

Лекция 12. Методы выявления аномалий и выбросов

Тема: Z-оценка, плотностные методы, автоэнкодеры

1. Введение

В процессе анализа данных часто встречаются **аномалии (outliers, выбросы)** — наблюдения, которые существенно отличаются от большинства данных. Аномалии могут быть результатом ошибок измерений, технических сбоев, мошеннических действий или просто отражать редкие, но важные явления.

Выявление таких объектов — ключевая задача **интеллектуального анализа данных (Data Mining)**, особенно в областях, где критично определять отклонения:

- банковское мошенничество,
 - кибербезопасность,
 - диагностика оборудования,
 - медицинские исследования,
 - контроль качества.
-

2. Определение и типы аномалий

Аномалия (выброс) — это наблюдение, которое не соответствует ожидаемому распределению данных или поведению системы.

Типы аномалий:

1. **Точечные (Point anomalies)** — отдельные экземпляры данных, отличающиеся от других (например, ошибка измерения).
 2. **Контекстные (Contextual anomalies)** — выбросы, зависящие от контекста (например, высокая температура зимой).
 3. **Коллективные (Collective anomalies)** — группы наблюдений, которые вместе представляют аномальное поведение (например, серия подозрительных транзакций).
-

3. Статистические методы: Z-оценка (Z-score)

Z-оценка — классический статистический способ определения выбросов, основанный на нормальном распределении.

Она измеряет, насколько далеко наблюдение отклоняется от среднего значения в единицах стандартного отклонения.

$$Z_i = \frac{x_i - \mu}{\sigma} \quad Z_i = \frac{x_i - \mu}{\sigma}$$

где

- x_i — значение признака,
- μ — среднее значение,
- σ — стандартное отклонение.

Правило:

Если $|Z| > 3$, элемент обычно считается выбросом.

Преимущества:

- простота и наглядность;
- не требует сложных вычислений.

Недостатки:

- чувствителен к ненормальному распределению;
- не работает для многомерных данных.

4. Плотностные методы выявления аномалий

Плотностные методы оценивают, насколько **плотно** данные сгруппированы в пространстве признаков.

Аномалии располагаются в областях **низкой плотности** по сравнению с остальными точками.

4.1. Метод k-ближайших соседей (k-NN Outlier Detection)

Для каждой точки вычисляется среднее расстояние до k ближайших соседей. Если расстояние значительно больше, чем у большинства, точка считается аномальной.

Плюсы:

- интуитивно понятный метод;
- хорошо работает в небольших размерностях.

Минусы:

- чувствителен к выбору k ;

- высокая вычислительная сложность при больших данных.
-

4.2. Метод локальной факторной выбросности (LOF — Local Outlier Factor)

LOF сравнивает **локальную плотность** точки с плотностью её соседей. Если плотность точки значительно меньше, чем у соседей, ей присваивается высокий коэффициент выбросности.

LOF > 1 означает, что точка может быть выбросом.

Преимущества:

- учитывает локальные особенности данных;
- эффективен при наличии кластеров разной плотности.

Недостатки:

- сложность настройки параметров;
 - требует значительных вычислений при больших выборках.
-

5. Модельные и машинные методы: автоэнкодеры

Автоэнкодеры (Autoencoders) — это тип **нейронных сетей**, используемых для обучения **компактного представления данных** (кодирования).

Идея заключается в том, чтобы сеть научилась восстанавливать исходные данные после сжатия.

Структура автоэнкодера:

- **Энкодер** — преобразует входные данные в сжатое представление (код, латентное пространство);
- **Декодер** — восстанавливает данные из этого представления.

$$x' = f(g(x)) \quad x' = f(g(x)) \quad x' = f(g(x))$$

где

$g(x)$ — кодирование,

$f(\cdot)$ — декодирование,

x' — восстановленные данные.

Принцип выявления аномалий:

- сеть обучается на «нормальных» данных;
- при подаче аномального примера ошибка восстановления будет значительно выше;
- если ошибка > порога, объект классифицируется как выброс.

Преимущества:

- подходит для сложных многомерных данных;
- может выявлять нелинейные закономерности;
- хорошо работает с изображениями, временными рядами и звуками.

Недостатки:

- требует большого объема данных для обучения;
- чувствителен к архитектуре и гиперпараметрам.

6. Сравнение методов

Метод	Тип данных	Преимущества	Недостатки
Z-оценка	1D, нормальное распределение	Простота, высокая скорость	Не подходит для многомерных данных
k-NN / LOF	Многомерные	Учитывают локальную структуру	Медленные при больших объёмах данных
Автоэнкодеры	Высокая размерность, сложные данные	Гибкость, адаптивность	Требуют обучения и ресурсов

7. Применение в Data Mining

- **Финансы:** обнаружение мошеннических транзакций;
- **Кибербезопасность:** детектирование сетевых атак и вредоносных действий;
- **Промышленность:** выявление неисправностей оборудования по сенсорным данным;
- **Медицина:** определение патологий на основе медицинских изображений или сигналов;
- **Маркетинг:** поиск необычного поведения клиентов.

8. Заключение

Выявление аномалий — один из важнейших этапов Data Mining, так как даже единичное отклонение может иметь большое значение.

Современные подходы объединяют статистические, плотностные и нейросетевые методы, обеспечивая высокую точность и адаптивность.

В зависимости от природы данных и доступных ресурсов выбираются разные подходы — от простых Z-оценок до мощных автоэнкодеров, способных работать с многомерными и неструктурированными данными.

Список литературы

1. Хэн, Дж., Камбер, М., Пей, Дж. *Интеллектуальный анализ данных: концепции и методы*. — М.: Вильямс, 2019.
2. Aggarwal, C. *Outlier Analysis*. — Springer, 2017.
3. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. — O'Reilly Media, 2022.
4. Chandola, V., Banerjee, A., Kumar, V. *Anomaly Detection: A Survey*. ACM Computing Surveys, 2009.
5. Goodfellow, I., Bengio, Y., Courville, A. *Deep Learning*. — MIT Press, 2016.