

УДК 004.89

СОВРЕМЕННЫЕ ТЕХНОЛОГИИ ДЕТЕКТИРОВАНИЯ АНОМАЛИЙ В БОЛЬШИХ ДАННЫХ

Гришкевич М.М.¹, студент гр.НДДб-22-1, 3 курс

¹Иркутский национальный исследовательский университет, г.Иркутск

Аномалии (выбросы) — это наблюдения, которые значительно отклоняются от ожидаемого поведения данных. Их обнаружение играет ключевую роль в различных областях: [1]

- Финансы (обнаружение мошеннических транзакций)
- Кибербезопасность (выявление вторжений)
- Промышленность (прогнозирование отказов оборудования)
- Здравоохранение (диагностика редких заболеваний)
- Телекоммуникации (обнаружение сбоев в сетях)
- Розничная торговля (выявление аномальных паттернов покупок)
- Транспорт и логистика (мониторинг отклонений в маршрутах)

С ростом объема и сложности данных традиционные методы (например, статистические тесты) становятся менее эффективными, что стимулирует развитие новых технологий на основе машинного обучения (МО) и искусственного интеллекта (ИИ). Особую актуальность приобретают методы, способные:

- Обработать данные в реальном времени (поточная аналитика)
- Работать с высокоразмерными и неструктурированными данными
- Адаптироваться к изменяющимся распределениям данных (концептуальный дрейф)
- Обеспечивать интерпретируемость результатов для принятия решений

Современные системы обнаружения аномалий все чаще используют гибридные подходы, сочетающие: [2]

- Статистические методы для быстрой первичной фильтрации
- Алгоритмы машинного обучения для сложных зависимостей
- Методы глубокого обучения для работы с неструктурированными данными
- Техники объяснимого ИИ (XAI) для интерпретации результатов
- Особые вызовы представляют:
- Несбалансированность данных (аномалии редки)

2.1.2

- Концептуальный дрейф (изменение паттернов со временем)
- Необходимость минимизировать ложные срабатывания
- Требования к производительности в реальном времени:

Классификация методов детектирования аномалий:

- Статистические методы. Z-score (Метод основан на стандартном отклонении)

$$z = \frac{x - \mu}{\sigma}$$

где x — значение наблюдения,
 μ — среднее значение,
 σ — стандартное отклонение.

Точки с $|z| > 3$ обычно считаются аномальными.

- Межквартильный размах (IQR): Аномалии определяются за пределами:

$$[Q_1 - k \cdot IQR, Q_3 + k \cdot IQR]$$

где Q_1, Q_3 — первый и третий квартили,
 $IQR = Q_3 - Q_1$
 k — коэффициент (обычно 1.5).

- Машинное обучение. Isolation Forest. Алгоритм строит деревья, изолируя аномалии за меньшее число разбиений:

$$\sin(x, n) = 2^{\frac{E(h(x))}{c(n)}}$$

где $h(x)$ — глубина изоляции объекта
 $c(n)$ — средняя глубина для n точек,
 $s \approx 1$ — аномалия, $s \approx 0$ — норма.

- One-Class SVM. Оптимизирует гиперплоскость в пространстве признаков:

$$w \cdot \phi(x_i) \geq \rho - \xi_i, \quad \xi_i \geq 0$$

где $\phi(x)$ — функция ядра

- Глубокое обучение
- Рекуррентные нейронные сети (LSTM) – для временных рядов.

GAN (Generative Adversarial Networks) – генеративные модели для сложных распределений [3].

- Transformer-модели – обработка последовательностей и мультимодальных данных.

• Комбинация нескольких методов (например, Isolation Forest + Autoencoder) повышает точность и устойчивость к шуму.

- Инструменты и фреймворки
- Python-библиотеки: Scikit-learn, PyOD, TensorFlow, PyTorch.

- Big Data-решения: Apache Spark (MLlib), Kafka (потокковая аналитика).
- Специализированные платформы: Elastic Stack (ELK), Splunk, Datadog.
- Проблемы и перспективы:
 1. Масштабируемость – необходимость эффективной обработки данных в реальном времени.
 2. Интерпретируемость – объяснимость моделей (XAI) для критически важных решений [4].
 3. Адаптивность – обнаружение новых типов аномалий (zero-day атаки, дрейф данных).

Современные технологии детектирования аномалий эволюционируют в сторону более сложных и адаптивных моделей, сочетающих методы машинного и глубокого обучения. Однако ключевыми остаются вопросы масштабируемости и интерпретируемости, особенно в условиях больших и быстро меняющихся данных [5, 6].

Список литературы:

1. Чандела В., Банерджи А., Кумар В. Обнаружение аномалий: обзор методов [Anomaly detection: A survey] // ACM Computing Surveys. - 2009. - Т. 41, № 3. - С. 1-58. - DOI: 10.1145/1541880.1541882.
2. Зайцев А.В. Большие данные: технологии обработки и анализа. - СПб.: БХВ-Петербург, 2020. - 512 с. - ISBN 978-5-9775-4321-0.
3. Пидгурская, Е. С. Влияние цифровых технологий на психическое здоровье: возможности и риски // Донецкие чтения 2024: образование, наука, инновации, культура и вызовы современности : материалы IX Международной научной конференции, Донецк, 15–17 октября 2024 года
4. Барсегян А.А. и др. Методы и модели анализа данных. - М.: Финансы и статистика, 2019. - 432 с. - ISBN 978-5-279-03456-7
5. Hawkins S. et al. Outlier Detection Using Replicator Neural Networks // Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery. - 2002. - P. 170-180.
6. Рафф Л. и др. Глубокие методы одноклассовой классификации [Deep One-Class Classification] // Journal of Machine Learning Research. - 2021. - Т. 22. - С. 1-73