|  |
| --- |
| **ANOMALY DETECTION**  **Shinde Yash Vikas**  *Department of Computer Science and Engineering,*  *Lovely Professional University, Phagwara, Punjab.*  [*Yashshinde227722@gmail.com*](mailto:Yashshinde227722@gmail.com) |

**What are Anomalies ?**

In Data, Anomalies are patterns that do not comply with rules to understandable concept of normal behavior. Figure 1, displaying the required anomalies in a 2-dimentional data set. Hence, the data has two normal regions, *N*1 and *N*2, since almost all observations lie in these two regions. Data points that are significantly far away from the respected regions that are anomalies. For e.g., points *o*1 and *o*2, and points in region *O*3 in the given figure.

y



*N*1

*o*1

*o*2

*O*3

*N*2

x

Fig. 1. A simple example of anomalies in a 2-dimensional data set.

In the data, Anomalies might be obtained on basis of several factors, like malicious activity, e.g., credit card fraud, terrorist activities or system breakdown, but all of the causes have common traits that they are *interesting* to the analyst. The “interestingness” or real life importance of anomalies is a main property of anomaly detection.

Anomaly detection is somehow related to, but dissimilar from *noise removal* and *noise accommodation*, both of which deal with unwanted *noise* present in the data. We can determine noise as a event in data which is not of interest to the analyst, but gives a delay to analyze the data. Noise removal is operated due to the requirement of removal of dangerous objects that may acts as obstacles before any examination is performed on the data. Noise accommodation look upon injecting a statistical model estimation opposed to anomalous observations.

**Type of Anomaly**

Anomalies are categorized into 3 types:

***1.Point Anomalies:***If the single data object is treated as anomalous with respect to the remaining of data, then the object is said to be a point anomaly. Mostly, it is concentrated for the research part on detection and the easiest one. Let’s take a real-life instance that is credit card fraud detector. Now the dataset respected to once credit card transactions having only one feature defined i.e. amount spent under assumption then the transaction for which the amount spending is very high with respect to medium range of expense for that individual it will be a point anomaly.

***2.Contextual Anomalies***: In a specific context, if a data object is anomalous then, it is said to be contextual anomaly*.*

The concept of a context is produced by the structure in the data set and therefore, as a share of problem formulation it should be specified. Each data object is explained using following two sets of attributes:

1. ***Contextual attributes*:** It is used to find the neighborhood for that object which is defined. Let’s take an instance of spatial data sets and time-series data sets, contextual attributes consist of longitude and latitude of a location and time which gives the position of an object on the entire sequence respectively.
2. ***Behavioral attributes*:** On the opposite to contextual, the non-contextual characteristics of an object is called as Behavioral attributes. Taking the same dataset (spatial data set) which depicting the average rainfall of the entire world, the amount of rainfall at any location is said to be behavioral attribute.

# Monthly Temp

t1

t2

Mar Jun Sept Dec Mar Jun Sept Dec Mar Jun Sept Dec

# Time

Fig. 3. Contextual anomaly *t*2 in a temperature time series. Note that 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 as an anomaly

***3.Collective Anomalies:***If a collection of related data objects is anomalous with respect to the entire data set, it is called as a collective anomaly. In a collective anomaly, as single data objects may or may not be anomalies by themselves, but if they occur together as a whole collection then, the corresponding collection is anomalous.

**Challenges**

Anomalies are patterns that do not comply with rules to a understandable concept of normal behavior. So, we can say that anomaly detection approach is very straight forward in order to define a region representing medium behavior and decide any objects in the data that are under analysis does not belong to this normal region as an anomaly. Still, there are several factors that make this simple approach very difficult to deal:

— The determination of a normal region which holds all the possible normal behavior is very hard to apply. In addition, the boundary which got established between normal and anomalous behavior is often not accurate. Thus, an anomalous observation can be normal in actual that lies close to the boundary and vice-versa as well.

—When anomalies appear as a output of malicious actions, malicious adversaries then, they frequently acquire themselves to make the anomalous objects appear like normal, hence the problem of finding normal behavior become more complex.

— In many domains, evolution of normal behaviour is happening day by day and the present concept of it might not be completely formable in future.

— Output data availability of models used by the anomaly is often a big problem at the time of training or validation.

— The data which consist of noise works identical to the actual anomalies and therefore, it

is not easy to find the difference and to remove the same.

Hence, because of existence of the challenges mentioned above, in general, the anomaly detection problem is very hectic to solve. In addition, most of the detection techniques can able to form some particular formulations only. The formulation is obtained by various factors such as nature of the data, availability of labeled data, type of anomalies to be detected, etc. Required concepts are acquired from statistics, machine learning, data mining, information theory, etc by the Researchers in order to apply them on different problem formulations*.*

Research Areas Cloud Computing

Machine Learning Data Mining

Spectral Theory

Anomaly Detection Technique

|  |  |
| --- | --- |
|  | |
| Nature of Data | Labels Anomaly Type Output |
| Problem Characteristics | |

Application Domains Intrusion Detection Fraud Detection Fault/Damage Detection

Medical Informatics

. . .

**Solution for Anomaly Detection in Clouds**

Anomaly detection: In order to efficiently detect the potential anomalies, we perform large-scale offline performance testing and also create an online detection method. i) Offline testing. The purpose is to find the key performance bottleneck and quantify comparison between difference hardware and software. We first propose a three-layer benchmarking methodology to fully evaluates cloud performance and then present a new benchmark suite that measures various scenarios, such as single machine virtualization, server consolidation, VM mapping, VM live migration, HPC virtual clusters and Hadoop virtual cluster. Finally, we introduce a performance testing toolkit to automate the benchmarking process. ii) Online detection. The purpose is to monitor applications in real time and quickly detect potential faults. We propose a quantile regression based online anomaly detection method and did a case study on 67 real Yahoo! anomaly traffic datasets.

**2)Anomaly inference:**

Researchers have put forward an inference method based upon graph dependency anomaly and the same dependency provides an interactive relationship and execution path to follow, and both the provided things are used to carry out fault localization. We have three ways to draw the dependency graph such as instrumentation, using the configuration files after the extraction and last but not list is analysing the network traffic of the dataset. In order to pay attention or observe the traffic, light-weight agents are created and using sampling technique, overheads also got decreased.

**3)Anomaly recovery:**

Checkpoint /Restart are known as traditional recovery methods which generate high overheads and due to this, it is applicable for the latency-sensitive applications. Still, we can produce two solution to this problem: using the cache-aware fault isolation virtual machine and the other is fault recovery using migration.

**1**.**Cash-Aware Fault Isolation**: Initially we need to determine about the isolation and then by keeping the aim of modification of fault isolation, we have to import virtual machine core scheduling process.

**2.Migration based Fault Recovery**: Taking the help of live migrating movements, fault recovery is being carried out effectively. It also has various demerits such as digital services and can be implemented in large-scale cloud datacenters as well.

**Applications of Anomaly Detection**

***Credit Card Fraud Detection:***

In this domain, anomaly detection techniques are applied to detect fraudulent credit card applications or fraudulent credit card usage (associated with credit card thefts). Detecting fraudulent credit card applications is similar to detecting insurance fraud.

The data typically comprises of records defined over several dimensions such as the user ID, amount spent, time between consecutive card usage, etc. The frauds are typically reflected in transactional records (point anomalies) and correspond to high payments, purchase of items never purchased by the user before, high rate of purchase, etc. The credit companies have complete data available and also have labeled records. Moreover, the data falls into distinct profiles based on the credit card user. Hence profiling and clustering based techniques are typically used in this domain.

The challenge associated with detecting unauthorized credit card usage is that it requires online detection of fraud as soon as the fraudulent transaction takes place. Anomaly detection techniques have been applied in two different ways to address this problem. The first one is known as *by owner* in which each credit card user is profiled based on his/her credit card usage history. Any new transaction is compared to the user’s profile and flagged as an anomaly if it does not match the profile. This approach is typically expensive since it requires querying a central data repository, every time a user makes a transaction. Another approach known as *by-operation* detects anomalies from among transactions taking place at a specific geographic location. Both *by-user* and *by-operation* techniques detect contextual anomalies. In the first case the context is a user, while in the second case, the context is the geographic location.

***Mobile Phone Fraud Detection:***

This is a monitoring type of activity problem. Here, we need to do scanning of large set of accounts and then illustrate the behaviour of each object. When the account appears to be misused by someone then, alarm should be issued.

Calling activity can be form in many ways, but frequently it gives details in the form of call records. Each call record includes array of features which can be discrete (calling city) or continuous (call duration).Ancient representation is not essential in this respective domain. Calls can be aggregated by time, for example into call-hours or call-days or user or area depending on the granularity desired. The anomalies correspond to high volume of calls or calls made to unlikely destinations.