**Онлайн-детектирование аномалий во временных рядах геофизических параметров на основе локальной линейной регрессии**

***Караваева К.Э.*1*,***

1*студент*

*Московский государственный университет имени М.В.Ломоносова,
физический факультет, Москва, Россия
E–mail*: *karavaeva.ke21@physics.msu.ru*

Современные системы мониторинга окружающей среды генерируют огромные объемы геофизических данных, поступающих в реальном времени с датчиков, измеряющих температуру, давление и другие параметры. Выявление аномалий в этих временных рядах играет ключевую роль в раннем обнаружении неисправностей оборудования и анализе климатических изменений.

Традиционные методы анализа часто предполагают постобработку данных, что снижает их эффективность для оперативного реагирования. В связи с этим возрастает необходимость в разработке методов онлайн-детектирования аномалий, способных работать в потоковом режиме, обрабатывая поступающие данные без задержек.

В данной работе рассматриваются современные подходы к онлайн-анализу временных рядов геофизических параметров, включая статистические методы, алгоритмы машинного обучения и гибридные решения. Оценивается их применимость к различным типам геофизических данных, анализируется точность и быстродействие.

Измерения приборов представляют собой временные ряды. Временной ряд – это последовательность наблюдений за определенным параметром в разные моменты времени. Принято классифицировать аномалии на точечные, коллективные и контекстуальные [1]. В этой работе будут рассматриваться точечные и групповые аномалии. Групповую аномалию будет определять как последовательность точечных аномалий.

Любой временной ряд можно представить в виде суммы слабо или медленно меняющейся части $m\_{t}$ (тренда), периодически изменяющейся циклической части $S\_{t}$ (сезонность) и остатка $ε\_{t}$

$f\left(t\right)=m\_{t}+S\_{t}+ε\_{t}.$ (1)

Далее мы будем рассматривать временные ряды температуры. Для таких рядов обычно характерны годовые и дневные колебания и линейный тренд.

Помимо работы с реальными геофизическими данными, важным этапом исследования является использование синтетических временных рядов. Они позволяют моделировать различные сценарии поведения геофизических параметров, контролировать характеристики данных и изучать работу алгоритмов детектирования аномалий в условиях, где реальные данные могут быть ограничены или зашумлены.

Синтетические ряды обеспечивают гибкость в выборе параметров, таких как уровень шума, амплитуда и частота колебаний, а также наличие заранее заданных аномалий. Это дает возможность тестировать и калибровать методы обнаружения аномалий на чистых данных, а затем адаптировать их к реальным измерениям. Кроме того, анализ синтетических данных позволяет глубже понять поведение алгоритмов и их устойчивость к различным видам шумов и выбросов, что особенно важно для онлайн-детектирования аномалий в потоках геофизических данных.

Временные ряды геофизических параметров часто имеют квазипериодическую природу и могут напоминать синусоидальные колебания. Это позволяет применять линейную регрессию для предсказания значений ряда на коротких интервалах

В работе предложен метод, который заключается в построении модели линейной регрессии на локальном участке ряда, меньшем, чем полный период колебаний. Она обучается на предыдущих значениях ряда и используется для прогнозирования будущих точек. Затем окно данных сдвигается по ряду, позволяя анализировать его в динамике. Если разница между предсказанным и фактическим значением превышает заданный порог, точка считается аномальной.

Длина обучающей выборки и длина прогноза в таком методе являются гиперпараметрами, которые подбираются в зависимости от характеристик конкретного временного ряда. Оптимальный выбор этих параметров позволяет учитывать особенности ряда, снижать влияние шума и повышать точность обнаружения аномалий. Такой подход особенно полезен для онлайн-обработки данных, поскольку он требует малых вычислительных ресурсов и может работать в потоковом режиме.

Кроме того, в данной работе исследуется зависимость метрик качества модели от выбора гиперпараметров – длины обучающей выборки и длины прогноза. Эти параметры оказывают значительное влияние на точность предсказания и, соответственно, на эффективность обнаружения аномалий.

Если длина обучающей выборки слишком мала, модель может недостаточно хорошо захватывать структуру временного ряда, что приведет к увеличению ошибок предсказания. С другой стороны, слишком большая обучающая выборка может сглаживать локальные особенности данных и снижать чувствительность к аномалиям. Аналогично, длина прогнозного интервала влияет на достоверность предсказания: слишком короткий прогноз может быть неинформативным, тогда как длинный прогноз приводит к накоплению ошибки.

В работе проводится серия экспериментов, в которых оценивается влияние этих гиперпараметров на метрики качества модели, такие как среднеквадратичная ошибка (MSE), средняя абсолютная ошибка (MAE) и полнота детектирования аномалий. Результаты анализа позволяют выбрать оптимальные значения параметров для конкретных типов временных рядов и обеспечить наилучший баланс между точностью предсказания и скоростью вычислений.

**Литература**

1. Schmidl S., Wenig P., Papenbrock T. Anomaly detection in time series: a comprehensive evaluation // Proceedings of the VLDB Endowment. 2022. V. 15. P. 1779-1797.
2. Cook A. A., Mısırlı G., Fan Z. Anomaly Detection for IoT Time-Series Data: A Survey // IEEE Internet of Things Journal. 2020. V. 7. N. 7. P. 6481-6494.
3. Kuchibhotla A. K. et al. All of linear regression // arXiv preprint arXiv:1910.06386. 2019.
4. ShipmonD. T., Gurevitch J. M., Piselli P. M., Edwards S. T. Time Series Anomaly Detection; Detection of anomalous drops with limited features and sparse examples in noisy highly periodic data // arXiv:1708.03665. 2017