Introduction: Anomaly Detection

Anomaly detection is a technique used to identify unusual patterns that do not conform to expected behavior, called outliers. It has many applications in business, from **intrusion detection** (identifying strange patterns in network traffic that could signal a hack) to system health monitoring (spotting a malignant tumor in an MRI scan), and from **fraud detection** in credit card transactions to fault detection in operating environments.

In data mining, anomaly detection (also outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.

A cluster analysis algorithm may be able to detect the micro clusters formed by these patterns

- Anomaly detection is a form of classification.
- Is the process to localize objects that are different from other objects (anomalies).
- The set of data points that are considerably different than the remainder of the data are anomalies/outliers.
- Anomaly detection is the process of detecting something unusual relative to something expected.
- The goal of anomaly detection is to identify cases that are unusual within data that is seemingly homogeneous.

What Are Anomalies?

Before getting started, it is important to establish some boundaries on the definition of an anomaly. Anomalies can be broadly categorized as:

- **Point anomalies:** A single instance of data is an anomalous if it's too far off from the rest. Business use case: *Detecting credit card fraud based on "amount spent."*
- **Contextual anomalies:** The abnormality is context specific. This type of anomaly is common in time-series data. Business use case: *Spending Rs.100 on food every day during the holiday season is normal, but may be odd otherwise.*
- **Collective anomalies:** A set of data instances collectively helps in detecting anomalies. Business use case: *Someone is trying to copy data form a remote machine to a local host unexpectedly, an anomaly that would be flagged as a potential cyber-attack.*

Anomaly detection is similar to — but not entirely the same as — noise removal and novelty detection.

- **Novelty detection** is concerned with identifying an unobserved pattern in new observations not included in training data — like a sudden interest in a new channel on YouTube during Christmas, for instance.
- **Noise removal** (NR) is the process of immunizing analysis from the occurrence of unwanted observations; in other words, removing noise from an otherwise meaningful signal.

Application of Anomaly detection

- Intrusion detection: suspicious traffic across the network, such as an unusually high rate of TCP connections.
- Fraud detection: Financial statement fraud, Credit card fraud, Insurance fraud, Corporate fraud, Securities and commodities fraud
- Fault detection,
- System health monitoring: use medical statistic reports for diagnosis
- Event detection in sensor networks,
- Medicine - use unusual symptoms or test result to indicate potential health problems
- Detecting ecosystem disturbances: try to predict events like hurricanes and floods

Why is Anomaly Detection important?

- to detect problems
- to detect new phenomenon
- to discover unusual behavior in data

Challenges

- How many outliers are there in the data?
- Method is unsupervised
- There are considerably more “normal” observations than “abnormal” observations (outliers/anomalies) in the data.

Anomaly Detection Techniques

i. **Unsupervised anomaly detection techniques** detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set.

ii. **Supervised anomaly detection techniques** require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier (the key difference to many other statistical classification problems is the inherent unbalanced nature of outlier detection).

iii. **Semi-supervised anomaly detection techniques** construct a model representing normal behavior from a given normal training data set, and then testing the likelihood of a test instance to be generated by the learnt model.

Anomaly Detection Schemes

General Steps

- Build a profile of the "normal" behavior, profile can be patterns or summary statistics for the overall population.
- Use the "normal" profile to detect anomalies, anomalies are observations whose characteristics differ significantly from the normal profile

Types of anomaly detection schemes

i. **Graphical based :**
   - Box plot (1-D),
   - Scatter plot (2-D),
   - Spin plot (3-D)

ii. **Statistical-based :**
   - Assume a parametric model describing the distribution of the data e.g., normal distribution.
   - A statistical test that depends on:
     - Data distribution.
     - Parameter of distribution (e.g., mean, variance).
     - Number of expected outliers (confidence limit)

       a. Grubbs' Test:
          - Detect outliers in univariate data.
          - Assume data comes from normal distribution.
          - Detects one outlier at a time, remove the outlier, and repeat.

       b. Likelihood Approach
          - Assume the data set D contains samples from a mixture of two probability distributions:
            - M (majority distribution)
            - A (anomalous distribution)

       General Approach:
       – Initially, assume all the data points belong to M
       – Let Lt(D) be the log likelihood of D at time t
       - Let Lt+1 (D) be the new log likelihood.
       - Compute the difference, $\Delta = Lt(D) - Lt+1(D)$
       - If $\Delta > c$ (some threshold), then Xt is declared as an anomaly and moved permanently from M to A

       Limitations of Statistical Approaches
          - Most of the tests are for a single attribute
          - In many cases, data distribution may not be known
          - For high dimensional data, it may be difficult to estimate the true distribution

iii. **Distance-based**: Data is represented as a vector of features. Three major approaches
   - Nearest-neighbor based
   - Density based
   - Clustering based

iv. **Model-based :**
   - An anomaly detection model predicts whether a data point is typical for a given distribution or not.
   - An atypical data point can be either an outlier or an example of a previously unseen class.
   - Normally, a classification model must be trained on data that includes both examples and counter-examples for each class so that the model can learn to distinguish between them.
     For example, a model that predicts side effects of a medication should be trained on data that includes a wide range of responses to the medication.

v. **Convex Hull Method**
   - Extreme points are assumed to be outliers. Use convex hull method to detect extreme values.
   - Major limitation is if the outlier occurs in the middle of the data.

Issues

a. **Number of Attributes**: Since an object may have many attributes, it may have anomalous values for some attributes; an object may be anomalous even if none of its attribute values are individually anomalous.
b. **Global Vs Local Perspective**: An object may seem unusual with respect to all objects, but not with respect to its local neighbors.
c. **Degree of Anomaly**: Some objects are more extreme anomalies than others;
d. **One at Time Vs Many at Once**: Is it better to remove anomalous objects one at a time or identify a collection of objects together?
e. **Evaluation**: Finding a good measure of evaluation for the process of anomaly detection when class labels are available and when class labels are not available.

f. **Efficiency**: calculate the computational cost of the process of anomaly detection scheme.

Base Rate Fallacy

- The base-rate fallacy is people's tendency to ignore base rates in favor of individuating information when such is available rather than integrate the two. This tendency has important implications for understanding judgment phenomena in many clinical, legal, and social-psychological settings.
- Base rate fallacy, also called base rate neglect or base rate bias, is a formal fallacy. If presented with related base rate information and specific information, the mind tends to ignore the former and focus on the latter.

DIFFERENT TYPES OF INTRUSION DETECTION SYSTEMS

Intrusion detection is defined as real-time monitoring and analysis of network activity and data for potential vulnerabilities and attacks in progress. One major limitation of current intrusion detection system (IDS) technologies is the requirement to filter false alarms lest the operator (system or security administrator) be overwhelmed with data. IDSes are classified in many different ways, including active and passive, network-based and host-based, and knowledge-based and behavior-based:

- **ACTIVE AND PASSIVE IDS**

An ***active*** IDS (now more commonly known as an intrusion prevention system — IPS) is a system that's configured to automatically block suspected attacks in progress without any intervention required by an operator. IPS has the advantage of providing real-time corrective action in response to an attack but has many disadvantages as well. An IPS must be placed in-line along a network boundary; thus, the IPS itself is susceptible to attack. Also, if false alarms and legitimate traffic haven't been properly identified and filtered, authorized users and applications may be improperly denied access. Finally, the IPS itself may be used to effect a *Denial of Service* (DoS) attack by intentionally flooding the system with alarms that cause it to block connections until no connections or bandwidth are available.

A ***passive*** IDS is a system that's configured only to monitor and analyze network traffic activity and alert an operator to potential vulnerabilities and attacks. It isn't capable of performing any protective or corrective functions on its own. The major advantages of passive IDSes are that these systems can be easily and rapidly deployed and are not normally susceptible to attack themselves.

- **NETWORK-BASED AND HOST-BASED IDS**

A ***network-based*** IDS usually consists of a network appliance (or sensor) with a Network Interface Card (NIC) operating in promiscuous mode and a separate management interface. The IDS is placed along a network segment or boundary and monitors all traffic on that segment.

A ***host-based*** IDS requires small programs (or *agents*) to be installed on individual systems to be monitored. The agents monitor the operating system and write data to log files and/or trigger alarms. A host-based IDS can only monitor the individual host systems on which the agents are installed; it doesn't monitor the entire network.

- **KNOWLEDGE-BASED AND BEHAVIOR-BASED IDS**

A ***knowledge-based*** (or ***signature-based***) IDS references a database of previous attack profiles and known system vulnerabilities to identify active intrusion attempts. Knowledge-based IDS is currently more common than behavior-based IDS. Advantages of knowledge-based systems include the following:

- It has lower false alarm rates than behavior-based IDS.
- Alarms are more standardized and more easily understood than behavior-based IDS.

Disadvantages of knowledge-based systems include these:

- Signature database must be continually updated and maintained.
- New, unique, or original attacks may not be detected or may be improperly classified.

A ***behavior-based*** (or ***statistical anomaly–based***) IDS references a baseline or learned pattern of normal system activity to identify active intrusion attempts. Deviations from this baseline or pattern cause an alarm to be triggered. Advantages of behavior-based systems include that they

- Dynamically adapt to new, unique, or original attacks.
- Are less dependent on identifying specific operating system vulnerabilities.

Disadvantages of behavior-based systems include

- Higher false alarm rates than knowledge-based IDSes.
- Usage patterns that may change often and may not be static enough to implement an effective behavior-based IDS.

Example

A group of policemen have breathalyzers displaying false drunkenness in 5% of the cases in which the driver is sober. However, the breathalyzers never fail to detect a truly drunk person. 1/1000 of drivers are driving drunk. Suppose the policemen then stop a driver at random, and force the driver to take a breathalyzer test. It indicates that the driver is drunk. We assume you don't know anything else about him or her. How high is the probability he or she really is drunk?

Many would answer as high as 0.95, but the correct probability is about 0.02.

To find the correct answer, one should use Bayes' theorem. The goal is to find the probability that the driver is drunk given that the breathalyzer indicated he/she is drunk, which can be represented as

$$p(drunk|D)$$

where "D" means that the breathalyzer indicates that the driver is drunk.

Using Bayes' Theorem ,

$$p(drunk|D) = \frac{p(D|drunk)\,p(drunk)}{p(D)}$$

We have,

$$p(drunk) = 0.001$$
$$p(sober) = 0.999$$
$$p(D|drunk) = 1.00$$
$$p(D|sober) = 0.05$$

$$p(D) = p(D|drunk)\,p(drunk) + p(D|sober)\,p(sober)$$
$$p(D) = 0.05095$$

Putting values into Bayes' Theorem, we get

$$p(drunk|D) = 0.019627.$$

A more intuitive explanation: in average, for every 1000 drivers tested,

- 1 driver is drunk, and it is 100% certain that for that driver there is a true positive test result, so there is 1 true positive test result
- 999 drivers are not drunk, and among those drivers there are 5% false positive test results, so there are 49.95 false positive test results therefore the probability that one of the drivers among the 1 + 49.95 = 50.95 positive test results really is drunk is . The validity of this result does, however, hinge on the validity of the initial assumption that the policemen stopped the driver truly at random, and not because of bad driving. If that or another non-arbitrary reason for stopping the driver was present, then the calculation also involves the probability of a drunk driver driving competently and a non-drunk driver driving competently.