МЕТОД АНАЛИЗА СОСТОЯНИЙ РАСПРЕДЕЛЁННЫХ СИСТЕМ НА ОСНОВЕ ТЕЛЕМЕТРИЧЕСКИХ ДАННЫХ И МАШИННОГО ОБУЧЕНИЯ

Котько Е.Н.[1], магистрант, e.kotko@bsuir.by

2026

1. Белорусский государственный университет информатики и радиоэлектроники

Ключевые слова: телеметрические данные, распределённые системы, предиктивный мониторинг, машинное обучение, прогнозирование отказов, Random Forest, временные ряды.

Аннотация: Разработан метод анализа состояний распределённых систем на основе телеметрических данных и машинного обучения. Предложен подход к преобразованию потоковых метрик в обучающее пространство с использованием скользящего окна наблюдений, позволяющий учитывать временную динамику системы. Для классификации применяется алгоритм Random Forest, устойчивый к шуму и нелинейным зависимостям метрик. Результатом является формирование датасета для задач предиктивного анализа отказов и оценки состояния распределённой инфраструктуры.

Введение. Современные распределённые системы характеризуются высокой динамичностью и значительным объёмом телеметрических данных, формируемых в процессе взаимодействия сервисов и обработки пользовательских запросов [1]. В таких условиях традиционные подходы мониторинга, ориентированные на фиксацию текущего состояния инфраструктуры, не обеспечивают достаточной эффективности для выявления ранних стадий деградации и прогнозирования отказов.

Существующие системы наблюдаемости в основном ограничиваются сбором и визуализацией метрик, не используя накопленные телеметрические данные для построения моделей предиктивного анализа. Это создаёт разрыв между мониторингом и задачами машинного обучения, что снижает возможности интеллектуального анализа состояния распределённой инфраструктуры.

В развитие ранее выполненного исследования [2] предложен подход к интеграции мониторинга и машинного обучения за счёт формирования обучающего набора на основе потоковых телеметрических данных и моделирования деградационных сценариев. Реализовано преобразование временных рядов в признаковое пространство с использованием скользящего окна наблюдений, обеспечивающее учёт временной динамики состояния системы.

Архитектура обработки телеметрических данных распределённой системы. Разработана многоуровневая архитектура обработки телеметрических данных, обеспечивающая полный цикл: от сбора и агрегации метрик до формирования обучающего набора и прогнозирования состояния распределённой системы. Подход расширяет классическую наблюдаемость за счёт использования телеметрии как источника данных для задач предиктивного анализа [3].

На рисунке 1 представлена многоуровневая архитектура обработки телеметрических данных распределённой системы.

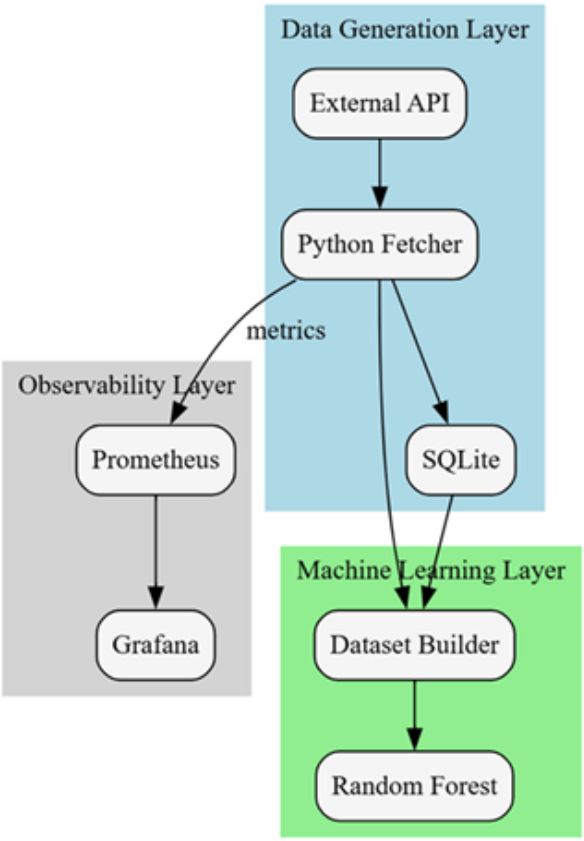

Рисунок 1 – Многоуровневая архитектура системы обработки телеметрических данных и формирования обучающего набора

Система включает три взаимосвязанных уровня, формирующих единый конвейер обработки данных.

Observability Layer реализует сбор и визуализацию телеметрии с использованием Prometheus и Grafana, фиксируя ключевые метрики состояния системы (latency, errors, resource usage) [4].

Data Generation Layer обеспечивает формирование потоковых данных и моделирование деградационных сценариев (сетевые задержки, ошибки API, перегрузка БД, утечки памяти) с сохранением данных в структурированное хранилище [5].

Machine Learning Layer выполняет преобразование временных рядов в признаки с использованием скользящего окна и формирует обучающий набор для классификации состояния системы. Для анализа применяется алгоритм Random Forest, устойчивый к шуму и нелинейным зависимостям метрик [6].

Предложенная архитектура объединяет мониторинг, генерацию данных и машинное обучение в единый конвейер обработки телеметрии обеспечивая переход от наблюдения состояния к его прогнозированию.

Математическая модель состояния системы. Состояние распределённой системы формализуется как многомерный вектор телеметрических метрик:

$$X(t) = \{latency, errors, db\_errors, requests, memory\_usage\}, \quad (1)$$

где *latency* – задержка обработки запросов; *errors* – количество ошибок; *db_errors* – ошибки базы данных; *requests* – число запросов; *memory_usage* – использование памяти.

Для учёта временной динамики используется скользящее окно наблюдений [7] длиной k:

$$\bar{X}(t) = \{X(t-k), \ldots, X(t)\}, \quad (2)$$

где k – размер окна (число шагов наблюдения).

На рисунке 2 представлена модель формирования агрегированного состояния системы на основе скользящего временного окна, обеспечивающего объединение последовательности наблюдений в единое контекстное представление [8].

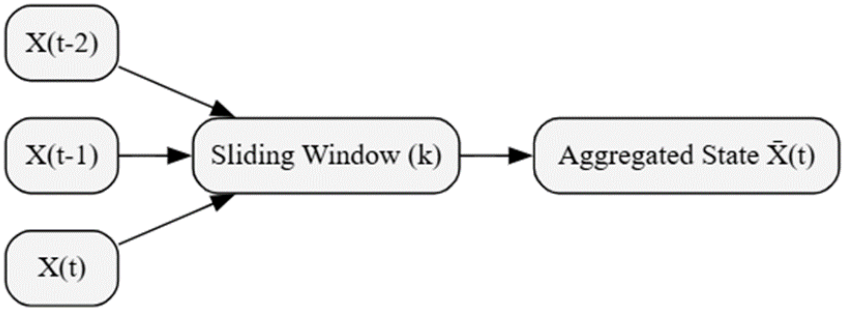

Рисунок 2 – Модель формирования агрегированного состояния системы на основе скользящего временного окна

Предложенный подход учитывает временные зависимости между наблюдениями и формирует устойчивое представление состояния распределённой системы для последующего анализа и прогнозирования.

Формирование датасета. Обучающий набор формируется на основе потоковых телеметрических данных распределённой системы мониторинга. Для преобразования временных рядов в признаки используется скользящее окно фиксированной длины k=5, позволяющее учитывать локальную временную динамику состояния системы [9].

Внутри окна агрегируются ключевые метрики: latency, errors, db_errors, requests и memory_usage. На их основе формируется признаковое

представление, включающее усреднённые и суммарные характеристики временного интервала.

Целевая переменная определяется по пороговой модели деградации: состояние системы считается предаварийным при превышении критических значений не менее чем по двум метрикам, иначе – нормальным [9].

Для расширения вариативности данных используется контролируемая генерация отказов, моделирующая типовые сценарии деградации (сбои API, ошибки базы данных, сетевые задержки, утечки памяти), что позволяет сформировать сбалансированные траектории поведения системы.

Сформированный датасет представляет собой временно-структурированное описание состояния системы и используется для обучения и тестирования моделей машинного обучения.

Модель машинного обучения. Для прогнозирования состояния распределённой системы используется алгоритм Random Forest, обеспечивающий устойчивость к шуму, нелинейным зависимостям и коррелированности телеметрических признаков [10].

Бинарная классификация формализуется как: 0 – нормальное состояние, 1 – предаварийное.

Обучение модели выполняется на датасете, сформированном на основе скользящего окна и агрегированных телеметрических признаков. Разделение выборки осуществляется в пропорции 70/30 с фиксацией random_state для воспроизводимости результатов.

Модель представляет собой ансамбль из 100 решающих деревьев, а итоговое решение формируется методом голосования, что снижает влияние шумовых и локальных аномалий в данных.

На рисунке 3 представлен полный цикл формирования обучающей выборки и обучения модели Random Forest, включающий этапы подготовки данных, обучения и получения предсказаний.

Оценка качества модели выполняется с использованием стандартных метрик: accuracy характеризует общую точность классификации, precision – точность выявления предаварийных состояний, recall – полноту их обнаружения, а confusion matrix позволяет анализировать структуру ошибок классификации, включая ложноположительные и ложноотрицательные срабатывания.

Ключевая особенность подхода заключается в обучении модели не на статических данных, а на временных агрегированных состояниях системы, что позволяет учитывать динамику деградационных процессов.

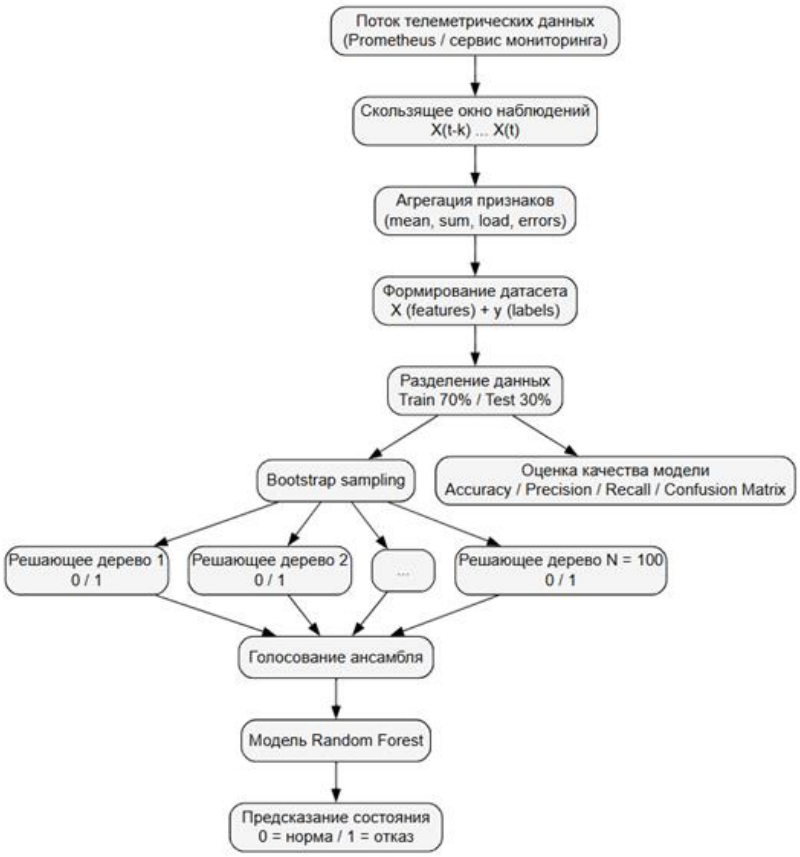

Рисунок 3 – Полный процесс формирования обучающей выборки и обучения модели Random Forest

Итоговая модель сохраняется в формате joblib и используется в режиме потокового прогнозирования состояния распределённой инфраструктуры.

Заключение. Разработан метод анализа состояний распределённых систем на основе телеметрических данных и машинного обучения, ориентированный на выявление деградационных процессов в условиях высокой динамичности инфраструктуры. Предложенный подход обеспечивает преобразование потоковых телеметрических данных в признаковое пространство с использованием скользящего окна наблюдений, что позволяет учитывать временную структуру поведения системы.

Ключевым результатом работы является формализация состояния распределённой системы в виде обучаемого многомерного представления, пригодного для задач классификации и предиктивного анализа. В отличие от классических подходов наблюдаемости, метод обеспечивает использование мониторинговых данных не только для контроля, но и для построения моделей оценки состояния.

Применение алгоритма Random Forest обеспечивает устойчивую классификацию состояний системы при наличии шумов, нелинейных зависимостей и коррелированных телеметрических метрик. Это позволяет выявлять предаварийные состояния на основе комплексного анализа поведения системы.

Практическая значимость метода заключается в возможности интеграции предложенного подхода в системы предиктивного мониторинга распределённой инфраструктуры, обеспечивающие переход от реактивного реагирования к проактивному управлению отказами.

Список использованных источников

1. Станолевич, В. С. Современные подходы к мониторингу ит-инфраструктуры в условиях распределенных и облачно-ориентированных архитектур / В. С. Станолевич, К. А. Батенков // International Journal of Humanities and Natural Sciences. – 2026. – . – № vol. 2-1 (113). – С. 254
2. Пискун, Е. С. Анализ производительности и устойчивости системы мониторинга платформы электронной коммерции на основе Prometheus и Grafana / Е. С. Пискун, Е. Н. Котько // Системный анализ и прикладная информатика. – 2025. – № 3. – С. 29–33.
3. Тед, Я. Изучаем OpenTelemetry. Современный мониторинг систем / Я. Тед, П. Остин. – Астана : Sprint Book, 2025. – 240 с.
4. Чарити, М. Обеспечение наблюдаемости ПО / М. Чарити, Ф. Лиз, М. Джордж. – Астана : Sprint Book, 2025. – 352 с.
5. Радченко И.А, Николаев И.Н. Технологии и инфраструктура Big Data. – СПб: Университет ИТМО, 2018. – 52 с.
6. Рашка, С. Машинное обучение с PyTorch и Scikit-Learn / С. Рашка, Ю. Лю, В. Мирджалили. – Астана : Фолиант, 2024. – 688 с.
7. Скользящее окно [Электронный ресурс]. – Режим доступа: https://basegroup.ru/deductor/function/algorithm/sliding-window. – Дата доступа: 08.05.2026.
8. Тихомиров, В. А. Процессная модель формирования агрегированных требований к сложным информационным системам / В. А. Тихомиров, А. В. Пушина // Программные продукты и системы. – 2010. – . – Т. 2. – С. 91
9. Гульчеев, В. А. Секреты датасетов: практическое руководство по анализу и обработке данных / В. А. Гульчеев. – Москва : ЛитРес : Самиздат, 2023. – 42 с.
10. RandomForestClassifier [Электронный ресурс]. – Режим доступа: https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Дата доступа: 02.05.2026.