Anomaly Detection: A Survey

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains.

Anomaly detection finds extensive use in a wide variety of applications such as fraud detection for credit cards, insurance or health care, intrusion detection for cyber-security, fault detection in safety critical systems, and military surveillance for enemy activities.

The importance of anomaly detection is due to the fact that anomalies in data translate to significant (and often critical) actionable information in a wide variety of application domains. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination.

A straightforward anomaly detection approach is to define a region representing normal behavior and declare any observation in the data which does not belong to this normal region as an anomaly. But several factors make this apparently simple approach very challenging:

- Defining a normal region which encompasses every possible normal behavior is very difficult. In addition, the boundary between normal and anomalous behavior is often not precise. Thus an anomalous observation which lies close to the boundary can actually be normal, and vice-versa.

- When anomalies are the result of malicious actions, the malicious adversaries often adapt themselves to make the anomalous observations appear like normal, thereby making the task of defining normal behavior more difficult.

- In many domains normal behavior keeps evolving and a current notion of normal behavior might not be sufficiently representative in the future.

- Often the data contains noise which tends to be similar to the actual anomalies and hence is difficult to distinguish and remove.

A key aspect of any anomaly detection technique is the nature of the input data. Input is generally a collection of data instances. Each data instance can be described using a set of attributes. The attributes can be of different types such as binary, categorical or continuous.

Data instances can be related to each other. Some examples are sequence data, spatial data, and graph data. In sequence data, the data instances are linearly ordered, e.g., time-series data. When the spatial data has a temporal (sequential) component it is referred to as spatio-temporal data, e.g., climate data. In graph data, data instances are represented as vertices in a graph and are connected to other vertices with edges.

Anomalies can be classified into the following three categories:

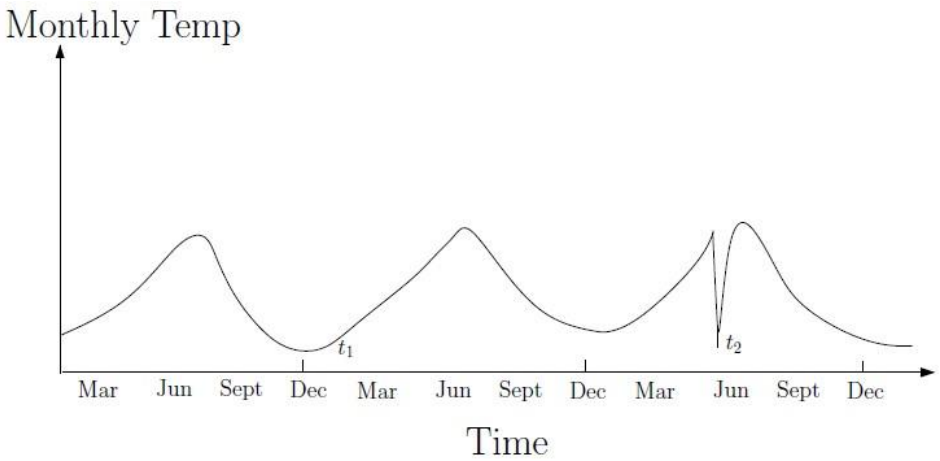
1. **Point Anomalies**. If an individual data instance is considered as anomalous with respect to the rest of the data, then the instance is called a point anomaly. This is the simplest type of anomaly and is the focus of the majority of research on anomaly detection.

2. **Contextual Anomalies**. If a data instance is anomalous in a specific context (but not otherwise), then it is termed a contextual anomaly (also referred to as a conditional anomaly). Each data instance is defined using the following two sets of attributes:

   - **Contextual attributes**. The contextual attributes are used to determine the context (or neighborhood) for that instance. For example, in spatial data sets, the longitude and latitude of a location are the contextual attributes.

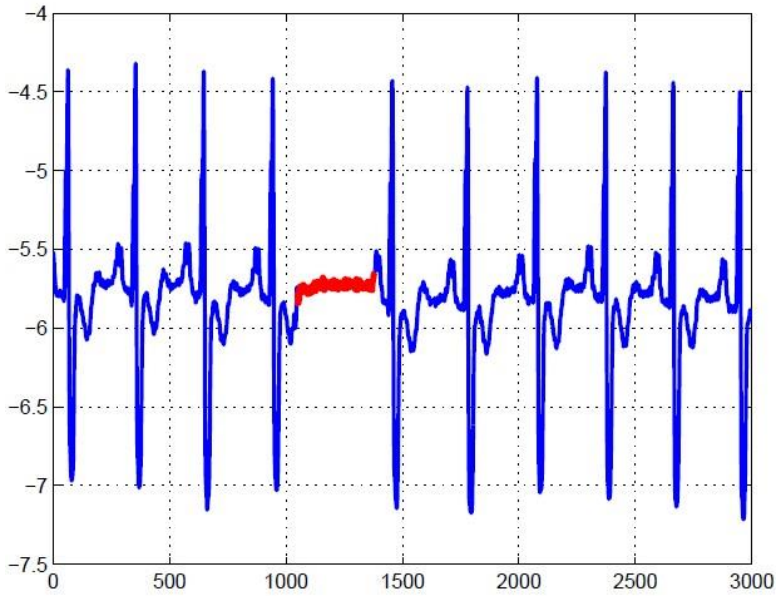
   - **Behavioral attributes**. The behavioral attributes define the noncontextual characteristics of an instance. For example, in a spatial data set describing the average rainfall of the entire world, the amount of rainfall at any location is a behavioral attribute.

   The anomalous behavior is determined using the values for the behavioral attributes within a specific context. A data instance might be a contextual anomaly in a given context, but an identical data instance (in terms of behavioral attributes) could be considered normal in a different context. This property is key in identifying contextual and behavioral attributes for a contextual anomaly detection technique.



A temperature of 35F might be normal during the winter (at time t1) at that place, but the same value during summer (at time t2) would be an anomaly.

3. **Collective Anomalies.** If a collection of related data instances is anomalous with respect to the entire data set, it is termed 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 highlighted region denotes an anomaly because the same low value exists for an abnormally long time. Note that that low value by itself is not an anomaly.

Anomaly detection techniques can operate in one of the following three modes:

- **Supervised anomaly detection**. Techniques trained in supervised mode assume the availability of a training data set which has labeled instances for normal as well as the anomaly class.

- **Semi-Supervised anomaly detection**. Techniques that operate in a semi-supervised mode assume that the training data has labeled instances for only the normal class.

- **Unsupervised anomaly detection**. Techniques that operate in unsupervised mode do not require training data, and thus are most widely applicable. The techniques in this category make the implicit assumption that normal instances are far more frequent than anomalies in the test data.

An important aspect for any anomaly detection technique is the manner in which the anomalies are reported. Typically, the outputs produced by anomaly detection techniques are one of the following two types:

- **Scores**. Scoring techniques assign an anomaly score to each instance in the test data depending on the degree to which that instance is considered an anomaly. Thus the output of such techniques is a ranked list of anomalies. An analyst may choose to either analyze top few anomalies or use a cut-off threshold to select the anomalies.

- **Labels**. Techniques in this category assign a binary label (normal or anomalous) to each test instance.

Intrusion detection refers to detection of malicious activity (break-ins, penetrations, and other forms of computer abuse) in a computer related system. The key challenge for anomaly detection in this domain is the huge volume of data. The anomaly detection techniques need to be computationally efficient to handle these large sized inputs. Another issue which arises because of the large sized input is the false alarm rate. Since the data amounts to millions of data objects, a few percent of false alarms can make analysis overwhelming for an analyst.

Denning classifies intrusion detection systems into host based and network based intrusion detection systems.

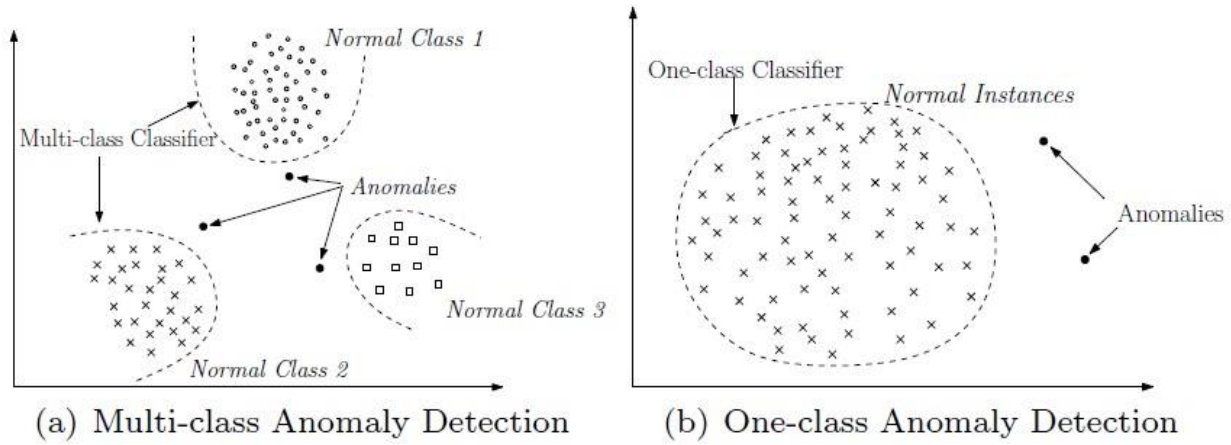
- **Host Based Intrusion Detection Systems**. Such systems deal with operating system call traces. The intrusions are in the form of anomalous subsequences (collective anomalies) of the traces. The anomalous subsequences translate to malicious programs, unauthorized behavior and policy violations. It is the co-occurrence of events which is the key factor in differentiating between normal and anomalous behavior. A key characteristic of the data in this domain is that the data can be typically profiled at different levels such as program level or user level.

- **Network Intrusion Detection Systems**. These systems deal with detecting intrusions in network data. The intrusions typically occur as anomalous patterns (point anomalies) though certain techniques model the data in a sequential fashion and detect anomalous subsequences (collective anomalies). The data available for intrusion detection systems can be at different levels of granularity, e.g., packet level traces and netflow data.

Classification-based anomaly detection techniques operate under the following general assumption: A classifier that can distinguish between normal and anomalous classes can be learnt in the given feature space. Classification-based anomaly detection techniques can be grouped into two broad categories: multi-class and one-class anomaly detection techniques.

- **Multi-class classification**-based anomaly detection techniques assume that the training data contains labeled instances belonging to multiple normal classes. Such anomaly detection techniques learn a classifier to distinguish between each normal class against the rest of the classes. A test instance is considered anomalous if its not classified as normal by any of the classifiers.

- **One-class classification**-based anomaly detection techniques assume that all training instances have only one class label. Such techniques learn a discriminative boundary around the normal instances using a one-class classification algorithm. Any test instance that does not fall within the learnt boundary is declared as anomalous.



(a) Multi-class Anomaly Detection (b) One-class Anomaly Detection

Classification-Based Detection Techniques:

A **basic multi-class anomaly detection** technique using neural networks operates in two steps. First, a neural network is trained on the normal training data to learn the different normal classes. Second, each test instance is provided as an input to the neural network. If the network accepts the test input, it is normal and if the network rejects a test input, it is an anomaly.

Rule based anomaly detection techniques learn rules that capture the normal behavior of a system. A test instance that is not covered by any such rule is considered as an anomaly. Rule based techniques have been applied in multi-class as well as one-class setting.

A **basic multi-class rule based technique** consists of two steps. First step is to learn rules from the training data using a rule learning algorithm. Each rule has an associated confidence value which is proportional to ratio between the number of training instances correctly classified by the rule and the total number of training instances covered by the rule. Second step is to find, for each test instance, the rule that best captures the test instance. The inverse of the confidence associated with the best rule is the anomaly score of the test instance.

Association rule mining has been used for one-class anomaly detection by generating rules from the data in an unsupervised fashion. Association rules are generated from a categorical data set. To ensure that the rules correspond to strong patterns, a support threshold is used to prune out rules with low support.

Disadvantages of classification-based techniques are as follows:

- Multi-class classification-based techniques rely on availability of accurate labels for various normal classes, which is often not possible.

- Classification-based techniques assign a label to each test instance, which can also become a disadvantage when a meaningful anomaly score is desired for the test instances.

The concept of **nearest neighbor analysis** has been used in several anomaly detection techniques. Such techniques are based on the following key assumption: Normal data instances occur in dense neighborhoods, while anomalies occur far from their closest neighbors. Nearest neighbor based anomaly detection techniques require a distance or similarity measure defined between two data instances. Nearest neighbor-based anomaly detection techniques can be broadly grouped into two categories: (1) Techniques that use the distance of a data instance to its *kth* nearest neighbor as the anomaly score. (2) Techniques that compute the relative density of each data instance to compute its anomaly score.

Density-based anomaly detection techniques estimate the density of the neighborhood of each data instance. An instance that lies in a neighborhood with low density is declared to be anomalous while an instance that lies in a dense neighborhood is declared to be normal. Density based techniques perform poorly if the data has regions of varying densities. To handle the issue of varying densities in the data set, a set of techniques have been proposed to compute density of instances relative to the density of their neighbors.

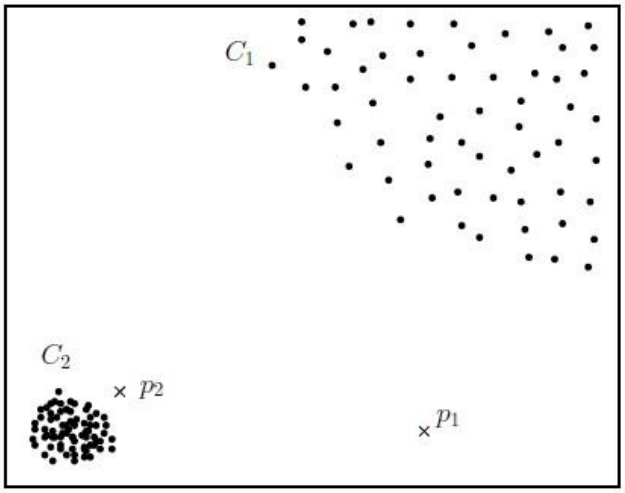


Fig. 7. Advantage of local density based techniques over global density based techniques.

Advantages of nearest neighbor based techniques: Adapting nearest neighbor based techniques to a different data type is straight-forward, and primarily requires defining an appropriate distance measure for the given data. The disadvantages of nearest neighbor based techniques: For unsupervised techniques, if the data has normal instances that do not have enough close neighbors or if the data has anomalies that have enough close neighbors, the technique fails to label them correctly, resulting in missed anomalies.

Clustering is used to group similar data instances into clusters. The first category of clustering based techniques rely on the following assumption: Normal data instances belong to a cluster in the data, while anomalies do not belong to any cluster. Techniques based on the above assumption apply a known clustering based algorithm to the data set and declare any data instance that does not belong to any cluster as anomalous. Note that if the anomalies in the data form clusters by themselves, the above discussed technique will not be able to detect such anomalies.

To address this issue a category of clustering based techniques have been proposed that rely on the following assumption: Normal data instances belong to large and dense clusters, while anomalies either belong to small or sparse clusters. Techniques based on the above assumption declare instances belonging to clusters whose size and/or density is below a threshold as anomalous.

Distinction between clustering-based and nearest neighbor-based techniques: the key difference between the two techniques is that clustering based techniques evaluate each instance with respect to the cluster it belongs to, while nearest neighbor based techniques analyze each instance with respect to its local neighborhood.

Disadvantages of clustering based techniques: several clustering algorithms force every instance to be assigned to some cluster. This might result in anomalies getting assigned to a large cluster, thereby being considered as normal instances by techniques that operate under the assumption that anomalies do not belong to any cluster.

Statistical anomaly detection techniques are based on the following key assumption: Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model. Statistical techniques fit a statistical model (usually for normal behavior) to the given data and then apply a statistical inference test to determine if an unseen instance belongs to this model or not. Instances that have a low probability to be generated from the learnt model, based on the applied test statistic, are declared as anomalies.

The **basic regression** model based anomaly detection technique consists of two steps. In the first step, a regression model is fitted on the data. In the second step, for each test instance, the residual for the test instance is used to determine the anomaly score. The residual is the part of the instance which is not explained by the regression model. The magnitude of the residual can be used as the anomaly score for the test instance.

Histogram Based. The simplest non-parametric statistical technique is to use histograms to maintain a profile of the normal data. Such techniques are also referred to as frequency based or counting based. A basic histogram based anomaly detection technique for univariate data consists of two steps. The first step involves building a histogram based on the different values taken by that feature in the training data. In the second step, the technique checks if a test instance falls in any one of the bins of the histogram. If it does, the test instance is normal, otherwise it is anomalous.

Histogram based techniques are relatively simple to implement, but a key shortcoming of such techniques for multivariate data is that they are not able to capture the interactions between different attributes. An anomaly might have attribute values that are individually very frequent, but their combination is very rare, but an attribute-wise histogram based technique would not be able to detect such anomalies.

Contextual anomalies require that the data has a set of contextual attributes (to define a context), and a set of behavioral attributes (to detect anomalies within a context). Some of the ways in which contextual attributes can be defined are:

1. **Spatial**: The data has spatial attributes, which define the location of a data instance and hence a spatial neighborhood.

2. **Graphs**: The edges that connect nodes (data instances) define neighborhood for each node.

3. **Sequential**: The data is sequential, i.e., the contextual attributes of a data instance is its position in the sequence.

4. **Profile**: Often times the data might not have an explicit spatial or sequential structure, but can still be segmented or clustered into components using a set of contextual attributes. These attributes are typically used to profile and group users in activity monitoring systems. The users are then analyzed within their group for anomalies.

The key advantage of contextual anomaly detection techniques is that they allow a natural definition of an anomaly in many real life applications where data instances tend to be similar within a context. Such techniques are able to detect anomalies that might not be detected by point anomaly detection techniques that take a global view of the data. The disadvantage of contextual anomaly detection techniques is that they are applicable only when a context can be defined.

Collective anomalies are a subset of instances that occur together as a collection and whose occurrence is not normal with respect to a normal behavior. The individual instances belonging to this collection are not necessarily anomalies by themselves, but it is their co-occurrence in a particular form that makes them anomalies. Collective anomaly detection problem is more challenging than point and contextual anomaly detection because it involves exploring structure in the data for anomalous regions.

Anomalous sequence in a set of sequences. The objective of the techniques in this category is to detect anomalous sequences from a given set of sequences. Key challenges faced by techniques in this category are: (1) The sequences might not be of equal length. (2) The test sequences may not be aligned with each other or with normal sequences. For example, the first event in one sequence might correspond to the third event in another sequence.

Anomalous subsequences in a long sequence. The objective of techniques belonging to this category is to detect a subsequence within a given sequence which is anomalous with respect to the rest of the sequence. Such anomalous subsequences have also been referred as discords. The underlying assumption is that the normal behavior of the time-series follows a defined pattern. A subsequence within the long sequence which does not conform to this pattern is an anomaly.