Title: Anomaly detection

URL: https://en.wikipedia.org/wiki/Anomaly_detection

PageID: 8190902

Categories: Category:Data mining, Category:Data security, Category:Machine learning, Category:Reliability engineering, Category:Statistical outliers

---

Supervised learning

Unsupervised learning

Semi-supervised learning

Self-supervised learning

Reinforcement learning

Meta-learning

Online learning

Batch learning

Curriculum learning

Rule-based learning

Neuro-symbolic AI

Neuromorphic engineering

Quantum machine learning

Classification

Generative modeling

Regression

Clustering

Dimensionality reduction

Density estimation

Anomaly detection

Data cleaning

AutoML

Association rules

Semantic analysis

Structured prediction

Feature engineering

Feature learning

Learning to rank

Grammar induction

Ontology learning

Multimodal learning

Apprenticeship learning

Decision trees

Ensembles Bagging Boosting Random forest

Bagging

Boosting

Random forest

k -NN

Linear regression

Naive Bayes

Artificial neural networks

Logistic regression

Perceptron

Relevance vector machine (RVM)

Support vector machine (SVM)

BIRCH

CURE

Hierarchical

k -means

Fuzzy

Expectation–maximization (EM)

DBSCAN

OPTICS

Mean shift

Factor analysis

CCA

ICA

LDA

NMF

PCA

PGD

t-SNE

SDL

Graphical models Bayes net Conditional random field Hidden Markov

Bayes net

Conditional random field

Hidden Markov

RANSAC

k -NN

Local outlier factor

Isolation forest

Autoencoder

Deep learning

Feedforward neural network

Recurrent neural network LSTM GRU ESN reservoir computing

LSTM

GRU

ESN

reservoir computing

Boltzmann machine Restricted

Restricted

GAN

Diffusion model

SOM

Convolutional neural network U-Net LeNet AlexNet DeepDream

U-Net

LeNet

AlexNet

DeepDream

Neural field Neural radiance field Physics-informed neural networks

Neural radiance field

Physics-informed neural networks

Transformer Vision

Vision

Mamba

Spiking neural network

Memtransistor

Electrochemical RAM (ECRAM)

Q-learning

Policy gradient

SARSA

Temporal difference (TD)

Multi-agent Self-play

Self-play

Active learning

Crowdsourcing

Human-in-the-loop

Mechanistic interpretability

RLHF

Coefficient of determination

Confusion matrix

Learning curve

ROC curve

Kernel machines

Bias–variance tradeoff

Computational learning theory

Empirical risk minimization

Occam learning

PAC learning

Statistical learning

VC theory

Topological deep learning

AAAI

ECML PKDD

NeurIPS

ICML

ICLR

IJCAI

ML

JMLR

Glossary of artificial intelligence

List of datasets for machine-learning research List of datasets in computer vision and image processing

List of datasets in computer vision and image processing

Outline of machine learning

v

t

e

In data analysis , anomaly detection (also referred to as outlier detection and sometimes as novelty detection) is generally understood to be the identification of rare items, events or observations which deviate significantly from the majority of the data and do not conform to a well defined notion of normal behavior. Such examples may arouse suspicions of being generated by a different mechanism, or appear inconsistent with the remainder of that set of data.

Anomaly detection finds application in many domains including cybersecurity , medicine , machine vision , statistics , neuroscience , law enforcement and financial fraud to name only a few. Anomalies were initially searched for clear rejection or omission from the data to aid statistical analysis, for example to compute the mean or standard deviation. They were also removed to better predictions from models such as linear regression, and more recently their removal aids the

performance of machine learning algorithms. However, in many applications anomalies themselves are of interest and are the observations most desirous in the entire data set, which need to be identified and separated from noise or irrelevant outliers.

Three broad categories of anomaly detection techniques exist. Supervised anomaly detection techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier. However, this approach is rarely used in anomaly detection due to the general unavailability of labelled data and the inherent unbalanced nature of the classes. Semi-supervised anomaly detection techniques assume that some portion of the data is labelled. This may be any combination of the normal or anomalous data, but more often than not, the techniques construct a model representing normal behavior from a given normal training data set, and then test the likelihood of a test instance to be generated by the model. Unsupervised anomaly detection techniques assume the data is unlabelled and are by far the most commonly used due to their wider and relevant application.

Definition

Many attempts have been made in the statistical and computer science communities to define an anomaly. The most prevalent ones include the following, and can be categorised into three groups: those that are ambiguous, those that are specific to a method with pre-defined thresholds usually chosen empirically, and those that are formally defined:

Ill defined

An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.

Anomalies are instances or collections of data that occur very rarely in the data set and whose features differ significantly from most of the data.

An outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.

An anomaly is a point or collection of points that is relatively distant from other points in multi-dimensional space of features.

Anomalies are patterns in data that do not conform to a well-defined notion of normal behaviour.

Specific

Let T be observations from a univariate Gaussian distribution and O a point from T. Then the z-score for O is greater than a pre-selected threshold if and only if O is an outlier.

History

Intrusion detection

The concept of intrusion detection, a critical component of anomaly detection, has evolved significantly over time. Initially, it was a manual process where system administrators would monitor for unusual activities, such as a vacationing user's account being accessed or unexpected printer activity. This approach was not scalable and was soon superseded by the analysis of audit logs and system logs for signs of malicious behavior.

By the late 1970s and early 1980s, the analysis of these logs was primarily used retrospectively to investigate incidents, as the volume of data made it impractical for real-time monitoring. The affordability of digital storage eventually led to audit logs being analyzed online, with specialized programs being developed to sift through the data. These programs, however, were typically run during off-peak hours due to their computational intensity.

The 1990s brought the advent of real-time intrusion detection systems capable of analyzing audit data as it was generated, allowing for immediate detection of and response to attacks. This marked a significant shift towards proactive intrusion detection.

As the field has continued to develop, the focus has shifted to creating solutions that can be efficiently implemented across large and complex network environments, adapting to the

ever-growing variety of security threats and the dynamic nature of modern computing infrastructures.

Applications

Anomaly detection is applicable in a very large number and variety of domains, and is an important subarea of unsupervised machine learning. As such it has applications in cyber-security, intrusion detection , fraud detection , fault detection, system health monitoring, event detection in sensor networks, detecting ecosystem disturbances, defect detection in images using machine vision , medical diagnosis and law enforcement.

Intrusion detection

Anomaly detection was proposed for intrusion detection systems (IDS) by Dorothy Denning in 1986. Anomaly detection for IDS is normally accomplished with thresholds and statistics, but can also be done with soft computing , and inductive learning. Types of features proposed by 1999 included profiles of users, workstations, networks, remote hosts, groups of users, and programs based on frequencies, means, variances, covariances, and standard deviations. The counterpart of anomaly detection in intrusion detection is misuse detection .

Fintech fraud detection

Anomaly detection is vital in fintech for fraud prevention.

Preprocessing

Preprocessing data to remove anomalies can be an important step in data analysis, and is done for a number of reasons. Statistics such as the mean and standard deviation are more accurate after the removal of anomalies, and the visualisation of data can also be improved. In supervised learning , removing the anomalous data from the dataset often results in a statistically significant increase in accuracy.

Video surveillance

Anomaly detection has become increasingly vital in video surveillance to enhance security and safety. With the advent of deep learning technologies, methods using Convolutional Neural Networks (CNNs) and Simple Recurrent Units (SRUs) have shown significant promise in identifying unusual activities or behaviors in video data. These models can process and analyze extensive video feeds in real-time, recognizing patterns that deviate from the norm, which may indicate potential security threats or safety violations. An important aspect for video surveillance is the development of scalable real-time frameworks. Such pipelines are required for processing multiple video streams with low computational resources.

IT infrastructure

In IT infrastructure management, anomaly detection is crucial for ensuring the smooth operation and reliability of services. These are complex systems , composed of many interactive elements and large data quantities, requiring methods to process and reduce this data into a human and machine interpretable format. Techniques like the IT Infrastructure Library (ITIL) and monitoring frameworks are employed to track and manage system performance and user experience. Detected anomalies can help identify and pre-empt potential performance degradations or system failures, thus maintaining productivity and business process effectiveness.

IoT systems

Anomaly detection is critical for the security and efficiency of Internet of Things (IoT) systems. It helps in identifying system failures and security breaches in complex networks of IoT devices. The methods must manage real-time data, diverse device types, and scale effectively. Garg et al. have introduced a multi-stage anomaly detection framework that improves upon traditional methods by incorporating spatial clustering, density-based clustering, and locality-sensitive hashing. This tailored approach is designed to better handle the vast and varied nature of IoT data, thereby enhancing security and operational reliability in smart infrastructure and industrial IoT systems.

Petroleum industry

Anomaly detection is crucial in the petroleum industry for monitoring critical machinery. Martí et al. used a novel segmentation algorithm to analyze sensor data for real-time anomaly detection. This approach helps promptly identify and address any irregularities in sensor readings, ensuring the reliability and safety of petroleum operations.

Oil and gas pipeline monitoring

In the oil and gas sector, anomaly detection is not just crucial for maintenance and safety, but also for environmental protection. Aljameel et al. propose an advanced machine learning-based model for detecting minor leaks in oil and gas pipelines, a task traditional methods may miss.

Methods

Many anomaly detection techniques have been proposed in literature. The performance of methods usually depend on the data sets. For example, some may be suited to detecting local outliers, while others global, and methods have little systematic advantages over another when compared across many data sets. Almost all algorithms also require the setting of non-intuitive parameters critical for performance, and usually unknown before application. Some of the popular techniques are mentioned below and are broken down into categories:

Statistical

Parameter-free

Also referred to as frequency-based or counting-based, the simplest non-parametric anomaly detection method is to build a histogram with the training data or a set of known normal instances, and if a test point does not fall in any of the histogram bins mark it as anomalous, or assign an anomaly score to test data based on the height of the bin it falls in. The size of bins are key to the effectiveness of this technique but must be determined by the implementer.

A more sophisticated technique uses kernel functions to approximate the distribution of the normal data. Instances in low probability areas of the distribution are then considered anomalies.

Parametric-based

Z-score ,

Tukey's range test

Grubbs's test

Density

Density-based techniques (k-nearest neighbor , local outlier factor , isolation forests , and many more variations of this concept)

Subspace-base (SOD), correlation-based (COP) and tensor-based outlier detection for high-dimensional data

One-class support vector machines (OCSVM, SVDD)

Neural networks

Replicator neural networks , autoencoders , variational autoencoders, long short-term memory neural networks

Bayesian networks

Hidden Markov models (HMMs)

Minimum Covariance Determinant

Deep Learning Convolutional Neural Networks (CNNs): CNNs have shown exceptional performance in the unsupervised learning domain for anomaly detection, especially in image and video data analysis. Their ability to automatically and hierarchically learn spatial hierarchies of features from low to high-level patterns makes them particularly suited for detecting visual anomalies. For instance, CNNs can be trained on image datasets to identify atypical patterns indicative of defects or out-of-norm conditions in industrial quality control scenarios. Simple

Recurrent Units (SRUs): In time-series data, SRUs, a type of recurrent neural network, have been effectively used for anomaly detection by capturing temporal dependencies and sequence anomalies. Unlike traditional RNNs, SRUs are designed to be faster and more parallelizable, offering a better fit for real-time anomaly detection in complex systems such as dynamic financial markets or predictive maintenance in machinery, where identifying temporal irregularities promptly is crucial. Foundation models : Since the advent of large-scale foundation models that have been used successfully on most downstream tasks, they have also been adapted for use in anomaly detection and segmentation. Methods utilizing pretrained foundation models include using the alignment of image and text embeddings (CLIP, etc.) for anomaly localization, while others may use the inpainting ability of generative image models for reconstruction-error based anomaly detection.

Convolutional Neural Networks (CNNs): CNNs have shown exceptional performance in the unsupervised learning domain for anomaly detection, especially in image and video data analysis. Their ability to automatically and hierarchically learn spatial hierarchies of features from low to high-level patterns makes them particularly suited for detecting visual anomalies. For instance, CNNs can be trained on image datasets to identify atypical patterns indicative of defects or out-of-norm conditions in industrial quality control scenarios.

Simple Recurrent Units (SRUs): In time-series data, SRUs, a type of recurrent neural network, have been effectively used for anomaly detection by capturing temporal dependencies and sequence anomalies. Unlike traditional RNNs, SRUs are designed to be faster and more parallelizable, offering a better fit for real-time anomaly detection in complex systems such as dynamic financial markets or predictive maintenance in machinery, where identifying temporal irregularities promptly is crucial.

Foundation models : Since the advent of large-scale foundation models that have been used successfully on most downstream tasks, they have also been adapted for use in anomaly detection and segmentation. Methods utilizing pretrained foundation models include using the alignment of image and text embeddings (CLIP, etc.) for anomaly localization, while others may use the inpainting ability of generative image models for reconstruction-error based anomaly detection.

Cluster-based

Clustering: Cluster analysis -based outlier detection

Deviations from association rules and frequent itemsets

Fuzzy logic-based outlier detection

Ensembles

Ensemble techniques , using feature bagging , score normalization and different sources of diversity

Others

Histogram-based Outlier Score (HBOS) uses value histograms and assumes feature independence for fast predictions.

Anomaly detection in dynamic networks

Dynamic networks, such as those representing financial systems, social media interactions, and transportation infrastructure, are subject to constant change, making anomaly detection within them a complex task. Unlike static graphs, dynamic networks reflect evolving relationships and states, requiring adaptive techniques for anomaly detection.

Types of anomalies in dynamic networks

Community anomalies

Compression anomalies

Decomposition anomalies

Distance anomalies

Probabilistic model anomalies

Explainable anomaly detection

Many of the methods discussed above only yield an anomaly score prediction, which often can be explained to users as the point being in a region of low data density (or relatively low density compared to the neighbor's densities). In explainable artificial intelligence , the users demand methods with higher explainability. Some methods allow for more detailed explanations:

The Subspace Outlier Degree (SOD) identifies attributes where a sample is normal, and attributes in which the sample deviates from the expected.

Correlation Outlier Probabilities (COP) compute an error vector of how a sample point deviates from an expected location, which can be interpreted as a counterfactual explanation: the sample would be normal if it were moved to that location.

Software

ELKI is an open-source Java data mining toolkit that contains several anomaly detection algorithms, as well as index acceleration for them.

PyOD is an open-source Python library developed specifically for anomaly detection.

scikit-learn is an open-source Python library that contains some algorithms for unsupervised anomaly detection.

Wolfram Mathematica provides functionality for unsupervised anomaly detection across multiple data types

Datasets

Anomaly detection benchmark data repository with carefully chosen data sets of the Ludwig-Maximilians-Universität München ; Mirror Archived 2022-03-31 at the Wayback Machine at University of São Paulo .

ODDS – ODDS: A large collection of publicly available outlier detection datasets with ground truth in different domains.

Unsupervised Anomaly Detection Benchmark at Harvard Dataverse: Datasets for Unsupervised Anomaly Detection with ground truth.

KMASH Data Repository at Research Data Australia having more than 12,000 anomaly detection datasets with ground truth.

See also

Change detection

Statistical process control

Novelty detection

Hierarchical temporal memory

References

v

t

e

Adware

Advanced persistent threat

Arbitrary code execution

Backdoors

Bombs Fork Logic Time Zip

Fork
Logic
Time
Zip
Hardware backdoors
Code injection
Crimeware
Cross-site scripting
Cross-site leaks
DOM clobbering
History sniffing
Cryptojacking
Botnets
Data breach
Drive-by download
Browser Helper Objects
Viruses
Data scraping
Denial-of-service attack
Eavesdropping
Email fraud
Email spoofing
Exploits
Fraudulent dialers
Hacktivism
Infostealer
Insecure direct object reference
Keystroke loggers
Malware
Payload
Phishing Voice
Voice
Polymorphic engine
Privilege escalation
Ransomware
Rootkits
Scareware
Shellcode

Spamming

Social engineering

Spyware

Software bugs

Trojan horses

Hardware Trojans

Remote access trojans

Vulnerability

Web shells

Wiper

Worms

SQL injection

Rogue security software

Zombie

Application security Secure coding Secure by default Secure by design Misuse case

Secure coding

Secure by default

Secure by design Misuse case

Misuse case

Computer access control Authentication Multi-factor authentication Authorization

Authentication Multi-factor authentication

Multi-factor authentication

Authorization

Computer security software Antivirus software Security-focused operating system

Antivirus software

Security-focused operating system

Data-centric security

Software obfuscation

Data masking

Encryption

Firewall

Intrusion detection system Host-based intrusion detection system (HIDS) Anomaly detection

Host-based intrusion detection system (HIDS)

Anomaly detection

Information security management Information risk management Security information and event management (SIEM)

Information risk management

Security information and event management (SIEM)

Runtime application self-protection

Site isolation

Computer security

Automotive security

Cybercrime Cybersex trafficking Computer fraud

Cybersex trafficking

Computer fraud

Cybergeddon

Cyberterrorism

Cyberwarfare

Electronic warfare

Information warfare

Internet security

Mobile security

Network security

Copy protection

Digital rights management

United States

Israel

Yale LUX