Anomaly Detection

2018 PHM Society Annual Conference Tutorial

Neil Eklund, Ph.D.
History of Anomaly Detection
Origins of the Notion of “Outlier”

In Europe, the concept of outlier arose from the analysis of geodetic (land survey) and astronomical data. Presumably also:

- Mesopotamia
- India
- Greece
- Egypt
- China
- Mesoamerica

Maire & Boscovich (1775) described outlier rejection by removing two observations that were “too much deviated” from the remaining data.
Bernoulli, 1777 – Uncertainty

“I see no way of drawing a dividing line between those that are utterly to be rejected and those that are to be wholly retained; it may even happen that the rejected observation is the one that would have supplied the best correction to the others.”
Peirce, 1852 – 1st Statistical Criterion

THE ASTRONOMICAL JOURNAL.

No. 45.

VOL. II. CAMBRIDGE, 1852, JULY 24. NO. 21.

CRITERION FOR THE REJECTION OF DOUBTFUL OBSERVATIONS.

BY BENJAMIN PEIRCE, LL. D.,
PERKINS PROFESSOR OF ASTRONOMY AND MATHEMATICS IN HARVARD UNIVERSITY.

1. In almost every true series of observations, some are found, which differ so much from the others as to indicate some abnormal source of error not contemplated in the theoretical

Let \( N \) be the whole number of observations.
Let \( n \) be the number of observations proposed to be rejected.
Let \( n' = N - n \) be the number of observations to be retained.

“…the proposed observations should be rejected when the probability of the system of errors obtained by retaining them is less than that of the system of errors obtained by their rejection multiplied by the probability of making so many, and no more, abnormal observations”

Although Innovation Frequently Begets Criticism…

Airy, 1856:

“And I have, not without surprize to myself, been led to think that the whole theory is defective in its foundation, and illusory in its results; that no rule for the exclusion of observations can be obtained by any process founded purely upon a consideration of the discordance of these observations.”

Glaisher, 1872:

“Professor Pierce's [sic] criterion for the rejection of doubtful observations seems to me to be destitute of scientific precision.”

1882 – Newcomb’s Revolutionary Ideas

Newcomb was the first to go beyond simple rejection of outliers:

• rejected outliers when necessary, but usually only based on external evidence or really huge deviations

• critical of outlier rejection criteria, he developed a new estimation procedure which weighted “more discordant” observations less heavily

• Introduced a mixture of normal densities as a model for a heavy-tailed distribution

“It is evident that if we have a collection of observations of different degrees of probable error, in which, however, there is no way of distinguishing those of great probable error from those of small probable error, the law of the errors will not be that usually adopted, but there will be a comparative excess of large residuals.”

Edgeworth (1887) – Least Absolute Values Regression

Early robust regression technique: $\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} |r_i|$
Rejection of multiple outliers in a single step (i.e., non-iterative) using a statistical criterion.
Shewhart (1926)

• Bell Labs – buried telephone equipment reliability issues

• Indicate process “in control”, i.e., is stable, with variation only coming from sources common to the process

• Still in use today, one of “Seven Basic Tools of Quality”

Irrespective of the care taken in defining the production procedure, the manufacturer realizes that he cannot make all units of a given kind of product identical. This is equivalent to assuming the existence of non-assignable causes of variation in quality of product. Of course, random fluctuations in such factors as humidity, temperature, wear and tear of machinery and the psychological and physiolog-
Grubbs (1969)

**SAMPLE CRITERIA FOR TESTING OUTLYING OBSERVATIONS**

By Frank E. Grubbs

University of Michigan and Ballistic Research Laboratories

1. **Summary.** The problem of testing outlying observations, although an old one, is of considerable importance in applied statistics. Many and various types

First still commonly used univariate outlier detection approach

• Assumes a Gaussian distribution for the training data & works only with univariate continuous data

• Computes the distance of the test data points from the estimated sample mean and declares any point with a distance above a certain threshold to be an outlier

• Requires a threshold parameter to determine the length of the tail that includes the outliers, typically three standard deviations from the mean

Fox (1972)

- First to consider outliers within time series
- Two broad categories of outlier are defined:
  - “Additive Outliers” where a single point is affected
  - “Innovative Outliers” where an innovation to the process affects both an observation and the subsequent series.
- Basically a likelihood ratio criteria, comparing the estimated error for an observation with the estimated standard error of that discrepancy

What is an Anomaly?
In engineering, terms like outlier, spurious observation, contaminant, gross error, and others are used with different and overlapping meanings.

Techniques for anomaly detection (AD) are also described in the literature as:

- Outlier Detection
- Novelty Detection
- Rare Class Mining
- Chance Discovery
- Exception Mining
- Noise Removal
Three Common Types of Anomaly

1. **Point anomalies**: A single data sample is defined a point anomaly if it is considered anomalous relative to the rest of the data.

2. **Contextual anomalies**: If data samples are considered anomalous only in a specific context they are called contextual anomalies.

3. **Collective anomalies**: Data samples are considered anomalous only when a group of correlated samples are anomalies relative to the entire dataset, even when the samples within the collective would not be considered an anomaly on its own.

$X_1$ and $X_2$ are **point anomalies**, also frequently called **global anomalies**.

---

$X_3$ can be seen as a normal record since it is not too far away from the cluster $C_2$.

However, if we focus only on the cluster $C_2$ and compare it with $X_3$ while neglecting all the other instances, it can be seen as an anomaly!

$X_3$ is therefore a *contextual anomaly*, or *local anomaly*, since it is only anomalous when compared with its close-by neighborhood.

It depends on the application, whether local anomalies are of interest or not.

---

Is \( C_3 \) three anomalies or a (small) regular cluster?

This is sometimes referred to as a *micro cluster* and anomaly detection algorithms should rate them more anomalous than the normal instances, but less anomalous than the obvious anomalies.

This simple example already illustrates that anomalies are not always obvious and a score is much more useful than a binary label assignment.

Point $t_2$ is a *contextual anomaly* in this temperature time series.

The temperature at time $t_1$ is same as that at time $t_2$ but occurs in a different context and hence is not considered anomalous.
If a collection of related data instances is anomalous with respect to the entire data set, it is termed as a *collective anomaly*. The individual data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together as a collection is anomalous.

The red segment is a collective anomaly corresponding to an atrial premature contraction in an human electrocardiogram output.

Note that that values by themselves are not an anomaly.

Point v. Contextual v. Collective

• While point anomalies can occur in any data set, collective anomalies can occur only in data sets in which data instances are related

• Occurrence of contextual anomalies depends on the availability of context attributes in the data

• A point anomaly or a collective anomaly can also be a contextual anomaly if analyzed with respect to a context

Thus a point anomaly detection problem or collective anomaly detection problem can be transformed to a contextual anomaly detection problem by incorporating the context information

Taxonomy of AD Approaches
Three Broad Approaches to Anomaly Detection

1. **Supervised Anomaly Detection**: Similar to any pattern classification approach, an ordinary classifier can be trained in advance, and applied at run-time. There are still properties of the problem, such as highly unbalanced data, which might drive classifier selection, but this is in many ways the easiest (and best) case. That said, it is also the least common and least practical, due to both lack of labeled data, and the likelihood of novel anomalies occurring at run-time.

2. **Semi-supervised Anomaly Detection**: In this much more common scenario, we use known “normal” data without any labeled anomalies (although perhaps actually containing some anomalous data) to train a classifier to learn the normal region. At run-time, anomalies are detected by looking for deviation from the trained model. Some examples are one-class Support Vector Machines (SVM), autoencoding neural networks, and density estimation methods such as Gaussian mixture models and kernel density estimation.

3. **Unsupervised Anomaly Detection**: The most widely used approach, which requires no labels – the algorithms evaluates new cases based on the intrinsic properties of the data. Typically, this is based on distances or densities to determine abnormality.

This tutorial focuses on semi-supervised and unsupervised AD.
Algorithm Output – Label vs Score

Two ways for an algorithm to identify an anomaly:

1. **Label**: nominal tag, “normal” or “anomalous”

   Output of an AD algorithm can be binary, e.g., one/zero – normal/anomalous. This approach is typical of supervised methods, and is often helpful for unsophisticated end-users because it feels like it removes uncertainty. However, it is the least informative approach.

2. **Score**: integer rank $[0 \cdots n]$ or raw real-valued $[0.0 \cdots 1.0]$ classifier output

   More informative as it indicates degree of atypicality – e.g., recall the micro-cluster example above. More common in semi and un-supervised algorithms, and most informative approach – much more useful to other downstream algorithms. A score can always be converted to a label via a threshold.

Supervised Anomaly Detection

Semi-supervised Anomaly Detection

Unsupervised Anomaly Detection
Statistical Approaches
Statistical Tests

Concept:

• Assume a particular kind of statistical distribution, e.g., Gaussian
  • Normal data occurs in the high probability region of the distribution
  • Outliers deviate strongly from this distribution

• Compute the parameters assuming all data points have been generated by such a statistical distribution (e.g., mean and standard deviation)

• Outliers are points that have a low probability to be generated by the overall distribution

• Many different tests, differing principally in:
  • data distribution (e.g. log-normal)
    • Including mixed distributions
  • dimensionality
  • parametric and non-parametric
Mahalanobis Distance

- Mahalanobis distance is a measure of the distance between a point and the center of a distribution
- Multi-dimensional generalization of the idea of measuring how many standard deviations away the point is from the mean

\[
MDist(p, q) = (p - q) \Sigma^{-1} (p - q)^T
\]

- Where \( \Sigma \) is the covariance matrix:

\[
\Sigma_{j,k} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)
\]
Mahalanobis Distance

Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)
B: (0, 1)
C: (1.5, 1.5)

$$\text{MDist}(A,B) = 5$$
$$\text{MDist}(A,C) = 4$$
Statistical Tests – Issues

Robustness

• Mean and standard deviation are very sensitive to outliers...

• Mean and standard deviation calculated over entire data set, including outliers

• MDist values (and similar distance metrics) are influenced by these outliers
Statistical Tests – Issues

Curse of dimensionality

- Say we want to label values with a less than 1% chance of occurring in any dimension
- Normal data in 1 dimension… \( (1 - .01)^1 = .99 \)
- Normal data in 3 dimensions… \( (1 - .01)^3 = .97 \)
- Normal data in 10 dimensions… \( (1 - .01)^{10} = .90 \)
- Normal data in 20 dimensions… \( (1 - .01)^{20} = .82 \)
- Normal data in 50 dimensions… \( (1 - .01)^{50} = .61 \)
- Normal data in 100 dimensions… \( (1 - .01)^{100} = .37 \)
Curse of Dimensionality – Pairwise Distances

Sample of $10^5$ instances drawn from a uniform $(0, 1)$ distribution, normalized $(1/\sqrt{d})$
Nearest Neighbor Approaches
Nearest Neighbor Approaches

Concept:

• “Outliers are objects far away from other objects”

• Distance based methods
  • Anomalies are data points most distant from other points

• Density based methods
  • Anomalies are data points in low density regions
**k-Nearest Neighbor**

- For each data point \( d \) compute the distance to the \( k^{th} \) nearest neighbor \( d_k \)
- Sort all data points according to the distance \( d_k \)
- Outliers are points that have the largest distance \( d_k \) and therefore are located in the more sparse neighborhoods
- Label top \( n \)% distance \( d_k \) are identified as outliers, where \( n \) is user set parameter
- Poorly suited to datasets that have modes with varying density
- As a rule of thumb, \( k \) should be in the range \( 10 < k < 50 \)
Visualization of the results of the k-NN anomaly detection algorithm for $k=10$. Anomaly score is represented by the bubble size whereas the color shows the labels of an artificially generated dataset.

**k-Nearest Neighbor**

**Pros:**

- Easier to define a proximity measure for a dataset than determine its statistical distribution
- Quantitative measure of degree to which object is an outlier
- Deals naturally with multiple modes, *assuming the same density*

**Cons:**

- Computational complexity – $O(n^2)$
- Score sensitive to choice of $k$
- Breaks down if data has variable density
The Problem of Inhomogeneous Data
Solution: Local Outlier Factor (LOF)
LOF is the most well-known local anomaly detection algorithm, and first to account for contextual anomalies.

1. Find $k$ nearest neighbors

2. Using these $k$-nearest-neighbors, $N_k$, the local density for a record is estimated by computing the local reachability density (LRD)
   
   • Except for some very rare situations in highly dense clusters, this is the Euclidean distance

3. LOF score is computed by comparing the LRD of a record with the LRDs of its $k$ neighbors

\[
LOF(x) = \frac{\sum_{o \in N_k(x)} \frac{LRD_k(o)}{LRD_k(x)}}{|N_k(x)|}
\]

Local Outlier Factor (LOF)

relative density (LOF) outlier scores
Local Outlier Factor (LOF)

Pros:
• Quantitative measure of degree to which object is an outlier
• Can work well even if data has variable density

Cons:
• $O(n^2)$ complexity
• Must choose parameters
  • $k$ for nearest neighbor
  • $d$ for distance threshold
Connectivity-Based Outlier Factor (COF)

• Similar to LOF, but different density estimation.

• In LOF, the k-nearest-neighbors are selected based on the Euclidean distance.
  • Assumes, that the data is distributed in a spherical way around the instance
  • If this assumption is violated, for example if features have a direct linear correlation, the density estimation is incorrect

• COF calculates local density of the neighborhood using an shortest-path approach, called the chaining distance, the minimum of the sum of all distances connecting all k neighbors and the instance

• Where features are obviously correlated, this density estimation approach performs much more accurately
The spherical density estimation of LOF fails to recognize the anomaly, whereas COF detects the non-linear anomaly for $k = 4$
High-dimensional Approaches
High-dimensional Data is Challenging!

Recall, curse of dimensionality:

- Relative contrast between distances decreases with increasing dimensionality
- Data are very sparse, almost all points are outliers
- Notion of neighborhood becomes meaningless

Potential solutions

- Use more robust distance functions and find full-dimensional outliers
- Find outliers in projections (subspaces) of the original feature space
High-dimensional Data is Challenging!

Angle-based Outlier Degree (ABOD)

- Angles are more stable than distances in high dimensional spaces (cf. e.g. the popularity of cosine-based similarity measures for text data)
- Object $p$ is an outlier if most other objects are located in similar directions
- Object $p$ is no outlier if many other objects are located in varying directions
Angle-based Outlier Degree (ABOD)

Basic assumption:

- Outliers are at the border of the data distribution
- Normal points are in the center of the data distribution

Model:

- Consider for a given point \( p \) the angle between \( \vec{px} \) and \( \vec{py} \) for any two \( x,y \) from the database
- Consider the spectrum of all these angles
- The broadness of this spectrum indicates outlierness of a point
- Measure the variance of the angle spectrum
- Weighted by the corresponding distances (for lower dimensional data sets where angles are less reliable)
Angle-based Outlier Degree (ABOD)

\[ ABOD(p) = \text{VAR}_{x,y \in DB} \left( \frac{\langle x_p, y_p \rangle}{\left\| x_p \right\|^2 \cdot \left\| y_p \right\|^2} \right) \]

inner point  
outlier  

Kriegel/Kröger/Zimek: Outlier Detection Techniques (KDD 2010)
Angle-based Outlier Degree (ABOD)

Computationally intensive – naïve algorithm is in $O(n^3)$

However…

- Approximate algorithm based on random sampling for mining top-$n$ outliers
- Compute ABOD based only on samples => lower bound of the real ABOD
- Filter out points that have a high lower bound
- Refine (compute the exact ABOD value) only for a small number of points
Isolation Forest

- Isolation Forest (iTree) “isolates” observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.
  - Isolation means “separating an instance from the rest of the instances.”
- The number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node of a tree structure.
- This path length, averaged over a forest of such random trees, is a measure of normality and our decision function.

Isolation Forest

\[ X_i \]

\[ \text{\textbullet} X_o \]
Isolation Forest - Algorithm
for each point of the data set:

1. Select the point to isolate
2. For each feature, set the range to isolate between the minimum and the maximum.
3. Choose a feature randomly.
4. Pick a value that’s in the range, again randomly:
   • If the chosen value keeps the point above, switch the minimum of the range of the feature to the value
   • If the chosen value keeps the point below, switch the maximum of the range of the feature to the value
5. Repeat steps 3 & 4 until the point is isolated.
6. Count how many times you’ve had to repeat steps 3 & 4; this quantity the isolation number
Isolating an inlier

https://quantdare.com/isolation-forest-algorithm/
Isolating an outlier

https://quantdare.com/isolation-forest-algorithm/
Isolation Forest

Very fast:

- iTrees can be built on partial models and exploit subsampling to an extent that is not feasible in existing methods.
- Since a large part of an iTree that isolates normal points is not needed for anomaly detection; it does not need to be constructed.
- A small sample size produces better iTrees because the swamping and masking effects are reduced.
  - Swamping refers to wrongly identifying normal instances as anomalies.
  - Masking is the existence of too many anomalies concealing their own presence.
- iForest utilizes no distance or density measures to detect anomalies!
- This eliminates major computational cost of distance calculation in all distance-based methods and density-based methods.

Very efficient:

- iForest has a linear time complexity with a low constant and a low memory requirement.
- iForest has the capacity to scale up to handle extremely large data size and high-dimensional problems with a large number of irrelevant attributes.

Deep Learning: Autoassociative Networks
Autoassociative Neural Networks

• Typical input data is correlated, i.e., has “features” which allow us to project it into a lower dimension

• Autoencoders: unsupervised feature learning

• Encoder is constrained, e.g., fewer neurons or sparsity

• Learns a compressed representation of the input

• Like principal components analysis (PCA), but nonlinear

Example: Vibration Data

- Use 10% validation set of normal data that the model has not been previously exposed to (i.e., was not trained on) and see how well the model can reconstruct it.
- The following plots show the raw data in blue, the reconstruction in orange, and the error (raw-reconstruction) in black.
- Because the model compresses the data significantly — it can represent a 500 point time series with just 4 values — as one might expect, some detail is lost.
Example: Vibration Data

Typical output for validation data

blue = raw ; orange = reconstruction

reconstruction error
Example: Vibration Data

- How well do the models reconstruct known faulty data that the model has not been previously exposed to (i.e., was not trained on)?
- Because the nature of the seeded fault data is different from the normal data, the models should reconstruct it less well than the normal data.
- The error between the raw data and the reconstruction can be used to create a condition indicator specific to that sensor and mounting location.
Example: Vibration Data

Typical output for mildly faulted data

blue = raw; orange = reconstruction
Example: Vibration Data

Typical output for severely faulted data
Example: Vibration Data

- The models are still able to characterize the performance of the system, even in the presence of a fault.
- However, fidelity of the reconstruction is diminished with increasing fault magnitude.
- The reconstruction works much better for a smaller fault than a larger fault.

This suggests that reconstruction error might be used as a condition indicator for helicopter bearings, gearboxes, and other rotating components.
Characterizing Performance

Performance of a model – i.e., how well the signal is reconstructed – can be characterized using mean squared reconstruction error:

$$MSRE = \frac{\sum_{i=1}^{n}(actual_i-estimate_i)^2}{n}$$

That is, the average of the squared error over some number of samples, in this case $n=100$ samples.
Mean Square Reconstruction Error

- Raw MSRE for accelerometer data
- It is easy to determine the condition of the gearbox using MSRE
- A much more illuminating way to look at the data is to consider the distribution of MSRE

[Graph showing MSRE over time with different colored dots representing normal, mild fault, and severe fault conditions]
Example: Vibration Data
Distribution of MSRE: easy to distinguish slight and severe fault
Deep Autoencoders
Deep Autoencoders
Extracted Features, Normal Bearing
Extracted Features, Abnormal Bearing

- Used autoassociative networks to construct features much better than handmade features


• Raw time series data from multiple sensors to build a robust model for anomaly detection.

• Tested on data from the NASA open database and demonstrates high fault detection rates (~97.8%) with zero false alarms.

Questions
Deep Learning II: Long Short-Term Memory
Long Short-Term Memory (LSTM)

• Long short-term memory (LSTM) networks are a variation of a recurrent neural network (RNN)
• Past state of the network is stored internally for later use

Fig. 1: (a) Long Short-term Memory Cell (b) Stacked Architecture

LSTM

LSTM have been shown to have good performance for AD on time series data.

Upside:
• Able to detect collective anomalies!

Downside:
• **Very** difficult to tune