Anomaly Detection: A Survey

Deepak Surana

Department of Computer Science, Nirma University
Sarkhej-Gandhinagar Highway, Ahmedabad - 382 481, India
deepaksurana18@gmail.com

***Abstract*— Anomaly Detection is the process of identifying unusual behavior. Anomalies are patterns in data that do not conform to a well-defined notion of normal behavior. Anomalies are the data points that are considerably different from rest of the data chunk. These normally represents important information because of which the analyst is interested in detecting these anomalies precisely and efficiently, and then deciding on the action in order to respond. There are situations where enormous impact can be made by identifying anomalies in a timely and right manner, e.g. early detection of fraud credit card transaction can prevent financial loss to both card holder as well as to the banking institution. Its applications has generated the interest for designing of efficient methods of detections anomalies and in turn anomaly detection systems.**

Anomaly detection systems, a subset of intrusion detection systems, model the typical system/network behavior which make them extremely effective in finding and foiling both known as well as unknown or "zero day" attacks. It is widely used in real-life domains, for example, to identify fraud, customer behavioral change, and cyber intrusions. Firstly, they are capable of detecting insider attacks, e.g. an alarm will be generated when a user is using any stolen account and thus performing actions that are beyond normal trends of that account holder. Secondly, the detection systems are custom made for each user profile thus it becomes very difficult for a fraud user to take out any movement without setting off an alarm. Finally, they are able to detect anomalies that are previously not known to the system. Anomaly detection systems search for anomalies instead of attacks.

***Keywords*— Anomaly Detection, training phase, detection phase, types of anomaly, Data Mining, Clustering, Classification.**

I. INTRODUCTION

In a smooth-going business, something that deviates from normal is usually not good. But if it's a glad mishap, despite everything you still need to check it out. Sounds easy, yet with immense amount of data this can turn out challenging, moreover the volume of incoming data is increasing quickly. More and more things are getting on to the Internet, and these information are constantly deciding their working and connected parts. This in turn, makes it difficult to find the anomaly manually. The solution is to develop an automated, self-adaptive anomaly detection system. Ironically, it is easy to build an anomaly detector somewhat as a result of the very thing creating the challenge: huge amounts of data. As in case of anomaly detection, having more data makes it easier to detect an aberration against the background of normal events.

Anomaly detection refers to the issue of finding patterns in data that don't conform to expected behavior. These non-conforming patterns are often termed as anomalies, outliers, exceptions, aberrations, discordant observations, surprises, peculiarities or contaminants in numerous application domains.

Anomaly detection system is modelled on the normal data and may or may not contain the previously known anomalies' dataset, then detects any deviation from the model in the observed data. With the input of available dataset as a training set, and a new test data, the goal of system is to determine whether the test data belongs to a "normal" or to an "anomalous" behavior. These systems do have an advantage of detecting new types of intrusions as deviations from normal profile. However, they got a weakness of high false positive and false negative rates.

Figure 1 shows anomalies in a trivial two-dimensional data-set. It illustrates two normal regions N_1 and N_2 which compasses most of the observations. Points that are considerably far away from these regions, o_1

and o_2, and points in region O_3, are regarded as anomalies. These unexpected behaviors cannot always be categorized as attacks but they can just be the events previously not known. Thus, they may or may not be harmful.

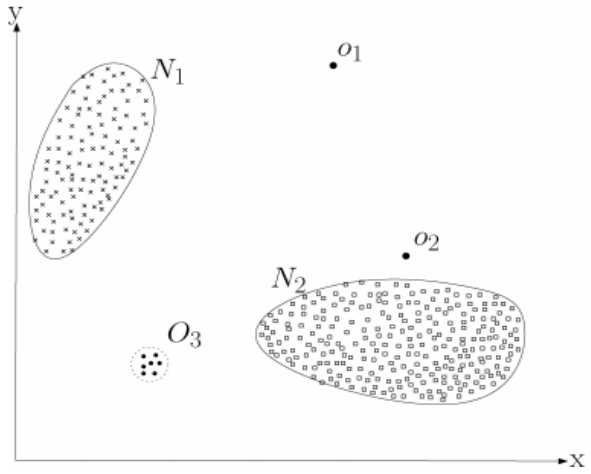

Fig. 1 A simple example of anomalies in a two-dimensional data set.

Anomaly detection expects that whenever there will be an intrusion, there will definitely a deviation from the normal pattern. Based on knowledge requirement for the system, anomaly detection can be categorized into static and dynamic anomaly detection. A static anomaly finder depends on the presumption that a part of the framework being observed does not change. For instance, the OS' software and data to bootstrap a PC never show signs of change. If in case the static part of the framework ever goes astray from its actual form, a mistake has happened or an intruder has changed the static portion of the framework. Dynamic anomaly detection normally works on review records or on monitored network traffic statistics. In order to detect anomalous behavior with respect to normal, thresholds are defined.

II. PROBLEM STATEMENT

To represent the issue of anomaly detection in any framework requires the presence of an underlying terminology of normality. The idea of "normal" is generally given by a formal model that defines relations between the key variables included in the system. Therefore, an event is classified as anomalous in light of the fact that its level of deviation with respect to the profile of characteristic behavior, indicated by the model of normality, is sufficiently high.

Formally, an anomaly detection system A can be expressed as a pair A = (M,D), where M is the model of normal behavior of the system and D is a similarity index which outputs the degree of deviation with regard to the trained model M when a input activity record is given. In this way, the elementary perspective of the system is constituted by two fundamental modules: the modelling subsystem and the detection subsystem. The modelling subsystem comes to the picture during the training phase and scans through the available events in order to train and finally result a model M which reflects the normal behavior the system. This resulted model is utilized by the detection subsystem to check the upcoming new events to get the deviation related to the trained model. These two operations are generally carried out independently. Moreover, take note that systems evolve and, along these lines, the model must be rebuild periodically so that the model can adapt to the new environment.

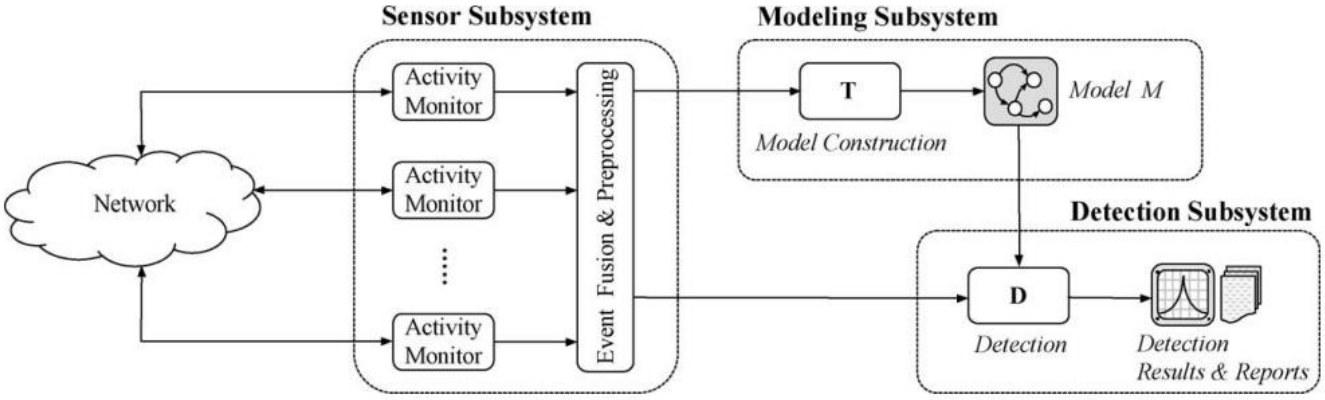

Fig. 2 Architecture of a typical anomaly detection system.

III. ANOMALY DETECTION

A. *Based on Data Labels*

1. Supervised Anomaly Detection
   - Training set consists of labels for both normal data as well as for anomalous data.
   - The minority class, i.e. of anomalous events contains a very small fraction of whole dataset as in rare class mining.
2. Semi-Supervised Anomaly Detection
   -Training set consists of labels for normal data only.
3. Unsupervised Anomaly Detection
   -No pre-defined labelled dataset is provided.
   -It works on the assumption that anomalies are very rare events as compared to normal events.

B. *Types of Anomaly*

1. Point Anomalies
2. Contextual Anomalies
3. Collective Anomalies.

- Point Anomalies:
  An individual instance of data is found to be anomalous with respect to the whole dataset. Figure 1 shows point anomalies o_1, o_2, and O_3.

- Contextual Anomalies:
  An individual instance of data is found to be anomalous within a context. For this, it requires to interpret data as a notion of context. It is also known as conditional anomalies. Figure 2 shows a contextual anomaly in monthly temperature of June in respect to the context of rest of the months.

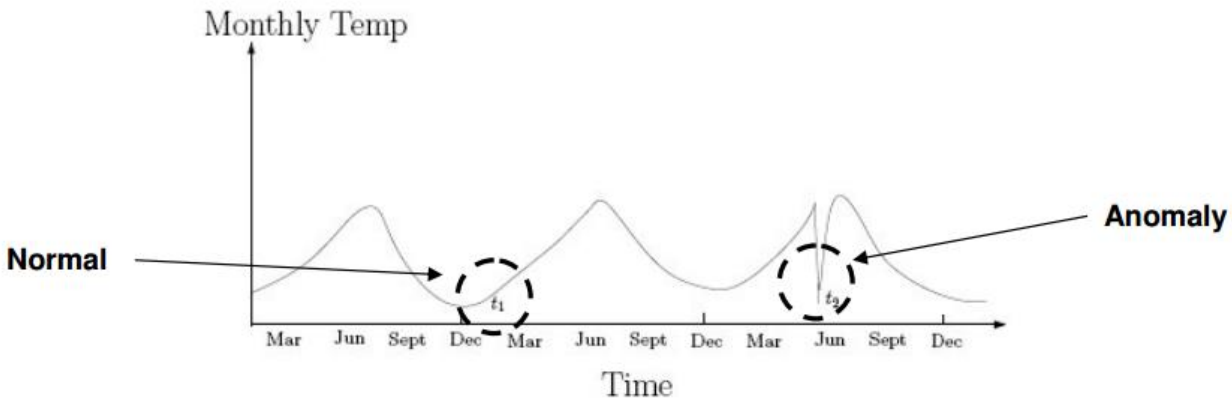

Fig. 3 A contextual anomaly instance in a monthly temperature data set.

- Collective Anomalies:
  A collection of related instances of data are found to be anomalous. As the data are related, it requires relationships among data instances such as sequential data, spatial data or graph data.
  However, the individual instances that make the collective anomaly are not considered as anomalies by themselves. Figure 3 illustrates collective anomalies.

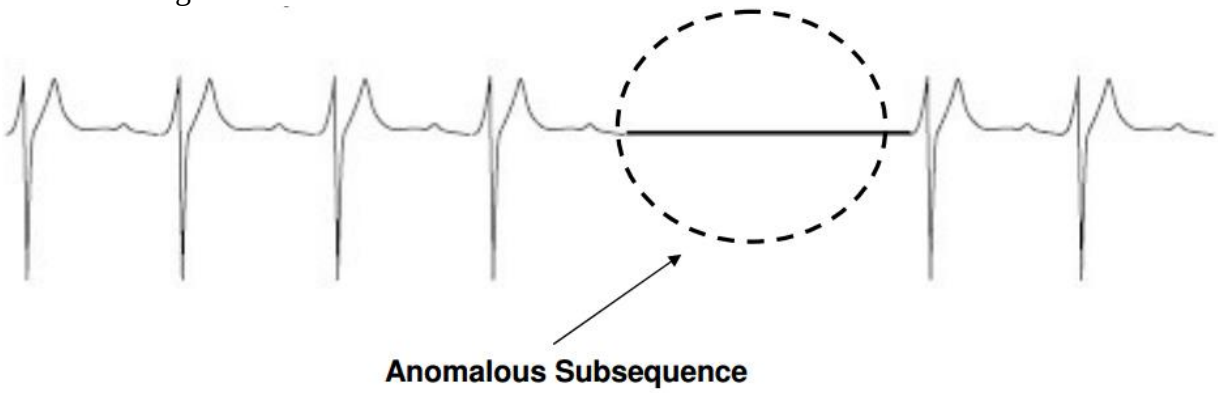

Fig. 4 A collective anomaly showing collection of instances from a data set.

C. Output of Anomaly Detection

- Label:
  - Each instance of a test dataset is assigned with either *normal* label or *anomaly* label.
  - It is of utmost importance mainly in classification-based approaches.
- Score:
  - Each instance of a test dataset is assigned an anomaly score.
  - In terms of this score, it gives ranked output.
  - The deciding parameter in terms of threshold value is needed.

D. Evaluation of Anomaly Detection

The fundamental thought behind anomaly detection is that intrusive activities are the activities that are anomalous. In case of intruder who is unknown to the user's profile trends, probability of detecting this intruder's activity as anomalous is very high. Although in ideal circumstances, all anomalous activities will be same as intrusive activities. In such a case, there will not be any false positives and false negatives.

However, it is not always the case that intrusive activities coincide with anomalous activities. As a result there can be four possibilities as follows, all with a non-zero probability:

- *Intrusive but not anomalous:* This case is referred as *false negatives* as the detection system falsely denied the presence of intrusion. The intrusion detection system fails to detect this intrusion because it was not found anomalous.
- *Not intrusive but anomalous:* This case is referred as *false positives* as the detection system falsely reports the presence of intrusion. Although the activity is not an intrusion, but since it is found as anomalous, the intrusion detection system detected this as intrusion.
- *Not intrusive and not anomalous:* This case is referred as *true negatives* as neither the activity is intrusive nor it is reported as intrusive as it non-anomalous.
- *Intrusive and anomalous:* This case is referred as *true positives* as the activity is intrusive and since it is anomalous too, it is also reported as intrusive.

To minimize false negatives, lower threshold values are set to define anomalies so that more anomalies can be captures. But, this results into larger number of false positives and reduces the efficacy and increases the overhead of investigating each such incident and discard false positive cases. That's why it leads to a trade-off between detection rate (minimize false negative, thus increase the chance of detective anomalies) and false alarm rate (higher false positives, overhead of checking).

IV. Anomaly Detection based on Data-Mining Techniques

Data mining is about analyzing large amounts of data in order to find correlations, patterns, and insights. These techniques make it possible to scan a huge data for characteristic rules. When this is applied to audit data monitoring a network, they can be used to detect anomalies and in turn attacks or intrusions.

Anomaly detection mainly involves the two pillars of data mining techniques: Clustering and Classification.

A. Clustering based Anomaly Detection techniques

Clustering is a task of dividing the data into different clusters of similar objects. This way, each cluster compasses all the objects that are similar to each other but dissimilar to objects of some other cluster. The advantage of using clustering techniques is their ability to learn from audit data provided, while not requiring any prior knowledge about the user profile and explicit description of different available classes to detect anomalies. Clustering and anomaly detection are closely related. From the perspective of a clustering algorithm, outliers are points not located in the clusters of a dataset, and in the viewpoint of anomaly detection, these points tell about intrusions or attacks.

The key assumption behind clustering for anomaly detection is that normal data instances are part of large and dense clusters, whereas anomalies do not belong to a significant cluster.

The general approach:

- Generating finite number of clusters from a given dataset.
- Analyse each data instance with respect to cluster closest to it.
- The data instance can be regarded as anomalous in following scenarios:
    - Data instances that are residuals from clustering, i.e. they do not fit into any of the clusters.
    - Data instances present as small clusters.
    - Data instances present in low density clusters.
    - Data instances that are far from other instances inside the same cluster.

The drawbacks with use of clustering techniques are:

- In case of datasets not possessing a natural clusters or the algorithm does not detect natural clusters, this may fail.
- Computationally expensive as does a lot of computation of distances between all pairs of data points. This may be alleviated to an extent by using indexing techniques.
- In higher dimensional dataset, data is sparse and it is possible to get same distance between two data records.

B. Classification based Anomaly Detection techniques

Classification is a task of identifying the class/category of new input data on the basis of a labelled training dataset that contains observations with their corresponding actual class label. Classification is a supervised learning technique in which a training set of already correctly identified events are available. An algorithm is then constructed to classify or predict this class label for a previously unseen new event. Similarly, in case of anomaly detection data is classified into two class labels as normal and anomalous.

The general approach:

- Use training set to identify all the class attributes and the respective classes.
- Decide on the attributes to use for classification.
- Learn a model using the given training set.
- Finally, use this trained/learned model to classify the class for new unknown data samples.

The techniques based on the required knowledge as part of the training dataset content can be categorized as:

- Supervised classification techniques:
  - As a part of training set, it requires instances with both normal and anomalous class.
  - Classifier is built to distinguish between normal and known anomalies.
- Semi-supervised classification techniques:
  - It requires knowledge of normal class instances only.
  - It deploys classification model that learns from the normal behavior and then detect anomalous instances as any deviations from the normal behavior.

The drawbacks with use of classification techniques are:

- Both supervised and semi-supervised requires labelled data in order to learn. They are not the unsupervised learning techniques.
- In supervised classification, it cannot detect any unknown/new and emerging anomalies.
- In semi-supervised classification, high false alarm rate. It can be the case when a previously unseen yet legitimate data record encountered is declared as anomaly.

V. CHALLENGES

At a theoretical level, an anomaly is characterized as a pattern that does not adjust to expected ordinary trend. A typical anomaly identification approach, therefore, is to define a region corresponding to the normal behavior and then declare any instance that does not belongs to this normal region as anomalous. In any case, a few elements make this apparently basic approach exceptionally difficult:

- To reflect every normal behavior in terms of a normal region is difficult. Moreover, as such there is no precise boundary between normal and anomalous behavior. Thus, it is possible that an anomalous instance close to the boundary may actually be a normal instance, and vice-versa.

- At the point when anomalies are the aftereffect of malicious activities, the intruder often adjust themselves to make the anomalous behavior seems typical, thus making the assignment of defining normal behavior more troublesome.
- In numerous areas normal behavior continues advancing and a present thought of ordinary event won't not be adequately illustrative later on.
- The precise idea of an anomaly is distinctive for diverse application areas. For case, in the pharma domain a little deviation from ordinary (e.g., changes in body temperature) may be an anomaly, while similar deviation in the stock market (e.g., fluctuation in the estimation of a stock) may be considered as ordinary. Therefore applying a system created in one area to another, is most certainly not direct.
- Availability of labelled instances for learning/testing of models utilized by anomaly detection system is generally a noteworthy issue.
- Often the information contains noise that has a tendency to be like the genuine anomaly and henceforth is hard to recognize and eliminate.

VI. CONCLUSIONS

Anomaly detection is an important problem in data mining with diverse range of applications from financial fraud detection, computer system intrusion detection to medical diagnosis. Anomalies corresponds to patterns present in data that deviates from the normal behavior. Detecting these patterns can detect critical information in data. Different models are deployed for different problems depending on the application domain. Although, it seems a simple two phase model: training phase and detection phase, it do have challenges ahead of it in order to replicate normal behavior in the model and to minimize the false rates with the presence of a trade-off between detection rate and false-alarm rate. Point anomalies detection have found their utility in lot of domains but contextual and collective anomalies detection techniques are still in infant stage with scope of exploring several domains.

REFERENCES

[1] An overview of anomaly detection techniques: Known solutions and technological trends by Jung-Min Park ,Animesh Patcha .

[2] Data Mining for Anomaly Detection. Aleksandar Lazarevic, *United Technologies Research Center*; Arindam Banerjee, Varun Chandola, Vipin Kumar, Jaideep Srivastava, *University of Minnesota*.

[3] Survey on Anomaly Detection using Data Mining Techniques. Shikha Agrawal, Jitendra Agrawal

[4] A Survey of Anomaly Detection Methods in Networks. Weiyu Zhang, Qingbo Yang, Yushui Geng

[5] Anomaly detection methods in wired networks: a survey and taxonomy. Juan M. Estevez-Tapiador*, Pedro Garcia-Teodoro, Jesus E. Diaz-Verdejo.

[6] Unsupervised Anomaly Detection in Numerical Datasets. Vineet Joshi MS (CS), Univ. Cincinnati

[7] Anomaly Detection: A Survey. VARUN CHANDOLA, ARINDAM BANERJEE, and VIPIN KUMAR *University of Minnesota*.

[8] Feature deduction and ensemble design of intrusion detection systems. Srilatha Chebrolua, Ajith Abrahama,b, Johnson P. Thomasa.

[9] A Survey of Outlier Detection Methodologies. Victoria J. Hodge and Jim Austin.

[10] A Survey on Efficient Data Mining Techniques for Network Intrusion Detection System. P.Kalarani , Dr.S. Selva Brunda

[11] http://madhukaudantha.blogspot.in/2014/06/anomaly-detection-survey.html

[12] Rare Class Mining: Progress and Prospect Shuli Han, Bo Yuan, Wenhuang Liu

[13] Anomaly Detection is the New Black, by Ted Dunning.