ANOMALY BASED INTRUSION AND DETECTION WITH MACHINE LEARNING

Vani Krishnaswamy
Assistant Professor,
Department of Computing-Decision and Computing Sciences
Coimbatore Institute of Technology, Coimbatore, India
vani.k@cit.edu.in

Dr.S.P Swornambiga
Associate Professor, Dept of Computer Applications,
CMS College of Science & Commerce,
Coimbatore, India
swornagoms@gmail.com

Abstract: Anomaly detection is the identification of rare events, items, or observations which are suspicious because they differ significantly from standard behaviours or patterns. Anomalies in data are also called standard deviations, outliers, noise, novelties, and exceptions. In the network anomaly detection/network intrusion and abuse detection context, interesting events are often not rare. Anomaly detection is any process that finds the outliers of a dataset; those items that don't belong. These anomalies might point to unusual network traffic, uncover a sensor on the fritz, or simply identify data for cleaning, before analysis. In today's world of distributed systems, managing and monitoring the system's performance is a chore—albeit a necessary chore. With hundreds or thousands of items to watch, anomaly detection can help point out where an error is occurring, enhancing root cause analysis and quickly getting tech support on the issue. Anomaly detection helps the monitoring cause of chaos engineering by detecting outliers, and informing the responsible parties to act.

Keywords: Network Intrusion, Detection Context, Anomaly Detection, Dataset, Sensors, Exceptions.

1.INTRODUCTION

It is critical for network admins to be able to identify and react to changing operational conditions. Any nuances in the operational conditions of data centres or cloud applications can signal unacceptable levels of business risk. On the other hand, some divergences may point to positive growth. Therefore, anomaly detection is central to extracting essential business insights and maintaining core operations. Consider these patterns—all of which demand the ability to discern between normal and abnormal behaviour precisely and correctly:

- An online retail business must predict which discounts, events, or new products may trigger boosts in sales which will increase demand on their web servers.

- An IT security team must prevent hacking and needs to detect abnormal login patterns and user behaviours.
- A cloud provider has to allot traffic and services and has to assess changes to infrastructure in light of existing patterns in traffic and past resource failures.

A evidence-based, well-constructed behavioural model can not only represent data behaviour, but also help users identify outliers and engage in meaningful predictive analysis. Static alerts and thresholds are not enough, because of the overwhelming scale of the operational parameters, and because it's too easy to miss anomalies in false positives or negatives.
To address these kinds of operational constraints, newer systems use smart algorithms for identifying outliers in seasonal time series data and accurately forecasting periodic data patterns.

Anomaly Classification

Network anomalies: Anomalies in network behaviour deviate from what is normal, standard, or expected. To detect network anomalies, network owners must have a concept of expected or normal behaviour. Detection of anomalies in network behaviour demands the continuous monitoring of a network for unexpected trends or events.

Application performance anomalies: These are simply anomalies detected by end-to-end application performance monitoring. These systems observe application function, collecting data on all problems, including supporting infrastructure and app dependencies. When anomalies are detected, rate limiting is triggered and admins are notified about the source of the issue with the problematic data.

Web application security anomalies: These include any other anomalous or suspicious web application behaviour that might impact security such as CSS attacks or DDOS attacks.

Detection of each type of anomaly relies on ongoing, automated monitoring to create a picture of normal network or application behaviour. This type of monitoring might focus on point anomalies/global outliers, contextual anomalies, and/or collective anomalies; the context of the network, the performance of the application, or the web application security is more important to the goal of the anomaly detection system.

Anomaly detection and novelty detection or noise removal are similar, but distinct. Novelty detection identifies patterns in data that were previously unobserved so users can determine whether they are anomalous. Noise removal is the process of removing noise or unneeded observations from a signal that is otherwise meaningful.

To track monitoring KPIs such as bounce rate and churn rate, time series data anomaly detection systems must first develop a baseline for normal behaviour. This enables the system to track seasonality and cyclical behaviour patterns within key datasets.

II IMPORTANCE OF ANAMOLY DETECTION

It is critical for network admins to be able to identify and react to changing operational conditions. Any nuances in the operational conditions of data centres or cloud applications can signal unacceptable levels of business risk. On the other hand, some divergences may point to positive growth.

Therefore, anomaly detection is central to extracting essential business insights and maintaining core operations. Consider these patterns—all of which demand the ability to discern between normal and abnormal behaviour precisely and correctly:

- An online retail business must predict which discounts, events, or new products may trigger boosts in sales which will increase demand on their web servers.
- An IT security team must prevent hacking and needs to detect abnormal login patterns and user behaviours.
- A cloud provider has to allot traffic and services and has to assess changes to infrastructure in light of existing patterns in traffic and past resource failures.

An evidence-based, well-constructed behavioural model can not only represent data behaviour, but also help users identify outliers and engage in meaningful predictive analysis. Static alerts and thresholds are not enough, because of the overwhelming scale of the operational parameters, and because it's too easy to miss anomalies in false positives or negatives.
To address these kinds of operational constraints, newer systems use smart algorithms for identifying outliers in seasonal time series data and accurately forecasting periodic data patterns.

There are three main classes of anomaly detection techniques: unsupervised, semi-supervised, and supervised. Essentially, the correct anomaly detection method depends on the available labels in the dataset.

Supervised anomaly detection techniques demand a data set with a complete set of "normal" and "abnormal" labels for a classification algorithm to work with. This kind of technique also involves training the classifier. This is similar to traditional pattern recognition, except that with outlier detection there is a naturally strong imbalance between the classes. Not all statistical classification algorithms are well-suited for the inherently unbalanced nature of anomaly detection.

Semi-supervised anomaly detection techniques use a normal, labelled training data set to construct a model representing normal behaviour. They then use that model to detect anomalies by testing how likely the model is to generate any one instance encountered.

Unsupervised methods of anomaly detection detect anomalies in an unlabelled test set of data based solely on the intrinsic properties of that data. The working assumption is that, as in most cases, the large majority of the instances in the data set will be normal. The anomaly detection algorithm will then detect instances that appear to fit with the rest of the data set least congruently.

III ANOMALY DETECTION TECHNIQUES

There are various anomaly detection techniques. Depending on the circumstances, one might be better than others for a particular user or data set. A generative approach creates a model based solely on examples of normal data from training and then evaluates each test case to see how well it fits the model. In contrast, a discriminative approach attempts to distinguish between normal and abnormal data classes. Both kinds of data are used to train systems in discriminative approaches.

Clustering-Based Anomaly Detection

Clustering-based anomaly detection remains popular in unsupervised learning. It rests upon the assumption that similar data points tend to cluster together in groups, as determined by their proximity to local centroids.

K-means, a commonly-used clustering algorithm, creates ‘k’ similar clusters of data points. Users can then set systems to mark data instances that fall outside of these groups as data anomalies. As an unsupervised technique, clustering does not require any data labelling.

Clustering algorithms might be deployed to capture an anomalous class of data. The algorithm has already created many data clusters on the training set in order to calculate the threshold for an anomalous event. It can then use this rule to create new clusters, presumably capturing new anomalous data.

However, clustering does not always work for time series data. This is because the data depicts evolution over time, yet the technique produces a fixed set of clusters.

Density-Based Anomaly Detection

Density-based anomaly detection techniques demand labelled data. These anomaly detection methods rest upon the assumption that normal data points tend to occur in a dense neighbourhood, while anomalies pop up far away and sparsely.

There are two types of algorithms for this type of data anomaly evaluation:

K-nearest neighbour (k-NN) is a basic, non-parametric, supervised machine learning technique that can be used to either regress or classify data based on distance metrics such as Euclidean, Hamming, Manhattan, or Minkowski distance.

Local outlier factor (LOF), also called the relative density of data, is based on reachability distance.

Support Vector Machine-Based Anomaly Detection

A support vector machine (SVM) is typically used in supervised settings, but SVM extensions can also be used to identify anomalies for some unlabelled data. A SVM is a neural network that is well-suited for classifying linearly separable binary patterns—obviously the better the separation is, the clearer the results.

Such anomaly detection algorithms may learn a softer boundary depending on the goals to cluster the data instances and identify the abnormalities properly. Depending on the situation, an anomaly detector like this might output numeric scalar values for various uses.

IV ANOMALY DETECTION WITH MACHINE LEARNING

Machine learning, then, suits the engineer's purpose to create an AD system that:

- Works better
- Is adaptive and on time
- Handles large datasets

Despite these benefits, anomaly detection with machine learning can only work under certain conditions.

Unstructured data

Applying machine learning to anomaly detection requires a good understanding of the problem, especially in situations with unstructured data.

Structured data already implies an understanding of the problem space. Anomalous data may be easy to identify because it breaks certain rules. If a sensor should never read 300 degrees Fahrenheit and the data shows the sensor reading 300 degrees Fahrenheit—there's your anomaly. There is a clear threshold that has been broken.

Fraud detection in the early anomaly algorithms could work because the data carried with it meaning. The data came structured, meaning people had already created an interpretable setting for collecting data. Their data carried significance, so it was possible to create random trees and look for fraud.

However, dark data and unstructured data, such as images encoded as a sequence of pixels or language encoded as a sequence of characters, carry with it little interpretation and render the old algorithms useless…until the data becomes structured. Structure can be found in the last layers of a convolutional neural network (CNN) or in any number of sorting algorithms.

Large datasets needed

Second, a large data set is necessary. A founding principle of any good machine learning model is that it requires datasets. Like law, if there is no data to support the claim, then the claim cannot hold in court.

Machine learning requires datasets; inferences can be made only when predictions can be validated. Anomaly detection benefits from even larger amounts of data because the assumption is that anomalies are rare.

Scarcity can only occur in the presence of abundance.

Talent required

Third, machine learning engineers are necessary. Obvious, but sometimes overlooked. Machine learning talent is not a commodity, and like car repair shops, not all engineers are equal.

It requires skill and craft to build a good Machine Learning model. The cost to get an anomaly detector from 95% detection to 98% detection could be a few years and a few ML hires.

Anomaly detection in three settings

In a 2018 lecture, Dr. Thomas Dietterich and his team at Oregon State University explain how anomaly detection will occur under three different settings. They all depend on the condition of the data. The three settings are:

1. Supervised
2. Clean
3. Unsupervised

Popular ML Algorithms for unstructured data are:

- Self-organizing maps (SOM)
- K-means
- C-means
- Expectation-maximization meta-algorithm (EM)
- Adaptive resonance theory (ART)
- One-class support vector machine

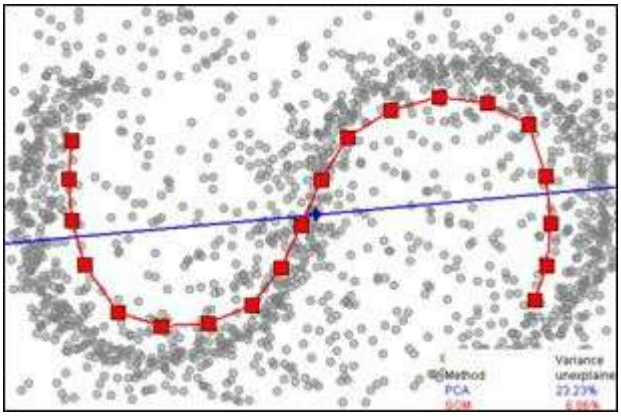

Fig 1.1 SOM Detection (Source)

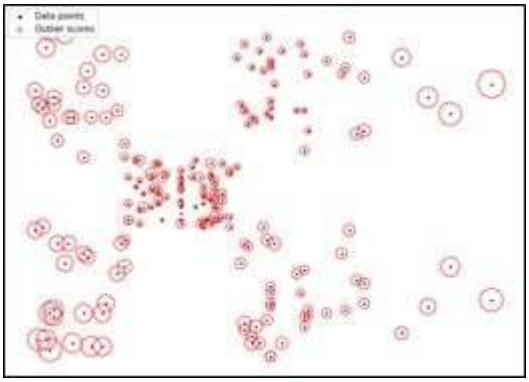

Fig 1.2 K-Means Clustering (Source)

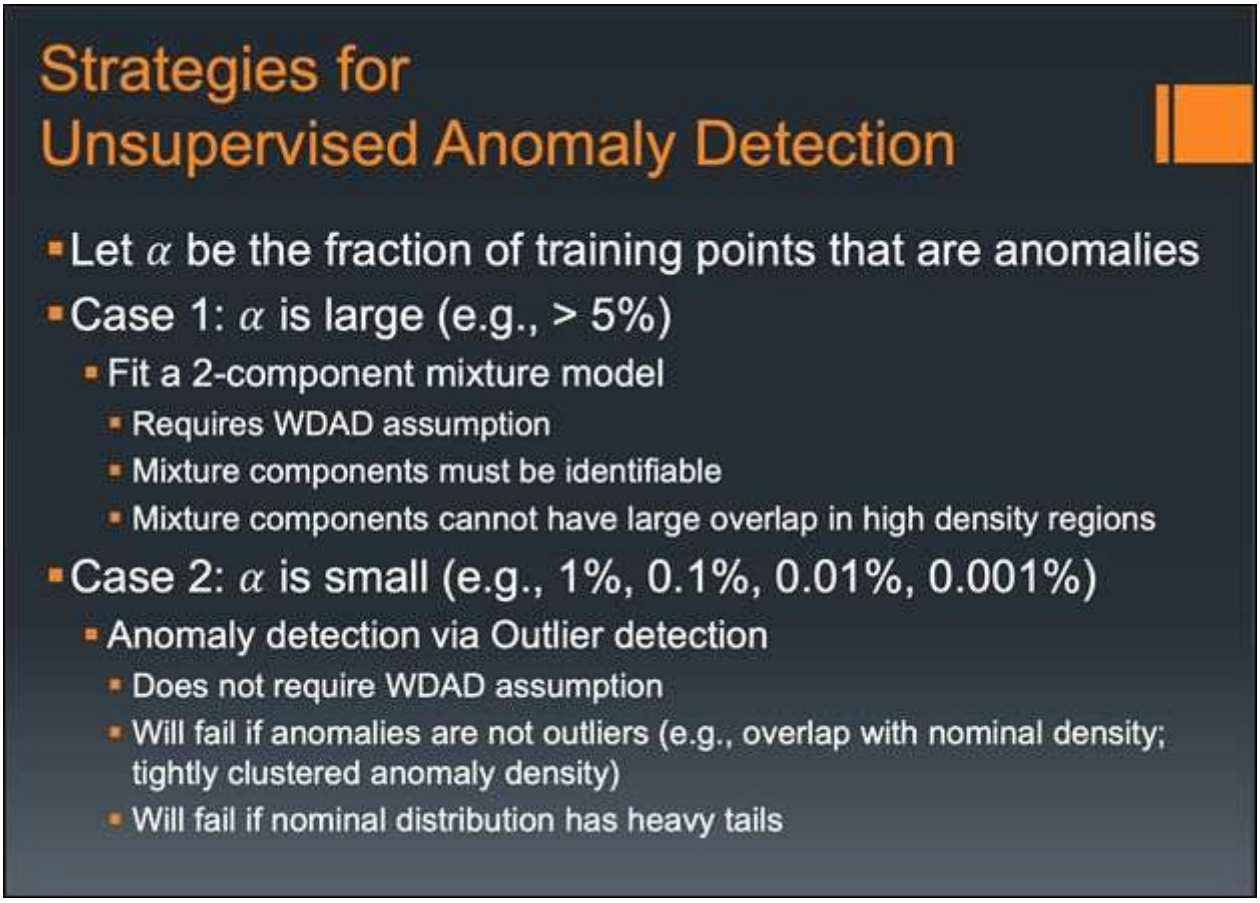

Fig 1.3 Strategies

Where machine learning isn't appropriate, top non-ML detection algorithms include:

- IFOR: Isolation Forest (Liu, et al., 2008)
- LODA: Lightweight Online Detector of Anomalies (Pevny, 2016)

Benchmarking anomaly detection

Engineers use benchmarks to be able to compare the performance of one algorithm to another's. Different kinds of models use different benchmarking datasets:

- Image classification has MNIST and IMAGENET.
- Language modelling has Penn TreeBank and Wiki Text-2.

In anomaly detection, no one dataset has yet become a standard. This has to do, in part, with how varied the applications can be. However, one body of work is emerging as a continuous presence—the Numenta Anomaly Benchmark. From the GitHub Repo:

"NAB is a novel benchmark for evaluating algorithms for anomaly detection in streaming, real-time applications. It is composed of over 50 labeled real-world and artificial time series data files plus a novel scoring mechanism designed for real-time applications."

Thus far, on the NAB benchmarks, the best performing anomaly detector algorithm catches 70% of anomalies from a real-time dataset.

Table 1.1 Image classification

Detector	Standard Profile	Reward Low FP	Reward Low FN
Perfect	100.0	100.0	100.0
Numenta HTM*	70.5-69.7	62.6-61.7	75.2-74.2
CAD OSE†	69.9	67.0	73.2
earthgecko Skyline	58.2	46.2	63.9

V ANOMALY DETECTION USE CASES

Some of the primary anomaly detection use cases include anomaly based intrusion detection, fraud detection, data loss prevention (DLP), anomaly based malware detection, medical anomaly detection, anomaly detection on social platforms, log anomaly detection, internet of things (IoT) big data system anomaly detection, industrial/monitoring anomalies, and anomalies in video surveillance.

An anomaly based intrusion detection system (IDS) is any system designed to identify and prevent malicious activity in a computer network. A single computer may have its own IDS, called a Host Intrusion Detection System (HIDS), and such a system can also be scaled up to cover large networks. At that scale it is called Network Intrusion Detection (NIDS).

This is also sometimes called network behaviour anomaly detection, and this is the kind of ongoing monitoring network behaviour anomaly detection tools are designed to provide. Most IDS depend on signature-based or

anomaly-based detection methods, but since signature-based IDS are ill-equipped to detect unique attacks, anomaly-based detection techniques remain more popular.

Fraud in banking (credit card transactions, tax return claims, etc.), insurance claims (automobile, health, etc.), telecommunications, and other areas is a significant issue for both private business and governments. Fraud detection demands adaptation, detection, and prevention, all with data in real-time.

Data loss prevention (DLP) is similar to prevention of fraud, but focuses exclusively on loss of sensitive information at an early stage. In practice, this means logging and analysing accesses to file servers, databases, and other sources of information in near-real-time to detect uncommon access patterns.

Malware detection is another important area, typically divided into feature extraction and clustering/classification stages. Sheer scale of data is a tremendous challenge here, along with the adaptive nature of the malicious behaviour.

Detecting anomalies in medical images and records enables experts to diagnose and treat patients more effectively. Massive amounts of imbalanced data means reduced ability to detect and interpret patterns without these techniques. This is an area that is ideal for artificial intelligence given the tremendous amount of data processing involved.

Detecting anomalies in a social network enables administrators to identify fake users, online fraudsters, predators, rumour-mongers, and spammers that can have serious business and social impact.

Log anomaly detection enables businesses to determine why systems fail by reconstructing faults from patterns and past experiences.

Monitoring data generated in the field of the Internet of things (IoT) ensures that data generated by IT infrastructure components, radio-frequency identification (RFID) tags, weather stations, and other sensors are accurate and identifies faulty and fraudulent behaviour before disaster strikes. The same is true of monitoring industrial systems such as high-temperature energy systems, power plants, wind turbines, and storage devices that are exposed to massive daily stress.

VI CONCLUSION

The unsupervised anomaly detection methods work best when you're not aware of the type of anomalies that may occur, especially with unstructured data. Supervised is best when sufficient data is available, and the nature of anomalies is consistent with the real world. According to the accuracy score Logistic regression works pretty well because predicting fraud transactions is a classification problem. So, this is one method to predict the fraud

transaction but also there are many methods and algorithms are there to solve this problem. In this article, we covered what is anomaly detection and how can we use machine learning to perform such tasks

REFERENCES

[1]. https://www.enjoyalgorithms.com/blog/introduction-to-anomaly-detection/

[2]. https://www.bmc.com/blogs/machine-learning-anomaly-detection/

[3]. https://www.projectpro.io/article/anomaly-detection-using-machine-learning-in-python-with-example/555

[4]. https://towardsdatascience.com/how-to-use-machine-learning-for-anomaly-detection-and-condition-monitoring-6742f82900d7

[5]. https://www.javatpoint.com/machine-learning-with-anomaly-detection