International Journal of Engineering & Technology, 7 (3.24) (2018) 449-453

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Anomaly Detection Techniques Causes and Issues

G Sandhya Madhuri*[1], Dr. M. Usha Rani[2]

[1]Research Scholar, [2]Professor , Dept. of Computer Science, SPMVV, Tirupati, India.
**Corresponding author E-mail:sandhyamadhuri@gmail.com*

Abstract:

Anomaly means something which is not normal. Any data point which deviates or placed in distance from all other normal data points is an anomaly. That is why anomalies are also called as outliers. Anomaly detection is also called as deviation detection because anomalous objects have attribute values that are different from all other normal data objects. In this paper we have discussed about various causes of anomalies, anomaly detection approaches and also issues that are to be taken care during finding out the best technique for anomaly detection.

Keywords:

1. Introduction

Anomalies either detect a problem or a new phenomenon to be investigated. Before we discuss various anomaly detection algorithms, let us see in detail about the following:

i. Applications for which anomalies occur
   a. Fraud Detection
   b. Intrusion Detection
   c. Ecosystem Disturbances
   d. Public Health
   e. Medicine
ii. Causes of anomalies
   a. Data from different classes
   b. Natural variation
   c. Data measurement
   d. Collection errors
iii. Different approaches to identify anomalies
   a. Model based Techniques
   b. Proximity based Techniques
   c. Density based Techniques.

i. Applications in which anomalies occur: In general most events or objects are, by default normal or ordinary. However we have keen awareness about the possibility of unusual or extra ordinary objects.
For Example: extremely dry or rainy season, anomalous values in experimental results. The following are the possibilities where we can see the occurrence of anomalous data.

a. Fraud Detection: The purchasing behaviour of a customer and a person who has stolen a credit card is different. Here the credit card companies identify the buying pattern of a customer. They can easily identify the theft by noticing a change in the typical customer buying pattern.

b. Intrusion Detection: An intrusion in a system or a network is detected by monitoring the unusual behaviour of a network or a system. However it is difficult to detect the behaviour of those that are designed to gather the information secretly.

c. Ecosystem Disturbances: The disturbances in the climatic condition have a direct effect on humans. Hurricanes, Floods are those that disturb the ecosystem. The aim is to predict the likelihood of the events.

d. Public Health: Many times Hospitals, medical clinics are required post the information about various statistical data regarding the public health. Any indication of the problem has to be immediately identified.

e. Medicine: Unusual symptoms of a patient's test results are also considered important situation to attend immediately.

2. Causes of Anomalies

There are various causes for anomalies. They are

a. Data from different classes: An object may be different because it is of a different class. Cases like credit card theft, Intrusion detection, outcome of disease, abnormal test result are good examples of anomalies occurring and identified using class labels.

These kinds of anomalies are of interest and are in the focus of anomaly detection in the field of Data Mining.

b. Natural Variation: In a Normal or Gaussian distribution the probability of a data object decreases rapidly. Such objects are considered as anomalies. These are also called as outliers. The distance between outliers and centre of the distribution is much more than all other data objects. All other objects are near to the centre of the distribution i.e. average object.

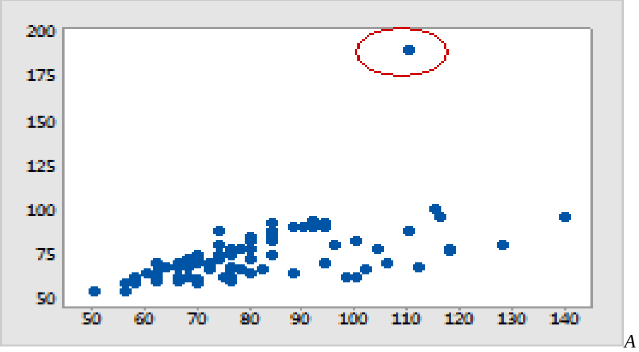
c. Data Measurement and Collection Errors: These kinds of errors occur when we collect erroneous data or if there is any deviation while measuring data. The ultimate goal to eliminate such errors as they reduce the quality of the data. Now the question is, How to reduce such errors. It is possible through a process called *Data Pre-Processing or Data Cleaning.*

ii. Approaches to Anomaly Detection: Anomaly detection is classifies into different approaches based on techniques followed.

a. Model Based Approach: In this kind of approach a model of the data is built. The objects that do not fit very well are considered as anomalies. For example let us assume that a model is built of certain data which is in the form of cluster. Then anomalies are those data objects that do not strongly belong to the cluster built. In a Regression Model the anomalies are far from predicted value.

However, the problem with such approach is when no training data is present to build a model or there is no statistical distribution of data. Therefore, in such cases we require techniques that do not require a data model to be built.

b. Proximity Based Techniques: This approach is based on the proximities. Consider a 2D or 3D scatter plot all the data objects are in one proximity. But Anomalous objects are away from them. See the graph given below.



A

2D scatter plot where in anomalies are out of the proximity

So for such type of anomalies distance based outlier detection techniques are used. These types of anomalies can be visually detected.

Density Based Techniques: Density of objects is easy to compute especially if a proximity measure between objects is available. Low density objects are those that relatively distant from neighbours. We call such objects as Anomalies.

The Use of Class Labels:

Basic Approaches to Anomaly detection: There are three basic approaches anomaly detection.

1. Supervised Anomaly Detection: In this supervised learning there must a training set for both data objects and expected anomalous objects. We have to observe that there can be more than one anomalous class.

2. Unsupervised Anomaly Detection: For situations where class labels are not available. We can give a score for each object that shows the degree to which the instance is anomalous. We also can observe that if there are many anomalies present which are similar to each other, then we can group them as normal group or the outliers score is low. So, we can say that for unsupervised anomaly detection to be successful it is must that anomalies are distinct.

3. Semi Supervised Anomaly Detection: Sometimes when there is training data with labelled normal objects and score given, but has no anomalous objects, then we can implement the semi supervised anomaly detection to find the anomalies. We use the normal objects to find the anomalies. But, the difficulty is sometimes it is not easy to find that representative set of normal objects using which we have to find out anomalies.

3. Anomaly Detection Methods

A. Types of Anomalies.

Anomalies can be widely classified as:

➢ Point anomalies**:** A single instance of knowledge is abnormal if it is too distant from the remainder. Business use case: DetectingMasterCard fraud based on "amount spent."

➢ Contextual anomalies: The abnormality is based on the specific context. We can find this kind of anomaly in time-series information. Business use case: Amount spent (100$) on petrol on a daily basis throughout the working days is normal, but will be found odd when spent on a vacation.

➢ Collective anomalies: A set of knowledge instances puttogether helps in detecting anomalies. Ex:Somebody is attempting to repeat information type a far off machine to an area host unexpectedly, An anomaly that will be marked as a possible cyber-attack

Anomaly detection is comparable to — however not entirely identical as — Noise Removal and Novelty Detection.

The some of the categories of approaches are given as follows:

- The Statistical Approach
- The Rule-Based Approach
- The Pattern-Matching Approach

B. Statistical Approach

Let's review the algorithm types from the statistical approaches with perspective of appliance to finding various types of outliers.

1) STL decomposition

This type STL (Seasonal-Trend decomposition procedure based on Loess) depends on the seasonal time series. In this technique the entire time series signal is split into three parts named as seasonal, trend and residue.

Once the deviation of residue is analysed there will be an introduction of some kind of threshold applied on it which gives us an anomaly detection algorithm. We should use median absolute deviation to obtain more robust way of detecting anomalies.

Ex: Twitter outlier detection library uses Generalised extreme student deviation test to check if the residual point is an anomaly.

Advantages of this technique are its simplicity and robustness. In all the different situations this technique will be able to interpret the anomalies. It is used mostly for additive outliers.

However, this type does not work well with the situations where there are changes occurring radically

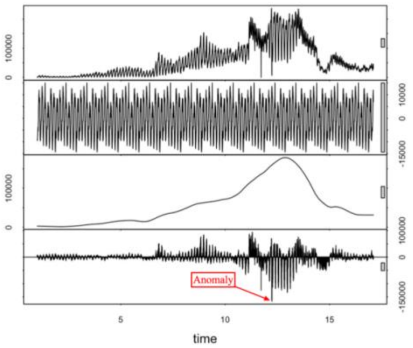


Figure 1: An example of the STL technique with original time series, seasonal, trend and residue parts graphically represented from top to bottom.

3) Classification and Regression Trees

‘Classification and regression trees’is considered as the best technique to detect outliers.

The classification and regression tree is constructed by splitting the data repetitively into unique data points by a simple rule. For each split the data is categorised into mutually distinguishable groups. Each group is homogeneous. After this division, splitting is applied on the groups separately. The idea is to divide the data into homogeneous groups by keeping the tree as small as possible.

The advantages of this implementation are that this technique is not bound to any sense of structure of the signal and you can introduce many feature parameters to perform the learning and get sophisticated models.

The disadvantage is the increasing number of features may start to influence your computational performance rapidly.

2) ARIMA

This method (ARIMA- Auto Regressive Integrated Moving Average) is a simple method. Even though ARIMA model design is simple, it is very robust and good method to forecast signals and find abnormalities in it.

The method used in it is that several points are taken from the past and used to generate a forecast of the next point by adding a random variable which is usually called as white noise. In such way the forecasted points will generate the future points.

The tough part in application of this method is that we should select the number of differences, number of auto - regressions and forecast the error regressions.

The important point to be remembered in ARIMA method is that each time we work with a new signal a new ARIMA model has to be built.

A significant constraint of this method is that the signal should not be dependent on time that means it should be static after differencing.

The anomaly detection is done by making an adjusted model of s signal by using outlier points and by comparing whether it is a better fit than the original model. This process is done by using t-stat method.

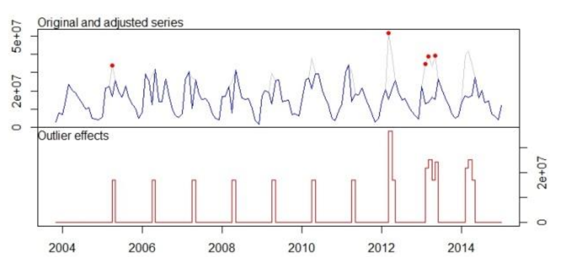


Figure 2: Two time series built using original ARIMA model and adjusted for outliers ARIMA model.

This method is suitable for detecting all the types of anomalies in every case where you can build an ARIMA model for our signal.

4) Exponential Smoothing

The Exponential smoothing methods of techniques are very identical to ARIMA approach. The basic exponential model is equivalent to ARIMA (0,1,1).

Using the exponential smoothing the anomaly detection perspective is as defined by Holt-Winters seasonal method where a seasonal period which might be equivalent to a week, month, year etc. is defined.

In this approach we should track various seasonal periods. For example, when we have week and year dependencies, we should select only one such as a week in this case as it is the shortest one. This is the drawback of this method because it will affect the forecasting horizon a lot.

The same statistical tests are used for detecting an outlier that is used in STLs.

C. Rule Based Approaches

Rule-based approach is a different way to detect anomalies. In this method a database is maintained which contains the set of rules which majorly govern the faulty system or an anomalous behaviour. This database is used to check whether an anomaly has occurred or not. The anomaly or fault is identified by observing a sequence of indications or symptoms that are predefined by the rules. Rule-based methods majorly depend on the human expertise and are not accommodative to new and changing environments.

There is a new method which is used as an extension to rule – based that is called as Case – based reasoning. In this method a history of all the faults occurred by a system all used to make decisions. It also can build new rules. This approach is adaptive for evolving environments. Case based approach fails in efficiency of computation time and complexity as relies heavily on previous information.

Rule-based packet classification is a powerful method for identifying anomalies in traffic in a network. This approach is mainly useful in the area of network security.

The two important applications in the network security management are: Detection of network attack traffic and detection of non-attack traffic. Most unwanted traffic is identified by rules that match known signatures. Rules might match on packet header or payload or both.

Signature based detection , such as Snort are widely implemented by organisations for network security but are limited to scaling factors such as concurrent implementations of rules at the gateways of the networks to detect the traffic anomalies.

An intrusion detection system which should inspect every packet is ideal but is almost impossible.

D. Pattern Matching Or Profiling

This approach uses online learning to build profiles or patterns of users and considers them for normal behaviour, and abnormalities from them are considered anomalies. These methods do not work well for changing behaviours over time.

Usually in organisations, examining the data for consistency comes once the info has been fed and its comprehensiveness and precision has been observed. The values of a given element will come in totally different shapes and sizes. This is regularly very accurate for fields requiring human input, wherever the values are entered in line with the notions of the user. Let us assume a column representing the pin code field is intelligible, it can be aforesaid that for each entry the values represent should be valid pin code since they match the expected format, length, and knowledge sort (numeric), therefore meeting the expectation of the system. Erroneously representing information in incorrect format results in inaccurate analytics, and in the massive data context, its sheer volume will intensify this inaccuracy.

There are various algorithms in pattern matching are exact – pattern matching, Knuth – Morris – Pratt, RE pattern matching, GREP.

1) Exact pattern Matching

In the exact pattern match we find the first match for the pattern of certain length in a text stream. This exact pattern match method is used in parsers, spam filters, digital libraries, screen scrappers, word processors web search engines etc.

The exact pattern matching has disadvantages like it can be slow when there is a repetitive pattern, but usually repetitive pattern is rare.

2) Knuth–Morris – Pratt (KMP) PatternMatching

Another type of pattern matching algorithm is Knuth – Morris – Pratt (KMP) which addresses almost all the draw backs of exact pattern making algorithm

In this technique we use DFA (Deterministic Finite Automata) from the pattern at the outset. Simulation of DFA is done with the text as input.

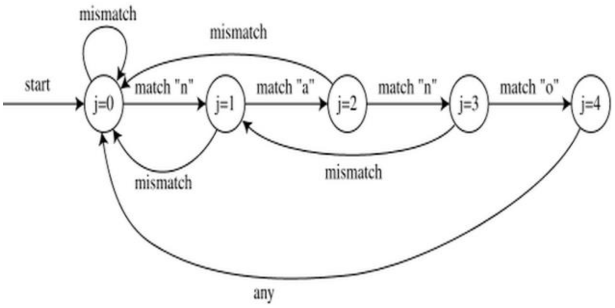


Figure 3: A DFA showing the pattern text match or mismatch

As seen in the above figure there is one stage for each character of the pattern. Whenever there is a mismatch it moves to the previous states.

Implementing KMP model shows that linear-time pattern matching is possible.

3) Regular Expression (RE) Pattern Matching

Regular Expression pattern matching (RE) method, unlike the exact pattern matching searches for occurrences of multiple patterns in a text file. RE pattern matching technique is applied to scan virus signatures, search for information in search engines, access information in digital libraries, search and replace in word processors, validation of data entry fields like name, age, DOB etc.

4) GREP Pattern Matching

Another technique in pattern matching is GREP implementation. The overview of GREP implementation is very much similar to KMP technique. In GREP we build DFA from RE unlike KMP where we build DFA from the pattern text. Simulate of DFA with text as input. RE is the concise way of describing the set of strings whereas DFA is a machine to recognise whether a given string is in the set or not.

The disadvantage of this technique is that the DFA that has to be built with RE will be exponentially large and so difficult to build. Therefore the revised GREP technique has NFA (Non Deterministic Finite Automata) with RE. However linear time guarantee is not considered here in this revised technique.

D. Other Techniques in Anomaly Detection

1) Model- Based Approach

The corporate setting of these modern days, have unusual users interacting with software who are potentially dangerous and can cost organizations millions of dollars. Anomalous behaviour can be caused by the presence of compromised user accounts, malicious users, or by novice users.

There are two general approaches in model based anomaly detection:

a. *Graph-based anomaly detection*
b. *Activity-based anomaly detection*

Graph-based approaches analyse company's structure like ego-networks of nodes, communities, and sub graphs. Whereas, activity-based methods focus on user behaviour like number of logins, file access.

2) Graph-based anomaly detection algorithms

To solve the abnormality detection problem, there were a lot of techniques that were developed since decades, particularly for identifying outliers and anomalies in amorphous collections of multi-dimensional data points. Also, data objects cannot always be called as points lying in a multi-dimensional space independently. On the other hand the objects might be inter – dependent or inter – linked, which should be considered while detecting anomalies; see the figure below. If we consider data instances in some of the disciplines like biology, physics, social sciences and information systems, they are all integrally inter – related. Graphs give a wonderful platform to capture these distantly – ranged correlations among mutually dependent data.

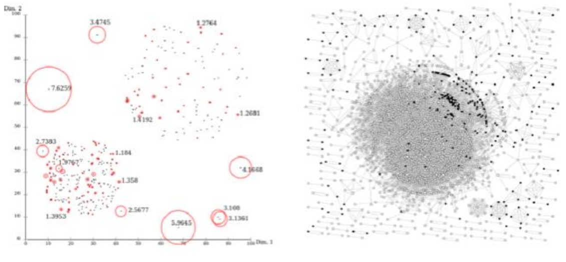

(a) Clouds of points (multi-dimensional) (b) Inter-linked objects (network)

Figure 4: Point based – outlier detection versus graph based anomaly detection.

For example, consider the reviewers data of some products of an organisation or a supermarket/online store. To detect the fraudulent reviewers, we can observe what ratings were by him/her to which product. Also, by observing how other reviewers rated the same product. This shows how trustworthy were their ratings. Hence it can be detected that among these distantly long correlated real – world data sets, anomaly detection is dissimilar than that of detecting outliers in multi – dimensional space.

The three vital reasons why graph based anomaly detection is important are

1. Data dependency: The data objects are either dependent on each other or related to one another.

2. Explicit representation: Using graph a natural representation of data with links and nodes between the related data objects is effectively represented.
3. Problem domain is relational: Interdependency of the anomalies is also one of the reasons to opt for graph based anomaly detection. For example: A computer failure due to malfunctioning can affect all the other machines which are dependent on it.

Challenges: In graph based anomaly detection it is difficult to span the large datasets which are streaming at a pace. Also the data is heterogeneous and complex. Incorporating the complex related data in a graphical representation where nodes and edges can be typed and a list of attributes associated with them scales up to a large graph. Also updating the changes of the graph overtime are essential.

D. Activity Based Approach for Anomaly Detection

In this kind of approach the behaviour of the user is observed. This type of analysis is similar to the dynamic analysis where a sample user or a machine activity is observed in a normal environment and the traces of the complete behaviour is noted. Keeping this as a benchmark the activity based detection is carried out.

Challenges: Usually in a network scenario gateways and access points which monitor the activity pattern of the user nodes it is difficult to maintain the network function normally for normal nodes when other nodes are not routed correctly is a big challenge.

E.Similarity based approach for anomaly detection

The main idea in this similarity based approach is that a normal point with similar characteristics of that of the neighbours is considered as non-anomalous. So a point is considered as not an outlier if it has closer proximity with the neighbours and neighbours are several.

The sum of similarity is calculated of the point to each other. This is considered as the degree of proximity which is taken as an opposite of the outlier factor that characterises the data points.

The key idea is that the normal point has several neighbours along with it which has similar characteristics. That means the point has a high degree of proximity when its neighbours are many.

4. Issues

There are various issues that are to be addressed before when dealing with anomalies.

How many Attributes are to be used to define anomalies: Consider a scenario when there is one attribute for an object, for some data values are anomalous but other values of other attributes are normal. Whether we have to decide based on those values of the attributes or not. For example, consider people who are 6 feet tall and people who weigh 90kgs. But it is uncommon to have people who are 6ft tall and 90kgs weight. So, there must be a definition of anomaly specifying how the multiple attributes are used to identify whether an object is an anomaly or not.

*Global vs Local:*Any object may seem unusual or anomalous with respect to all other objects, but may not be anomalous to its local neighbourhood. For example: A person with height 6 ½ft is abnormal when compared to normal people. However, it is normal when compared to the basketball players.

*Degree to which a point is anomaly:*To decide whether an object is anomalous or not is based on the divisive decision in a binary format in some techniques. But it is important to know that there are some extreme anomalies and very less in degree. This is possible by giving a score to each anomaly based on the degree of it being anomalous. This assessment is called as anomaly or outlier score.

Masking and Swamping: In some anomaly detecting techniques anomalies are identified one at a time. In such techniques there is a chance of missing out anomalies because of Masking. Masking means presence several anomalies masks all other objects. Whereas identifying multiple anomalies at once can involvement of Swamping. Swamping is where normal objects are considered as anomalies.

Precision, recall, false positive rate: If class labels are available to identify anomalies and normal data, then it is usually normal that the class of anomalies is smaller than that of normal class. In such cases measures like precision, recall and false positive rate are more important than the accuracy.

Efficiency: There are some significant differences between anomaly detection techniques. It is very important to choose the right technique. Based on the time complexities choosing an anomaly detection technique is important as finding an anomaly is not the only criteria it is also important to reduce the cost too.

5. Conclusion

In this paper we have discussed about various causes of anomalies and techniques to identify the anomalies. Whereas we leave the roadmap by discussing the issues that arise during the detection of anomalies as a further scope of identifying techniques that overcome the issues.

References

[1] D. Polla, "A Framework for Cooperative Intrusion Detection", Proc. 21st Nat'l Information Systems Security Conf., pp. 361-373, 1998.

[2] D. Zerkle, "A Data-Mining Analysis of RTID Alarms", Recent Advances in Intrusion Detection, 1999.

[3] G. Grinstein, "Workshop on Information Exploration Shootout Project and Benchmark Data Sets: Evaluating How Visualization Does in Analyzing Real-World Data Analysis Problems", Proc. IEEE Visualization 97 Conf., pp. 511-513, 1997.

[4] K. Cox, S. Eick, T. He, "3D Geographic Network Displays", ACM Sigmod Record, vol. 25, no. 4, pp. 50, Dec. 1996.

[5] E.E. Koutsofios, "Visualizing Large-Scale Telecommunication Networks and Services", Proc. IEEE Visualization 97 Conf., pp. 457-461, 1997.

[6] S.G. Eick, G.J. Wills, "Navigating Large Networks with Hierarchies", Visualization 93 Conf. Proc., pp. 204-210, 1993.

[7] R. Becker, S. Eick, A. Wilks, "Visualizing Network Data", Readings in Information Visualization: Using Vision To Think, pp. 215-227, 1999.

[8] T. Bray, "Measuring the Web", Readings in Information Visualization: Using Vision To Think, pp. 469-492, 1999

[9] Davidson, "What Your Database Hides Away", New Scientist no. 1855, pp. 28-31, Jan. 1993.

[10] R.F. Erbacher, D. Frincke, "Visualization in Detection of Intrusions and Misuse in Large-Scale Networks", Proc. Int'l Conf. Information Visualization 2000, pp. 294-299, 2000.

[11] Introduction to “Data Mining” book by Pang-Ning Tan, Michigan State University, Michael Steinbach, University of Minnesota, Vipin Kumar, University of Minnesota and Army High Performance Computing Research Center.