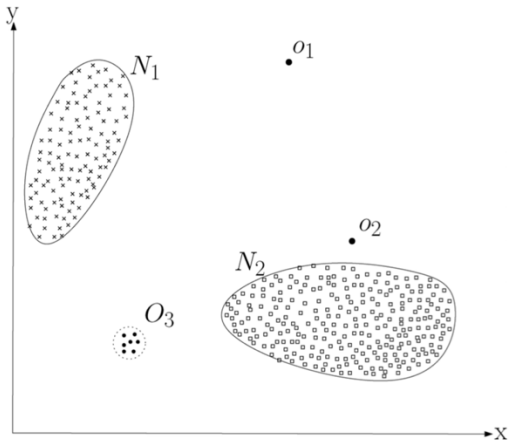
Outlier (Anomaly) Detection

Outliers refer to objects that deviate considerably from normal objects [Han, et.al.]. These outliers may be generated by different mechanisms so they do not belong to the original data set or the model. The focus of anomaly detection is to find any patterns in the datasets that do not conform to the expected behavior. Alternative terms for non-conform patterns are anomalies, outliers, discordant observation, peculiarities, etc. But the two most common terms are anomaly and outlier.

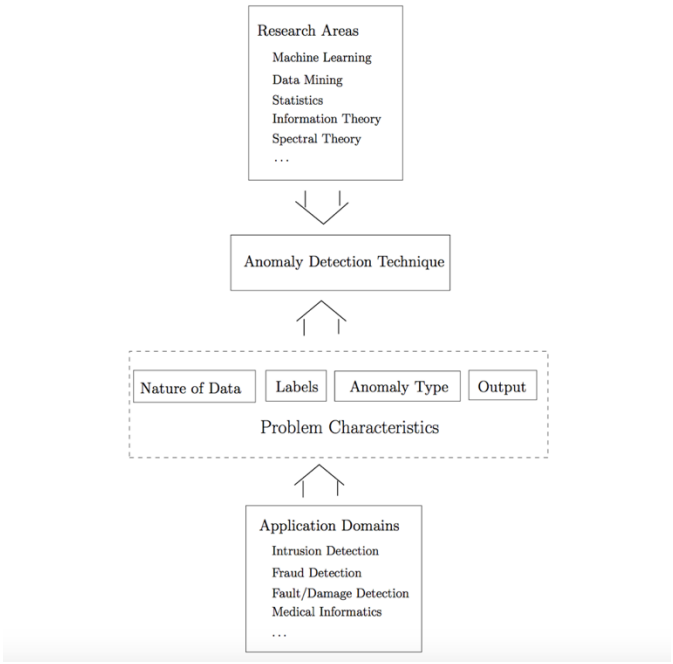


A simple example of outliers (O_1, O_2, O_3)
Source: Chandola et.al. [2009]

Outliers and noise data are not the same. Noise is random error or variance in a measured variable and it should be removed, whereas, outliers are interesting.

Anomaly detections are extensively used in many applications such as:

- Fraud detection for credit cards, insurance claims, and health care
  Abnormal patterns of using the credit cards, insurance claim.
- Medical and public health anomaly detection
  Unusual patients' symptoms or test results may imply health problems
- Intrusion detection for cyber-security
  Intrusion detection can be described as any malicious activities against systems, networks, and servers. Some characteristics of intrusion include high volume of data, false alarm rate in the input data set, missing labeled data, etc.
- Fault detection in safety critical systems
- Military surveillance for enemy activities

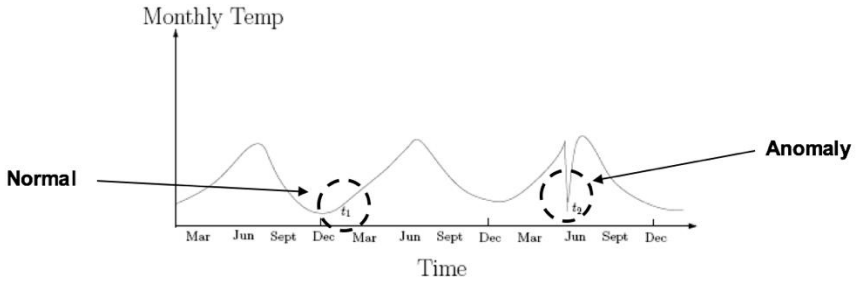


Components associated with anomaly detection Technique.
Source: Chandola et.al [2009]

Type of Outliers

- Global outlier (or point anomaly)
  An individual object that deviates significantly from the rest of the data set or not in the boundary of the normal data point regions is known as point anomaly. This is the simplest type of anomaly. Also, most research focuses on point anomalies. For example, intrusion detection in computer networks.

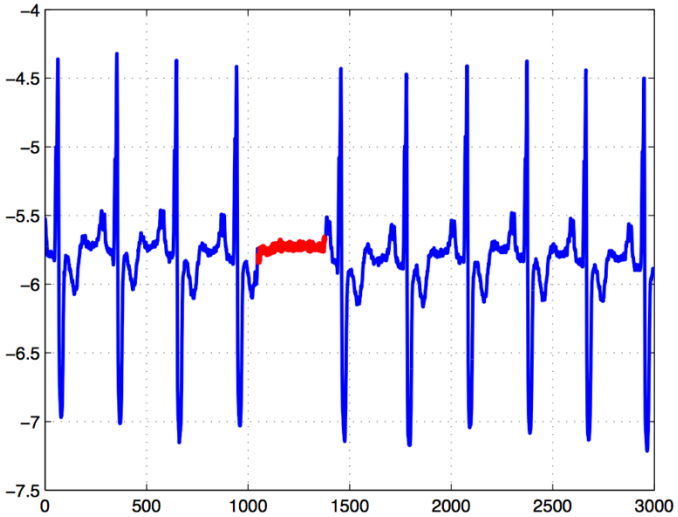
- Contextual outlier (conditional outlier)
  This refers to an object that has unusual pattern/behavior or deviates significant based on specific context or condition. The attributes of data objects can be categorized into 2 groups:
  - Contextual attributes: defines the context such as time and location.
  - Behavioral attributes: defines characteristics of the object for outlier evaluation such as temperature.



An example of contextual outliner
Source: Banerjee, et.al.

In general, context anomalies are used in time-series data and spatial data. For example, a person spends average of $500 per week but during Christmas week, the spending can go up to $3000. This spending may considered normal behavior in December but not in July. Here, the context attribute is time of spending. Another example is that the average temperature at night during winter is 30F considering normal, but the same temperature would be an anomaly during summer. A data instance is considered normal in one context but anomaly in another context.
This type of anomaly is considered a generalization of local outliers; whose density significantly deviates from its local area.

- Collective Anomalies
  This type of anomaly usually occurs in data sets where data instance is related. An individual data instance may not be anomaly by itself but when it occurs together as a collection, it becomes anomaly. For example, the low value in ECG signal is not anomaly by itself but if it happens for a long period of time, it may be anomaly.



An example of collective anomaly (low value in red color regions) from human electrocardiogram output
Source: Chandola et.al., [2009]

Outlier detection challenges:

- Modeling normal objects and outliers correctly is extremely complicated since the border between normal objects and outliers are not clear.
- Outlier detection also depends on the the applications. As a result, choice of distance measurements can be varied. For example, a small deviation may be outliers in medical data but it is larger fluctuations in marketing.
- Noise can blur or distort the distinction between normal objects and outliers. This can make the issue even more complicated.
- Understanding the degree of outliers or why they are outliers are important for outlier detection.

Labels

The label of a data instance is denoted as normal or anomaly. Labeling is usually performed manually, therefore, it requires extensive of efforts. Getting anomalous data labels are much more difficult than normal data labels.

Output of Anomaly Detection

Two different types of output produced by anomaly detection techniques are:

1) Scores. This technique assigns an anomalous score, which is the degree that an instance is considered as anomaly. The output is a list of anomalies ranking. Top few anomalies or cut off threshold is used to select anomalies for analysis.
2) Labels. A label is binary output; it is either normal or anomaly assigned to each instance in this technique.

Although outlier detection main concerns are on finding abnormal objects, their methods are based on the concept of supervised and unsupervised learning.

Outliers detection techniques can be categorized into two groups depend on [Han, 2012]:

1) Whether user-labeled examples of outliers can be obtained. If so, techniques such as supervised, semi-supervised and unsupervised methods are utilized.
2) Whether assumptions about normal data and outliers are deployed. If so, employing statistical, proximity-based, and clustering-based methods.

Supervised methods:

Outlier detection is modelled as a classification problem. Samples are divided into training and testing set. Objects that match the model are classified as normal, otherwise, they are outliers. Some challenges include imbalanced class, and try to spot as many outliers as possible. Imbalanced class happens because outliers are rare. This can be resolved by increasing outlier class using artificial outliers. When you try to catch outliers, recall is more important than accuracy. (e.g. not mislabeling normal objects are outliers).

Unsupervised methods:

Presuming that normal objects are clustered into multiple groups and an outlier is far away from any normal groups. Many clustering techniques can be used for unsupervised methods. Firstly, normal clusters are identified. Then locate outliers, which do not belong to any cluster.

Related problems include 1) differentiating noise from outliers can be difficult 2) it is costly since we have to spot normal clustering first before outliers and this is complicated since there are far fewer outliers than normal objects. An alternative and newer strategy is tackling outliers directly. In addition, this technique can not effectively detect collective outliers.

Semi-supervised methods:

In applications, where there is only a small number of labeled data such as labeled on outliers only, labeled on normal objects only, or both. These applications are considered as semi-supervised learning. If some labeled normal objects are available, use them and the proximate unlabeled objects to train a model for normal objects. Objects that are not matched with the model of normal objects are detected as outliers. If only some labeled outliers are available, it may not recognize outliers well enough. To increase the quality of outlier detection, a model from normal objects that learned from unsupervised models can be deployed.

Statistical methods:

Statistical methods or model-based methods presume that objects in the data set are generated from a stochastic process (e.g. Gaussian distribution). The basic concept is to learn a generative model, and notify objects in low probability regions as outliers. Effectiveness of these methods depends on whether the assumptions of the model are valid in the real data. Statistical methods can be further divided into parametric (assume that normal data is generated from a parametric distribution) and non-parametric methods (no or fewer assumptions about the data model using histogram or kernel density estimation to detect outliers).

I. Parametric method

1.1 Univariate (one variable or one attribute)

- z-scores

A z-score is based on z-statistic. The z-score can be computed from the following formula:

$$Z_i = \frac{x_i - \bar{x}}{s}$$

Where x_i is an observation i

$\bar{x}$ is the sample mean

s is the sample standard deviation

Some useful property of z-scores is that if the sample has a normal distribution (bell-shaped dataset), then z-scores also have a standard normal distribution. This means that 99.7% of z-scores should fall within the range $x \pm 3s$. Thus, the z-scores outside this range are considered outliers.

- Grubb's test (parametric method) for univariate outlier detection

This is a statistical method based on normal distribution. Compute a z-score of any object *x*. If z-score is larger than the specified value, *x* is considered as an outlier.

$$z \geq \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\frac{\alpha}{2N},N-2}}{N-2+t^2_{\frac{\alpha}{2N},N-2}}}$$

Where *N* is the number of objects in the data set.

1.2 Multivariate (two or more attributes)

For multivariate data, we can transform it into a univariate outlier detection problem. Apply methods such as computing Mahalaobis distance and Grubb's test to detect outliers, and chi-square test (e.g. an object is outlier if the value is large).

Proximity-based methods:

An object is an outlier if its proximity considerably deviates (far away) from most of other objects' proximity in the same data set. The performance of this method is greatly depending on the proximity measure which can be hard to obtain in some application (i.e. finding a group of outliers which close to each other can be challenging). There are two types of proximity-based methods: distance-based (i.e. an object is an outlier if there are not enough points in its neighborhood) and density-based (its density is much lower than its neighbors).

Clustering methods:

These methods assume that normal data belong to large and dense clusters, while outliers belong to small or sparse clusters, or not belonging to any clusters. There are many clustering-based outlier detection as a result of many clustering methods. Applying a straightforward clustering method to detect outliers can be costly. In addition, it does not scale well with large data sets.

Related techniques for outlier detection:

- (Interquartile) box plot is a simple but robust method. The outliers are values outside the range : $(Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR)$
  When Q_1(25^{th} percentile) is the first quartile
  Q_3(75^{th} percentile) is the third quartile, and
  IQR (Interquartile) is defined as $Q_3 - Q_1$

Evaluation of Anomaly Detection

Accuracy is not a sufficient metric (e.g. classifier which predict everything as the normal class will achieve the high accuracy rate). Therefore, the focus is on precision, recall, and F-measure.

Precision (P) = $\frac{TP}{TP+FP}$

This is also known as positive predictive value (PPV).

Recall (R) = $\frac{TP}{TP+FN}$

This is also known as detection rate (hit rate, sensitivity). It is the proportion between the correct number of detected anomalies and the total number of anomalies.

F-score = $\frac{2*R*P}{(R+P)}$

This is a harmonic mean of precision and recall.

For outlier detection examples:

Daroczi, G. (2015). Mastering Data Analysis with R. Packt Publishing. (E-book Chapter 12 – Outlier detection)

Toomey, D (2014). R for data science. Packt Publishing. (Chapter 1- Custer Analysis: Anomaly Detection)

Kejariwal, A. (2015). Introducing practical and robust anomaly detection in a time series (Blog). Retrieved from: https://blog.twitter.com/2015/introducing-practical-and-robust-anomaly-detection-in-a-time-series

References:

Chandola, V., Banerjee, A., and Kumar, V (2009). Anomaly Detection: A Survey. ACM Computing Surveys. Retrieved from: http://www-users.cs.umn.edu/~banerjee/papers/09/anomaly.pdf

Han, J. Kamber, M. and Pei, J (2012). Data Mining: Concepts and Techniques. UIUC. Retrieved from: http://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm

Kriegel, H., Kroger, P. and Zimek, A. (2010). Outlier Detection Technique. 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Tutorial Notes. Washington D.C. Retrieved from: http://www.dbs.ifi.lmu.de/~zimek/publications/KDD2010/kdd10-outlier-tutorial.pdf

Tan, Steinbach, and Kumar. Data Mining Anomaly Detection. Retrieved from: https://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap10_anomaly_detection.ppt

Banerjee, A. Chandola, V. Kumar, V et. Al. Anomaly Detection: A tutorial. University of Minnesota.