machine learning.

Елисеева Виктория Денисовна
ФГБОУ ВО Волгоградский государственный технический университет
viktoriya.eliseeva.2002@mail.ru
Косюра Надежда Александровна
ФГБОУ ВО Волгоградский государственный технический университет
nadejdakna2002@gmail.com

Аннотация. Целью исследования является разработка алгоритма предподготовки данных для подсистемы детектирования аномалий, ориентированной на повышение устойчивости и точности интеллектуального анализа в условиях шумных и неоднородных входных потоков. В работе предложен алгоритм, объединяющий этапы очистки, адаптивной нормализации, отбора признаков и нелинейной трансформации данных в единую последовательность обработки. Алгоритм учитывает специфику потоковых и высокоразмерных данных, характерных для задач кибербезопасности, промышленного мониторинга и анализа сетевой телеметрии. Проведён анализ принципов построения алгоритма и обоснован выбор используемых методов предподготовки. Показано, что предложенный подход позволяет снизить влияние шумов и выбросов, повысить информативность признакового пространства и создать условия для более стабильной работы моделей детектирования аномалий. Полученные результаты могут быть использованы при проектировании аналитических подсистем и конвейеров обработки данных в составе современных систем мониторинга и безопасности.

Ключевые слова: предподготовка данных, алгоритм обработки данных, детектирование аномалий, очистка данных, нормализация, отбор признаков, машинное обучение.

Введение

Современные системы детектирования аномалий широко применяются в задачах кибербезопасности, промышленного мониторинга, анализа сетевого трафика и технической диагностики. В условиях роста объёмов данных и усложнения их структуры ключевым фактором эффективности таких систем становится не столько выбор модели машинного обучения, сколько качество и согласованность предподготовки входных данных. Практика показывает, что значительная часть ошибок классификации и ложных срабатываний обусловлена недостаточно формализованными или фрагментарными процедурами очистки, нормализации и формирования признакового пространства. Несмотря на большое количество исследований, посвящённых отдельным методам предподготовки данных, в большинстве прикладных решений данный этап реализуется в виде набора разрозненных процедур, не объединённых в единый алгоритм. Отсутствие формализованного алгоритма предподготовки приводит к снижению воспроизводимости результатов, усложняет интеграцию аналитических моделей в автоматизированные конвейеры и ограничивает адаптацию систем детектирования аномалий к изменяющимся условиям среды и характеристикам входных потоков данных.

В этой связи актуальной является задача разработки алгоритма предподготовки данных, ориентированного на использование в подсистемах детектирования аномалий и учитывающего специфику многомерных, потоковых и зашумлённых данных. Такой алгоритм должен обеспечивать последовательную обработку входной информации, включающую устранение шумов и выбросов, приведение данных к сопоставимым масштабам, отбор информативных признаков и снижение размерности при сохранении значимых закономерностей. Целью настоящей работы является разработка алгоритма пред-

подготовки данных для подсистемы детектирования аномалий и обоснование его структуры с позиции повышения устойчивости и эффективности аналитических моделей. В рамках исследования формируется логика алгоритма, определяются ключевые этапы обработки данных и их взаимосвязь, а также обозначаются направления практического применения предложенного решения в системах мониторинга и безопасности.

Результаты и обсуждение

Вопросы предподготовки данных и отбора признаков в задачах детектирования аномалий рассматриваются в современных исследованиях как один из ключевых факторов эффективности интеллектуального анализа данных и машинного обучения. По мнению А.Н. Линдигрина (2021), корректность предварительной обработки сетевых данных, включающей фильтрацию шумов, устранение выбросов и балансировку классов, в значительной мере определяет предельные возможности аналитических моделей [1]. Автор подчёркивает, что неочищенные и слабо структурированные журналы событий приводят к росту ложных срабатываний и снижению надёжности систем обнаружения аномалий. Как отмечает В.С. Никулин (2020), сложность предподготовки существенно возрастает при работе с большими и неоднородными потоками данных, характерными для распределённых вычислительных систем [2]. В таких условиях требуется системный подход, включающий многоступенчатую фильтрацию, нормализацию и адаптивное преобразование признаков. По результатам его исследований можно полагать, что внедрение интеллектуальных методов на этапе предподготовки позволяет повысить достоверность диагностики и снизить количество ложных отклонений.

По результатам работы П.Ю. Гусева и А.В. Таволжанского (2024), процесс подготовки данных для задач предиктивной аналитики целесообразно рассматривать как итеративный алгоритм, в котором все этапы от очистки до масштабирования и балансировки должны быть формализованы и автоматизированы [3]. Авторы подчёркивают значимость гибридных подходов, сочетающих статистические методы и алгоритмы машинного обучения. Сходной позиции придерживается И.А. Попова (2022), указывая, что обучаемые модели могут использоваться не только для анализа, но и для коррекции пропусков и шумов в исходных данных, формируя основу самонастраивающихся систем предподготовки [4]. Исследование A. Boukerche, L. Zheng и O. Alfandi (2020) показывает, что методы обнаружения выбросов являются фундаментальным элементом предподготовки данных в системах детектирования аномалий [5]. Авторы предлагают таксономию стратегий обнаружения выбросов и отмечают ограниченность классических методов при работе с потоковыми и высокоразмерными данными. В этой связи особое внимание уделяется применению нейронных сетей и кластерных моделей для автоматической фильтрации шумов и выбросов.

В работах, посвящённых сетевой безопасности, подчёркивается необходимость учёта адверсариальных воздействий уже на этапе предподготовки данных. Так, E. Alhajjar, P. Maxwell и N. Bastian (2021) демонстрируют, что специально модифицированные входные данные способны снижать эффективность моделей NIDS, несмотря на использование сложных алгоритмов машинного обучения [6]. Авторы обосновывают целесообразность применения устойчивых методов предподготовки, включая стохастическое сглаживание и обучение с добавлением шумовых примеров. Широкий обзор методов обнаружения сетевых аномалий представлен и в работе M. Ahmed, A.N. Mahmood и J. Hu (2016), где этап предподготовки данных рассматривается как определяющий для корректной классификации [7]. Авторы выделяют классификационные, статистические, информационно-теоретические и кластеризационные подходы, указывая на проблему отсутствия стандартизированных процедур подготовки данных, что затрудняет сопоставимость результатов различных систем.

Проблема избыточных и нерелевантных признаков подробно рассмотрена M.A. Ambusaidi и соавторами (2016), которые предлагают алгоритм фильтрационного отбора признаков на основе взаимной информации [8]. Эксперименты на наборах данных KDD Cup 99 и NSL-KDD показывают, что применение метода FMIFS позволяет повысить точность классификации и сократить вычислительные затраты без потери качества анализа. Пример интеграции предподготовки данных в архитектуру реального времени представлен в работе Y. Dong, R. Wang и J. He (2019), где очистка, кодирование и нормализация данных выполняются в потоковом режиме с использованием Flume и Flink [9]. Использование автоэнкодеров для снижения размерности позволило достичь точности детектирования 94,32 %, однако авторы отмечают высокие вычислительные затраты и сложность обучения моделей.

Проблемы качества обучающих данных также подчёркнуты в исследовании M. Ring, S. Wunderlich и коллег (2019), где проведён анализ 34 сетевых датасетов [10]. Авторы показывают, что несбалансированность классов и недостаточное описание реальных сценариев атак существенно ограничивают эффективность моделей. Аналогичные выводы делают H. Hadian Jazi и соавторы (2017), указывая, что корректная выборка данных критична для обнаружения DoS-атак на уровне приложений [11]. Для задач снижения размерности Y. Hamid и M. Sugumaran (2020) предлагают использовать метод t-SNE, который позволяет выявлять сложные взаимосвязи между признаками и повышать эффективность класси-

фикаторов типа SVM [12]. Несмотря на рост точности, авторы отмечают значительные вычислительные затраты данного подхода.

Так, анализ представленных источников показывает, что развитие систем детектирования аномалий тесно связано с совершенствованием методов предподготовки данных. Современные тенденции направлены на переход от разрозненных процедур к формализованным и автоматизированным алгоритмам, интегрированным в аналитические конвейеры и MLOps-процессы. Данные выводы формируют основу для разработки алгоритма предподготовки данных, ориентированного на универсальное применение в подсистемах детектирования аномалий с возможностью адаптации под специфику конкретных задач. На следующем этапе исследования, опираясь на результаты проведённого анализа научных источников, осуществляется разработка алгоритма предподготовки данных для подсистемы детектирования аномалий. В основу алгоритма закладываются методы очистки, нормализации, отбора и трансформации признаков, выбор которых обосновывается с точки зрения их вычислительной эффективности, устойчивости к шумам и способности сохранять информативность данных.

В рамках исследования проводится сравнительный анализ производительности отдельных методов предподготовки, по результатам которого формируется универсальная модель алгоритма, допускающая адаптацию к различным прикладным задачам за счёт параметрических и структурных корректировок. Разработка такого алгоритма представляет собой основной результат работы и направлена на обеспечение воспроизводимости, масштабируемости и практической применимости подсистем детектирования аномалий в условиях разнородных и динамически изменяющихся входных данных. Итак, на основе проведённого анализа современных методов предподготовки данных целесообразным является переход от разрозненных процедур очистки и трансформации к формализованному алгоритму, ориентированному на использование в подсистемах детектирования аномалий. Разрабатываемый алгоритм должен обеспечивать последовательную обработку данных, учитывать специфику высокоразмерных и потоковых входных массивов, а также допускать адаптацию под различные прикладные области при сохранении общей структуры. В данном разделе предлагается алгоритм предподготовки данных, объединяющий ключевые этапы обработки в единую логическую схему и ориентированный на повышение устойчивости и воспроизводимости результатов детектирования аномалий (рисунок 1).

Входные данные:

1) сырые данные D, поступающие из журналов событий, телеметрии, датчиков или сетевых потоков.

Выходные данные:

2) подготовленный набор данных D', пригодный для обучения и функционирования подсистемы детектирования аномалий.

Работа разработанного алгоритма предподготовки данных начинается с приёма входных массивов и их первичной валидации, в ходе которой анализируется структурная целостность данных, корректность форматов и наличие обязательных атрибутов. На этом этапе формируется базовое представление о качестве входного потока и принимается решение о необходимости углублённой очистки. Далее алгоритм последовательно выполняет очистку данных от дублирующихся и некорректных записей, а также обработку пропусков с использованием статистических или обучаемых методов восстановления, что позволяет сохранить объём и информативность данных. После стабилизации структуры данных осуществляется выявление и обработка выбросов, способных исказить обучение моделей детектирования аномалий; в зависимости от плотности и размерности данных применяются статистические или кластерные подходы с возможностью адаптивной перенастройки порогов. На следующем этапе признаки приводятся к сопоставимым масштабам путём нормализации или стандартизации, что обеспечивает корректную работу алгоритмов машинного обучения и устойчивость модели при изменении распределений входных данных. Завершающая часть алгоритма направлена на формирование информативного признакового пространства за счёт отбора релевантных атрибутов и, при необходимости, нелинейной трансформации данных, после чего формируется итоговый набор данных, готовый для обучения и эксплуатации подсистемы детектирования аномалий.

Для формального описания задачи разработки алгоритма предподготовки данных в подсистеме детектирования аномалий целесообразно представить процесс обработки входной информации в виде отображения исходного пространства данных в преобразованное признаковое пространство, обеспечивающее повышение устойчивости и точности аналитических моделей. Формализация позволяет определить цель предподготовки данных, критерии качества и ограничения, накладываемые на процесс обработки.

Пусть исходный набор данных представляется в виде матрицы наблюдений:

$$X = \left\{ x_{ij} \right\}, i = 1,\ldots,N, j = 1,\ldots,M,$$

где N — количество наблюдений, а

M — число исходных признаков, характеризующих состояние анализируемой системы.

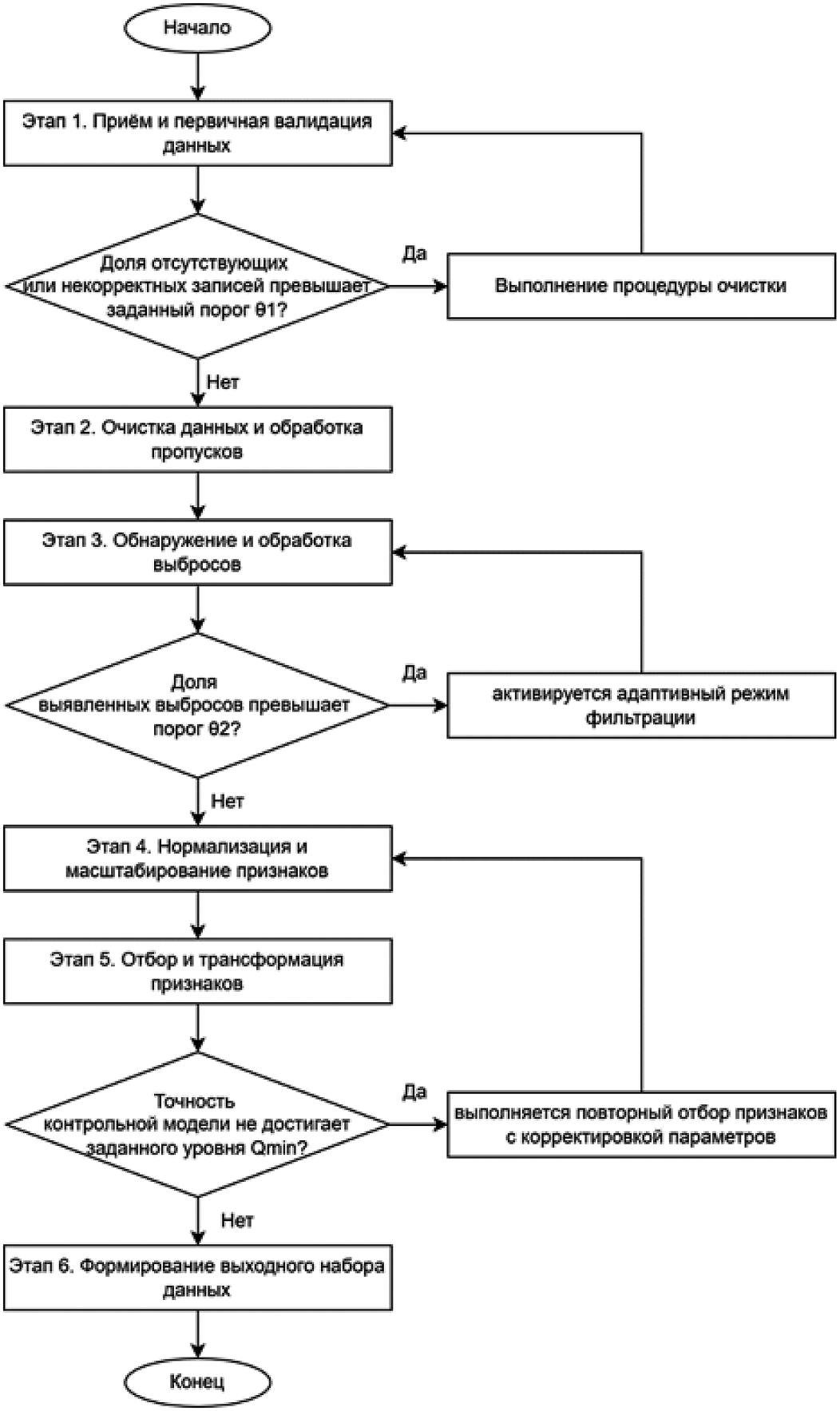

Рис. 1. Алгоритм предподготовки данных для подсистемы детектирования аномалий

Входные данные могут содержать шумы, выбросы и пропуски, что снижает эффективность алгоритмов детектирования аномалий. Процесс предподготовки данных задаётся в виде композиции последовательных преобразований:

$$X' = F_k \circ F_{k-1} \circ \ldots \circ F_1(X),$$

где F_1 — оператор очистки и обработки пропусков,

F_2 — оператор обнаружения и фильтрации выбросов,

F_k — оператор нормализации, отбора и трансформации признаков.

Результатом применения данных преобразований является набор данных X', обладающий улучшенными статистическими и структурными свойствами. Цель предподготовки данных формализуется как задача оптимизации качества детектирования аномалий:

$$\max_F Q(A(X')),$$

где где A — алгоритм детектирования аномалий,

$Q(\cdot)$ — функция качества, характеризующая точность, устойчивость или полноту обнаружения аномальных событий.

Так, алгоритм предподготовки данных рассматривается как самостоятельный объект оптимизации, обеспечивающий повышение эффективности последующего анализа за счёт формирования устойчивого и информативного признакового пространства. Для практической реализации разработанного алгоритма предподготовки данных необходимо выбрать конкретные методы обработки, соответствующие каждому этапу алгоритма. Выбор методов осуществлялся на основе анализа их вычислительной эффективности, устойчивости к шумам и выбросам, а также влияния на качество детектирования аномалий. С целью обоснования выбора был проведён сравнительный анализ наиболее распространённых и показательных методов предподготовки данных, применяемых в задачах обнаружения аномалий.

1. Выбор методов для этапов алгоритма.

На этапе очистки данных и обработки пропусков для анализа были выбраны методы удаления записей с пропусками, статистической аппроксимации (среднее и медиана) и обучаемой реконструкции на основе k-ближайших соседей (kNN). Данные методы различаются по вычислительной сложности и степени сохранения структуры данных.

Для этапа обнаружения и обработки выбросов рассматривались статистический метод на основе межквартильного размаха (IQR), алгоритм Local Outlier Factor (LOF) и кластерный подход на основе DBSCAN. Данные методы позволяют оценить устойчивость алгоритма при работе с выбросами различной природы.

На этапе нормализации и масштабирования признаков анализировались методы Min–Max-нормализации, Z-нормализации и робастного масштабирования, устойчивого к выбросам.

Для этапа отбора и трансформации признаков рассматривались корреляционный фильтр, метод отбора на основе взаимной информации (MI) и нелинейное снижение размерности с использованием автоэнкодеров.

2. Методика проведения эксперимента.

Экспериментальная оценка проводилась на основе стандартных наборов данных, применяемых в задачах детектирования аномалий, включая NSL-KDD и синтетические выборки с контролируемым уровнем шума и выбросов. Для оценки влияния методов предподготовки использовалась фиксированная модель детектирования аномалий (изолирующий лес), что позволило исключить влияние архитектуры модели на результаты эксперимента.

В качестве количественного показателя использовалась метрика F1-score, отражающая баланс между точностью и полнотой обнаружения аномалий, а также среднее время обработки одного пакета данных. Каждый эксперимент повторялся несколько раз с последующим усреднением результатов. Результаты количественного анализа методов предподготовки данных представлены в таблице 1.

Таблица 1.
Сравнительная оценка методов предподготовки данных

Этап алгоритма	Метод	F1-score	Время обработки, мс
Обработка пропусков	Удаление записей	0,78	12
	Среднее значение	0,82	15
	kNN-реконструкция	0,86	41
Обнаружение выбросов	IQR	0,81	18
	LOF	0,87	56
	DBSCAN	0,85	63
Нормализация	Min–Max	0,83	9
	Z-нормализация	0,85	11
	Робастное масштабирование	0,86	14
Отбор признаков	Корреляционный фильтр	0,84	7
	Взаимная информация	0,88	19
	Автоэнкодер	0,89	74

Для комплексной оценки эффективности выбранных методов предподготовки данных выполнена совмещённая визуализация результатов экспериментального анализа, отражающая как качество детектирования аномалий, так и вычислительные затраты применяемых подходов. На рисунке 2 представлены значения показателя F1-score (рис. 2, а) и среднего времени обработки данных (рис. 2, б), что позволяет наглядно сопоставить точностные и временные характеристики методов и выявить компромиссные решения, наиболее подходящие для использования в составе универсального алгоритма предподготовки данных.

Анализ данных, представленных на рисунке 2а, показывает, что наибольшие значения показателя F1-score достигаются при использовании интеллектуальных методов предподготовки данных, таких как автоэнкодеры, методы отбора признаков на основе взаимной информации и алгоритм LOF. В то же время рисунок 2б. демонстрирует, что указанные методы характеризуются повышенными вычислительными затратами по сравнению с простыми статистическими и фильтрационными подходами. Минимальное время обработки обеспечивают корреляционный фильтр и методы линейной нормализации, однако их применение сопровождается умеренным снижением качества детектирования аномалий.

Итак, результаты эксперимента показывают, что более простые методы предподготовки обладают высокой вычислительной эффективностью, однако уступают по качеству детектирования аномалий более сложным интеллектуальным подходам. В частности, kNN-реконструкция и LOF демонстрируют наилучшие значения F1-score, но сопровождаются ростом вычислительных затрат. Аналогичная тенденция наблюдается при использовании автоэнкодеров для отбора и трансформации признаков. С учётом необходимости универсальности и масштабируемости алгоритма в качестве базовых были выбраны методы, обеспечивающие компромисс между точностью и вычислительной сложностью, с возможностью их замены на более сложные при повышенных требованиях к качеству анализа.

3. Итоговые рекомендации по выбору методов предподготовки данных.

По результатам работы были сформированы следующие рекомендации для их применения в разработанном алгоритме предподготовки данных для подсистемы детектирования аномалий (таблица 2):

Проведённый анализ и экспериментальная оценка методов предподготовки данных позволили обосновать выбор конкретных методов для каждого этапа разработанного алгоритма. Полученные результаты подтверждают, что оптимальная предподготовка данных должна строиться на компромиссе между вычислительной эффективностью и качеством детектирования аномалий. Сформированная совокупность методов обеспечивает универсальность алгоритма и допускает его адаптацию под различные прикладные задачи за счёт параметрической настройки и замены отдельных этапов обработки. Это позволяет рассматривать предложенный алгоритм как практическую основу для построения подсистем детектирования аномалий в условиях разнородных и динамически изменяющихся данных.

Таблица 2.

Рекомендуемые методы для этапов алгоритма предподготовки данных

Этап алгоритма	Базовый метод	Альтернативный метод
Обработка пропусков	Статистическая аппроксимация	kNN-реконструкция
Обнаружение выбросов	IQR	LOF
Нормализация признаков	Z-нормализация	Робастное масштабирование
Отбор признаков	Взаимная информация	Автоэнкодер
Снижение размерности	Корреляционный фильтр	Нелинейные методы

Заключение

В работе рассмотрена и решена задача повышения эффективности систем детектирования аномалий за счёт разработки алгоритма предподготовки данных, ориентированного на обработку разнородных, зашумлённых и высокоразмерных входных потоков. На основе анализа современных научных исследований показано, что этап предподготовки данных является критически значимым для устойчивости и точности интеллектуальных моделей, при этом в большинстве существующих решений он реализуется фрагментарно и без формализованного алгоритмического описания. В рамках исследования предложен структурированный алгоритм предподготовки данных, включающий этапы валидации, очистки, обработки выбросов, нормализации и отбора признаков, а также обоснован выбор методов для каждого этапа на основе сравнительного анализа точности и вычислительных затрат. Полученные результаты подтверждают, что формализованный и адаптивный подход к предподготовке данных позволяет обеспечить воспроизводимость, масштабируемость и практическую применимость подсистем детектирования аномалий в различных предметных областях. Выводы по результатам работы:

— разработан алгоритм предподготовки данных для подсистемы детектирования аномалий, обладающий модульной структурой и допускающий адаптацию к различным типам данных и условиям эксплуатации;

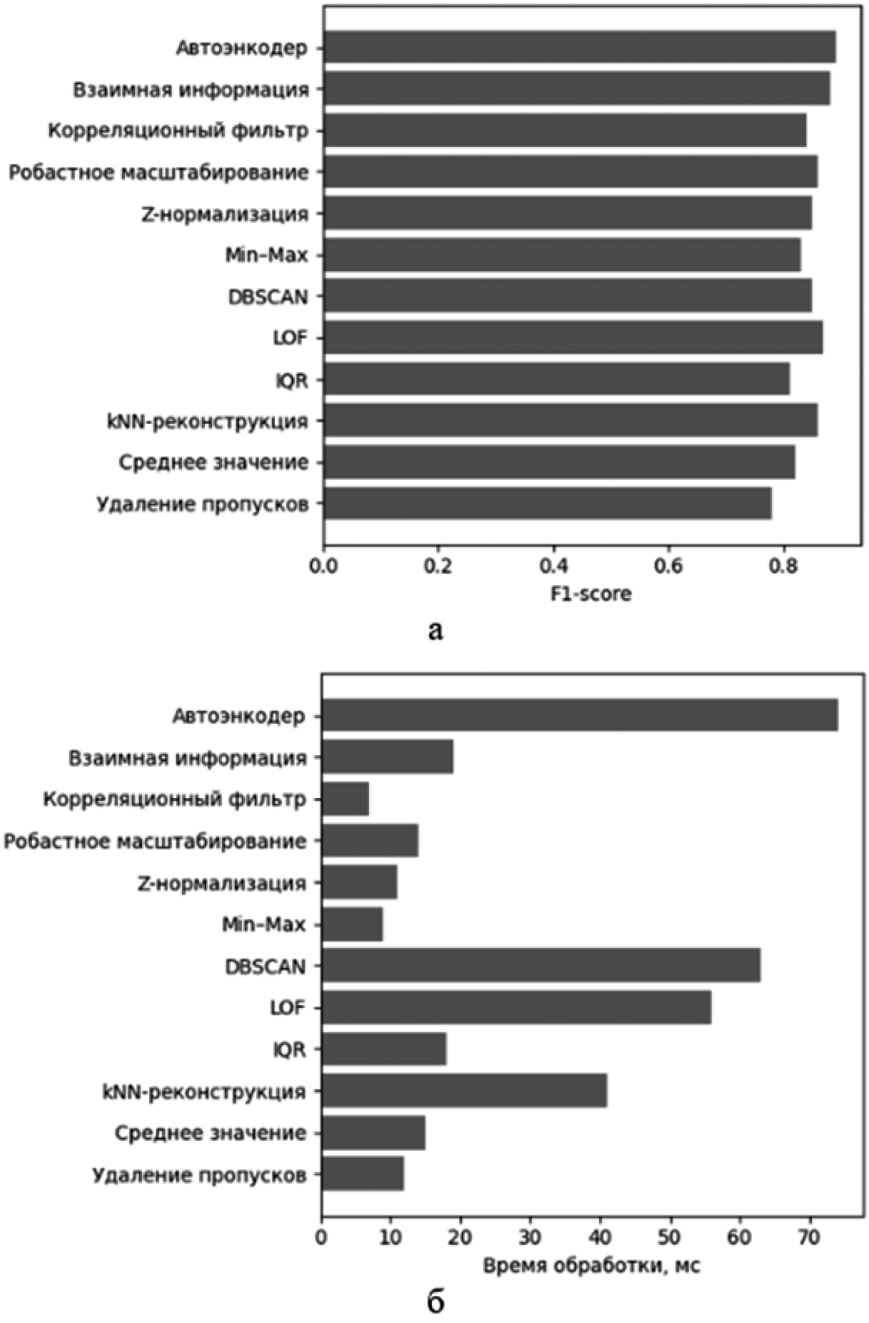

Рис. 2. Сравнительный анализ эффективности и вычислительных затрат методов предподготовки данных (а — сравнение методов по показателю F1-score, б — сравнение методов по времени обработки данных)

— проведён сравнительный анализ методов предподготовки данных, позволивший количественно оценить их влияние на качество детектирования аномалий и вычислительные затраты;
— обоснован выбор базовых и альтернативных методов для каждого этапа алгоритма, обеспечивающих компромисс между точностью анализа и быстродействием системы.

Перспективы дальнейших исследований связаны с расширением экспериментальной базы за счёт использования дополнительных наборов данных и моделей детектирования аномалий, а также с интеграцией предложенного алгоритма в реальные аналитические конвейеры и MLOps-среды. Практическая ценность работы заключается в возможности использования разработанного алгоритма в качестве универсальной основы для построения подсистем детектирования аномалий в задачах кибербезопасности, промышленного мониторинга и анализа сетевых данных, что позволяет повысить надёжность и устойчивость интеллектуальных систем в условиях роста объёмов и сложности обрабатываемой информации.

ЛИТЕРАТУРА

1. Линдигрин А.Н. Анализ специфики и проблематики процессов поиска аномалий в сетевых данных // Известия ТулГУ. Технические науки. 2021. №5. С. 304–309.
2. Никулин В.С. Методика подготовки данных для интеллектуального анализа надежности вычислительных комплексов // Вестник СибГУТИ. 2020. №3 (51). С. 26–37.
3. Гусев П.Ю., Таволжанский А.В. Алгоритмизация обработки и подготовки данных для построения моделей предиктивной аналитики // Вестник ВГТУ. 2024. №2. С. 14–19.
4. Попова И.А. Исследование алгоритмов машинного обучения для предварительной обработки данных в задачах регрессии // Искусственный интеллект в автоматизированных системах управления и обработки данных: Сборник статей Всероссийской научной конференции. В 2-х томах Москва. 2022. С. 237–242.
5. Boukerche A., Zheng L., Alfandi O. Outlier Detection: Methods, Models, and Classification // ACM Computing Surveys (CSUR). 2020. Vol. 53, No. 3. Article No. 55. P. 1–37.
6. Alhajjar E., Maxwell P., Bastian N. Adversarial Machine Learning in Network Intrusion Detection Systems // Expert Systems with Applications. 2021. Vol. 186. Article No. 115782. 30 December.
7. Ahmed M., Mahmood A.N., Hu J. A Survey of Network Anomaly Detection Techniques // Journal of Network and Computer Applications. 2016. Vol. 60. P. 19–31.
8. Ambusaidi M.A., He X., Nanda P., Tan Z. Building an Intrusion Detection System Using a Filter-Based Feature Selection Algorithm // IEEE Transactions on Computers. 2016. Vol. 65, No. 10. P. 2986–2998.
9. Dong Y., Wang R., He J. Real-Time Network Intrusion Detection System Based on Deep Learning // Proceedings of the IEEE International Conference on Software Engineering and Service Sciences (ICSESS). 2019. P. 1–4.
10. Ring M., Wunderlich S., Scheuring D., Landes D., Hotho A. A Survey of Network-Based Intrusion Detection Data Sets // Computers & Security. 2019. Vol. 86. P. 147–167.
11. Hadian Jazi H., Gonzalez H., Stakhanova N., Ghorbani A.A. Detecting HTTP-Based Application Layer DoS Attacks on Web Servers in the Presence of Sampling // Computer Networks. 2017. Vol. 121. P. 25–36.
12. Hamid Y., Sugumaran M. A t-SNE Based Non-Linear Dimension Reduction for Network Intrusion Detection // International Journal of Information Technology (Singapore). 2020. Vol. 12, No. 1. P. 125–134.