УДК 004.89

СОВРЕМЕННЫЕ ПОДХОДЫ К АНАЛИЗУ АНОМАЛИЙ В СТРУКТУРИРОВАННЫХ И НЕСТРУКТУРИРОВАННЫХ ДАННЫХ

Алексеева Д.К., студент гр. БКНАД222, II курс
Национальный исследовательский университет «Высшая школа экономики»,
г. Москва

Анализ аномалий является важной задачей в различных областях, таких как кибербезопасность, финансы, медицина и интернет вещей. В современном мире данные могут быть как структурированными, так и неструктурированными, что требует применения различных методов для их анализа. В этой статье рассмотрим современные подходы к анализу аномалий в обоих типах данных [1].

Анализ аномалий в структурированных данных:

1. Машинное обучение.

Машинное обучение является одним из наиболее распространенных подходов к анализу аномалий в структурированных данных. Ниже перечислены несколько методов:

- Кластеризация (Clustering): Алгоритмы, такие как K-means и DBSCAN, используются для выявления кластеров в данных. Объекты, которые не принадлежат ни одному кластеру или принадлежат к малочисленным кластерам, могут рассматриваться как аномалии.
- Методы на основе плотности (Density-Based Methods): Алгоритмы, такие как Local Outlier Factor (LOF), оценивают плотность вокруг точки и сравнивают её с плотностью соседних точек. Объекты с существенно более низкой плотностью считаются выбросами.
- Ансамблевые методы (Ensemble Methods): Методы, такие как Isolation Forest, строят множество решающих деревьев, которые изолируют объекты. Чем быстрее объект изолируется, тем более вероятно, что он является аномалией [2].

2. Статистические методы.

Статистические методы основываются на предположениях о распределении данных:

- Методы на основе Z-оценки (Z-Score): Оценка Z-score измеряет, насколько далеко каждое наблюдение от среднего значения. Наблюдения с высокими абсолютными значениями Z-score считаются аномалиями.

- Методы на основе гауссовского распределения (Gaussian Distribution): Предполагается, что данные следуют нормальному распределению, и аномалии идентифицируются как наблюдения, находящиеся за пределами определенного количества стандартных отклонений от среднего значения.

3. Байесовские методы. Байесовские методы, такие как Байесовские сети и наивный байесовский классификатор, используются для моделирования вероятностных зависимостей между переменными и выявления аномалий на основе отклонений от ожидаемых вероятностей.

Анализ аномалий в неструктурированных данных:

1. Обработка естественного языка (NLP)

Для анализа аномалий в текстовых данных используются методы NLP:

- Векторизация текста: Методы, такие как TF-IDF и Word2Vec, преобразуют текст в числовые векторы, которые затем анализируются на наличие аномалий с использованием алгоритмов машинного обучения [3].
- Теммные модели (Topic Modeling): Алгоритмы, такие как Latent Dirichlet Allocation (LDA), используются для выявления тем в текстах. Документы, которые не соответствуют ни одной из тем, могут рассматриваться как аномалии.

2. Анализ изображений

Для анализа аномалий в изображениях применяются методы компьютерного зрения:

- Глубокое обучение (Deep Learning): Сверточные нейронные сети (CNN) используются для извлечения признаков из изображений и выявления аномалий на основе их отклонений от нормальных паттернов.
- Autoencoders: Автоэнкодеры обучаются восстанавливать изображения и могут выявлять аномалии, основываясь на различиях между исходными и восстановленными изображениями.

3. Анализ звуковых данных

Для анализа аномалий в звуковых данных используются методы обработки сигналов и глубокого обучения: [4]

- Спектрограммы: Аудио преобразуется в спектрограммы, которые затем анализируются с использованием CNN или других методов глубокого обучения для выявления аномалий.
- Рекуррентные нейронные сети (RNN): Используются для анализа временных последовательностей в аудио и выявления аномалий на основе отклонений от нормальных паттернов.

Современные инструменты и технологии. Современные инструменты и платформы для анализа данных, такие как Apache Spark, TensorFlow и PyTorch, предоставляют мощные возможности для обработки больших

объемов структурированных и неструктурированных данных. Эти инструменты позволяют эффективно применять алгоритмы машинного обучения и глубокого обучения для выявления аномалий.

Современные подходы к анализу аномалий в структурированных и неструктурированных данных охватывают широкий спектр методов, от статистических и машинного обучения до передовых методов глубокого обучения. Выбор подходящего метода зависит от характера данных и конкретной задачи. Постоянное развитие технологий и методов анализа данных продолжает расширять возможности для выявления аномалий, что играет ключевую роль в обеспечении безопасности и качества данных во многих областях [5].

Список литературы:

1. Pang, G. Deep learning for anomaly detection: a review / G. Pang, C. Shen, L. Cao, A. van den Hengel // ACM Computing Surveys. – 2021. – Vol. 54, № 2. – Article 38. – DOI 10.1145/3439950.
2. Goldstein, M. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data / M. Goldstein, S. Uchida // PLoS ONE. – 2016. – Vol. 11, № 4. – Article e0152173. – DOI 10.1371/journal.pone.0152173.
3. Emmott, A. Systematic construction of anomaly detection benchmarks from real data / A. Emmott, S. Das, T. Dietterich, A. Fern, W.-K. Wong // Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description. – New York : ACM, 2013. – P. 16–21. – DOI 10.1145/2500853.2500858.
4. Akoglu, L. Graph based anomaly detection and description: a survey / L. Akoglu, H. Tong, D. Koutra // Data Mining and Knowledge Discovery. – 2015. – Vol. 29. – P. 626–688. – DOI 10.1007/s10618-014-0365-y.
5. Zimek, A. A survey on unsupervised outlier detection in high-dimensional numerical data / A. Zimek, E. Schubert, H.-P. Kriegel // Statistical Analysis and Data Mining. – 2012. – Vol. 5, № 5. – P. 363–387. – DOI 10.1002/sam.11161.