Introduction to Anomaly Detection

Fang Zhou

Anomalies and outliers are essentially the same thing:

objects that are different from most other objects

The techniques used for detection are the same.

Anomaly detection

- Historically, the field of statistics tried to find and remove outliers as a way to improve analyses.

- There are now many fields where the outliers / anomalies are the objects of greatest interest.
  - The rare events may be the ones with the greatest impact, and often in a negative way.

Causes of anomalies

- Data from different class of object or underlying mechanism
  - fraud vs. not fraud

- Natural variation
  - tails on a Gaussian distribution

Distinction Between Noise and Anomalies

- Noise doesn’t necessarily produce unusual values or objects

- Noise is not interesting

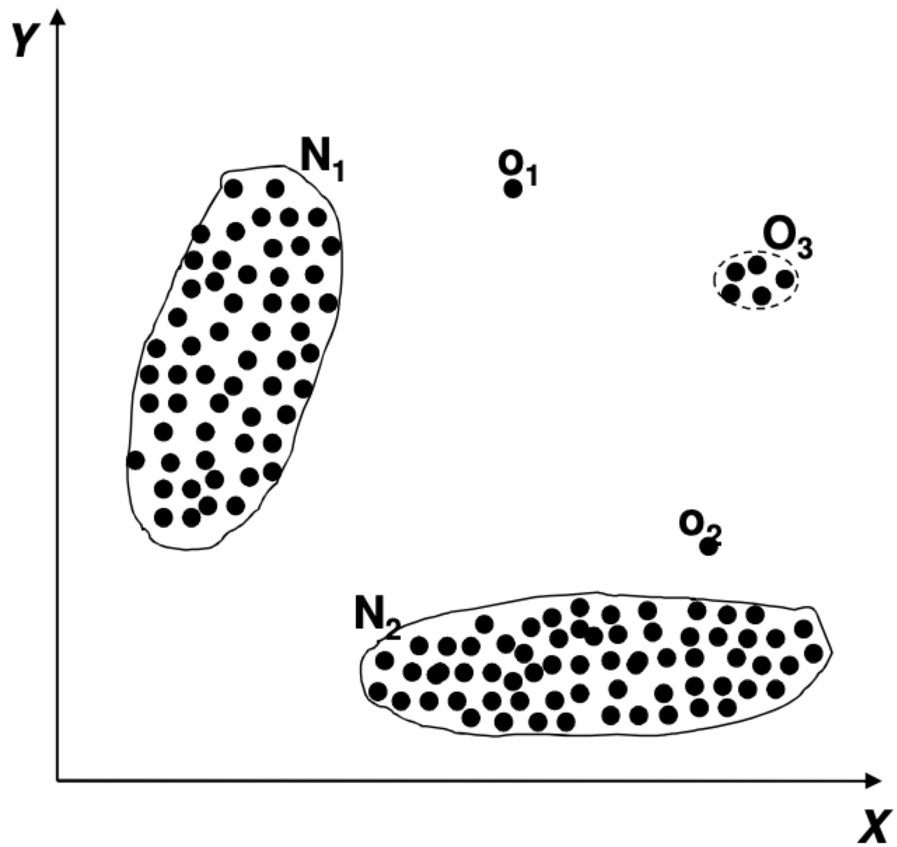
- Noise and anomalies are related but distinct concepts

Structure of anomalies

- Point anomalies
- Contextual anomalies
- Collective anomalies

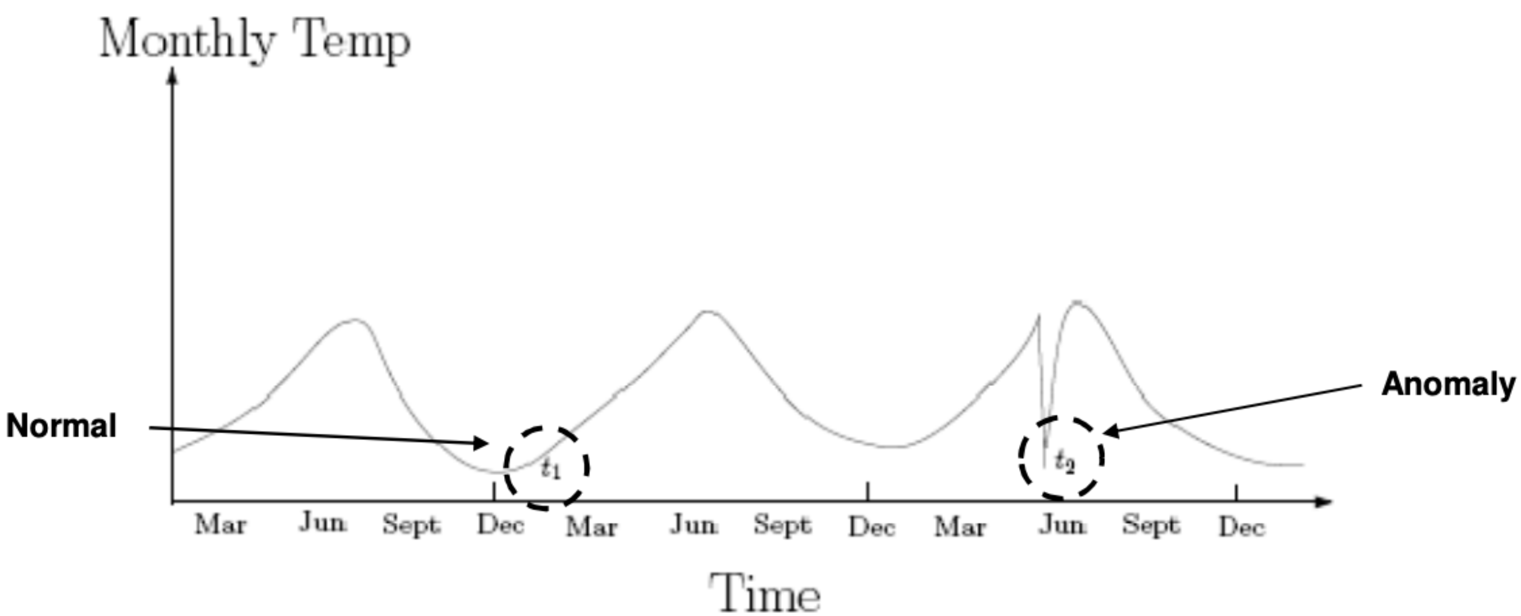
Point anomalies

- An individual data instance is anomalous with respect to the data



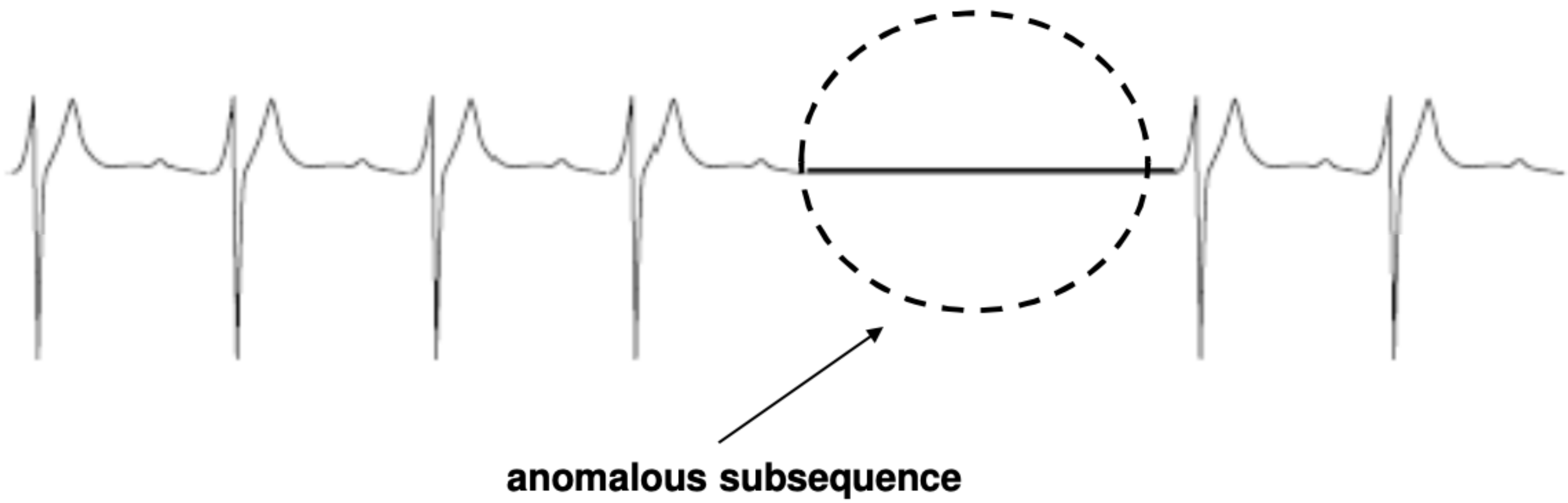
Contextual anomalies

- An individual data instance is anomalous within a context
- Requires a notion of context
- Also referred to as conditional anomalies



Collective anomalies

- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential data
  - Spatial data
  - Graph data
- The individual instances within a collective anomaly are not anomalous by themselves



Applications of anomaly detection

- Network intrusion
- Insurance / credit card fraud
- Healthcare informatics / medical diagnostics
- Industrial damage detection
- Image processing / video surveillance
- Novel topic detection in text mining
- …

Intrusion detection

- Intrusion detection
  - Monitor events occurring in a computer system or network and analyze them for intrusions
  - Intrusions defined as attempts to bypass the security mechanisms of a computer or network
- Challenges
  - Traditional intrusion detection systems are based on signatures of known attacks and cannot detect emerging cyber threats
  - Substantial latency in deployment of newly created signatures across the computer system
- Anomaly detection can alleviate these limitations

Fraud detection

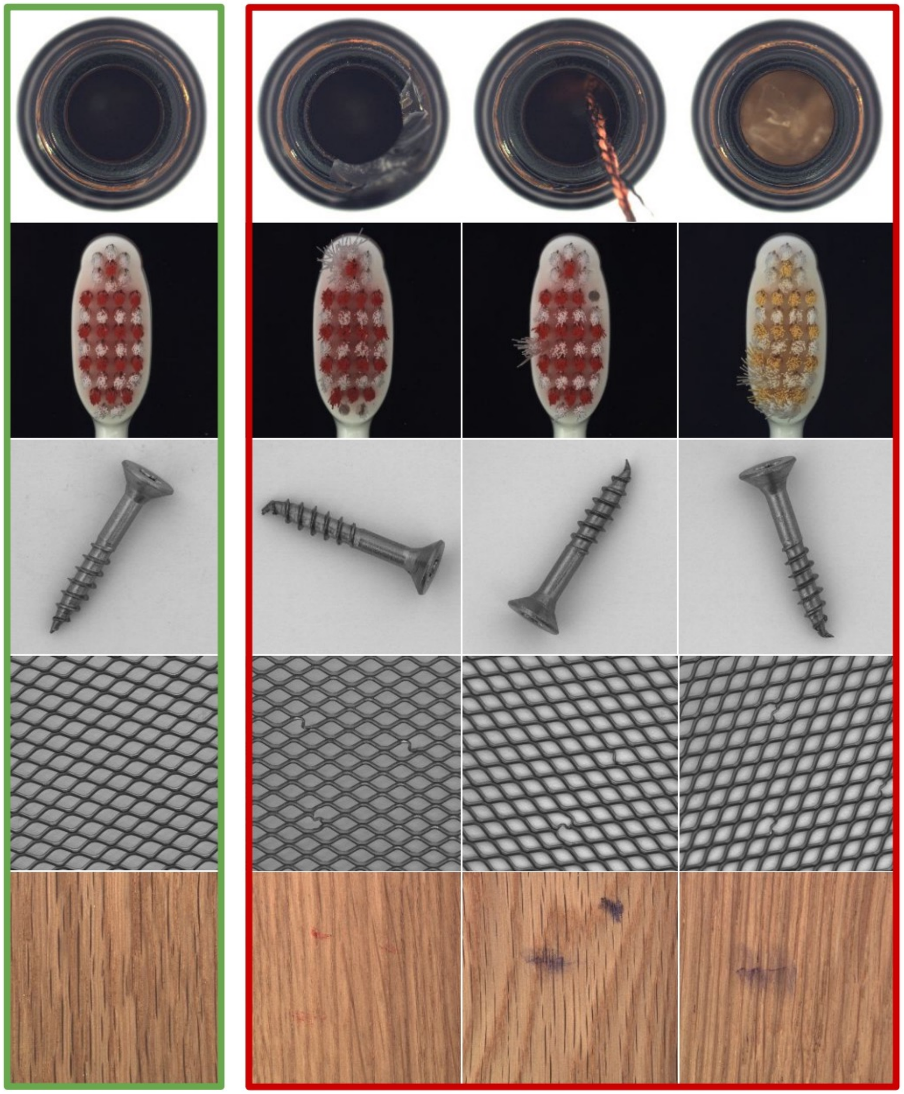
- Detection of criminal activities occurring in commercial organizations.
- Malicious users might be:
  - Employees
  - Actual customers
  - Someone posing as a customer (identity theft)
- Types of fraud
  - Credit card fraud
  - Insurance claim fraud
  - Mobile / cell phone fraud
  - Insider trading
- Challenges
  - Fast and accurate real-time detection
  - Misclassification cost is very high

Healthcare informatics

- Detect anomalous patient records
  - Indicate disease outbreaks, instrumentation errors, etc.
- Key challenges
  - Only normal labels available
  - Misclassification cost is very high
  - Data can be complex: spatio-temporal

Industrial damage detection

- Key challenges
  - Data is extremely large, noisy, and unlabeled
  - Most of applications exhibit temporal behavior
  - Detected anomalous events typically require immediate intervention



(a) Normal (b) Anomaly

Use of data labels in anomaly detection

- Supervised anomaly detection
  - Labels available for both normal data and anomalies
  - Similar to classification with high class imbalance
- Semi-supervised anomaly detection
  - Labels available only for normal data
  - Labels available only for anomalies
- Unsupervised anomaly detection
  - No labels assumed
  - Based on the assumption that anomalies are very rare compared to normal data

Output of anomaly detection

- Label
  - Each test instance is given a *normal* or *anomaly* label
  - Typical output of classification-based approaches

- Score
  - Each test instance is assigned an anomaly score
    - allows outputs to be ranked
    - requires an additional threshold parameter

Anomaly detection problem definition

3.1.1 Problem Statement. Given a training dataset $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N, \mathbf{x}_{N+1}, \cdots, \mathbf{x}_{N+K}\}$, with $\mathbf{x}_i \in \mathbb{R}^D$, where $\mathcal{U} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ is a large unlabeled dataset and $\mathcal{A} = \{\mathbf{x}_{N+1}, \mathbf{x}_{N+2}, \cdots, \mathbf{x}_{N+K}\}$ ($K \ll N$) is a small set of labeled anomaly examples that often do not illustrate every possible class of anomaly, our goal is to learn a scoring function $\phi : \mathcal{X} \to \mathbb{R}$ that assigns anomaly scores to data instances in a way that we have $\phi(\mathbf{x}_i) > \phi(\mathbf{x}_j)$ if $\mathbf{x}_i$ is an anomaly (despite it is a seen or unseen anomaly) and $\mathbf{x}_j$ is a normal instance.

Anomaly detection: Supervised

- Supervised methods → Classification of a class attribute with very rare class values

- Key issue: Unbalanced datasets
  - Suppose a intrusion detection problem.
  - Two classes: *normal* (99.9%) and *intrusion* (0.1%)
  - The default classifier, always labeling each new entry as *normal,* would have 99.9% accuracy!

Anomaly detection : Supervised

- Managing the problem of Classification with rare classes:
  - We need other evaluation measures as alternatives to accuracy (Recall, Precision, F-measure, ROC-curves)
  - Some methods manipulate the data input, oversampling those tuples with the outlier label (the rare class value)
  - Cost-sensitive methods (assigning high cost to the rare class value)