IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

OUTLIERS IN DATA MINNG

A closer look on outliers to understand its concepts

[1]Miss. Reshma Joseph

[1]Assistant Professor
[1]PG Department of Computer Science
[1]JPM Arts and Science College, Labbakkada,India

Abstract: Data Mining refers to the mining of information from a large amount of data. It is the process of discovering interesting patterns and knowledge from massive data. Clustering analysis is used to find these interesting patterns or to group similar and dissimilar data. This dissimilar data can be called outliers. Outlier detection refers to the problem of finding out the patterns in the massive datasets that do not show the accordance with the generalized expected behaviour. Outlier detection and analysis is sometimes known as outlier mining which means the mining of outliers.

Index Terms **- Outlier Detection, Causes of outliers, Global outliers, Contextual outliers, Collective outliers**

I. INTRODUCTION

Nowadays we are dealing with a large amount of data. We have to categorise similar and dissimilar data from massive data. Outlier Detection is one of the major topics in Data Mining; detection of outliers from a group of patterns is a popular issue in the field of data mining. An outlier is that pattern that is unlike concerning all the remaining patterns in the data set. It is a very important task in various application domains. Earlier outliers considered as noisy data has now become severe difficulty which has been discovered in various domains of research. Noise is anything that is not the "true" signal. It may have values close to your true signal. An outlier is something much different from the other values. The vast majority of time outliers are noise but sometimes a data point that is true signal can be an outlier. The discovery of outlier is useful in the detection of unpredicted and unidentified data, in certain areas like fraud detection of credit cards, calling cards, discovering computer intrusion and criminal behaviours etc. Several surveys, research and review articles cover outlier detection techniques in great details.

II. MOST COMMON CAUSES OF OUTLIERS IN A DATA SET

- Data entry errors (human errors)
- Measurement errors (instrument errors)
- Experimental errors (data extraction or experiment planning/executing errors)
- Intentional (dummy outliers made to test detection methods)
- Data processing errors (data manipulation or data set unintended mutations)
- Sampling errors (extracting or mixing data from wrong or various sources)
- Natural (not an error, novelties in data)

III. TYPES OF OUTLIERS

Outliers can be classified into three categories;

3.1 Global Outliers

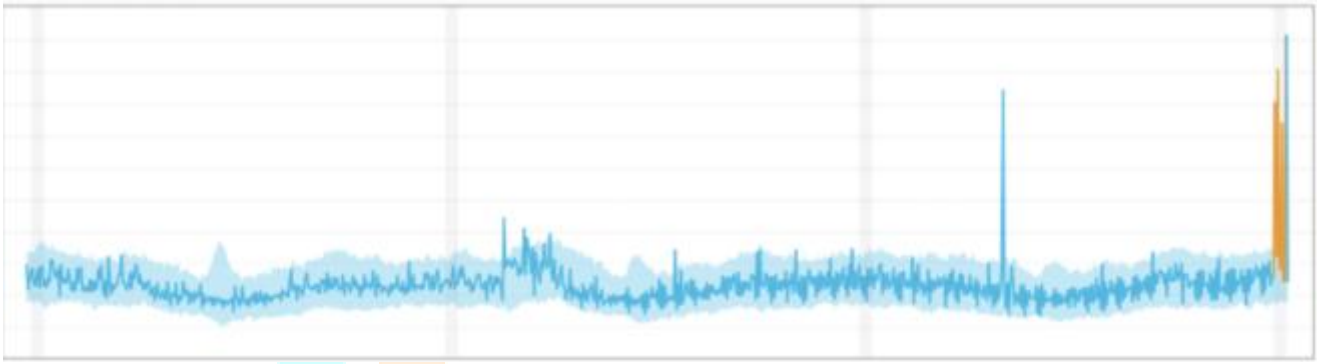
3.2 Contextual Outliers

3.3 Collective Outliers

3.1 Global Outliers

A data point is considered a global outlier if its value is far away from the entire set of data in which it is found. It deviates significantly from the rest of the data set. It is also called *point anomalies* and it is the simplest type of outlier. To find global outliers a major issue is its difficulty to find an appropriate measurement of deviation with respect to the application in question. Global outlier detection is useful in many applications like intrusion detection in computer network, trading transaction auditing system etc.

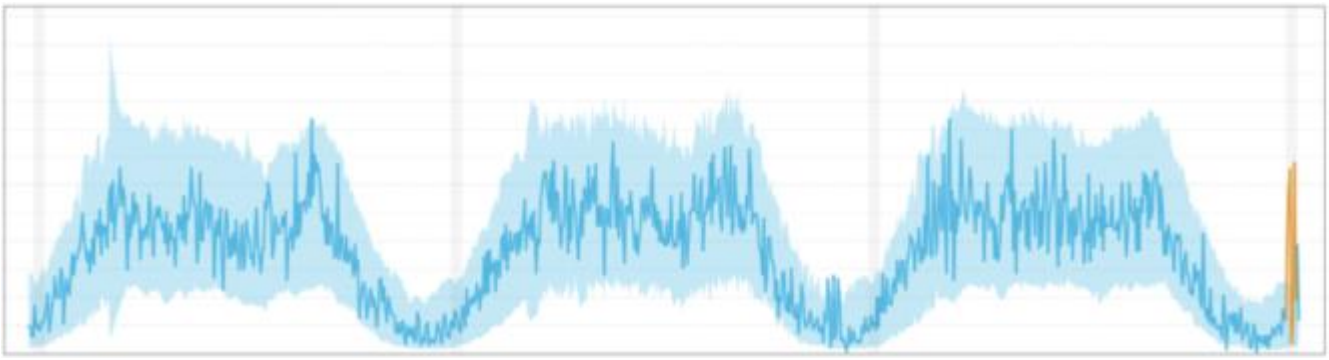
Global Anomaly:



3.2 Contextual Outliers

In a given data set, a data object is a contextual outlier if it deviates significantly concerning a specific context of the object. Contextual outliers are also known as conditional outliers because they are conditional on the selected context. Therefore in contextual outlier detection, a context or condition has to be specified as part of the problem definition. Global outlier detection can be regarded as a special case of contextual outlier detection where the set of contextual attributes is empty. In other words, global outlier detection uses the whole set of data as the context. Contextual outlier analysis provides flexibility to users to examine in different contexts, which can be highly desirable in many applications.

Values are not outside the normal global range, but are abnormal compared to the seasonal pattern.

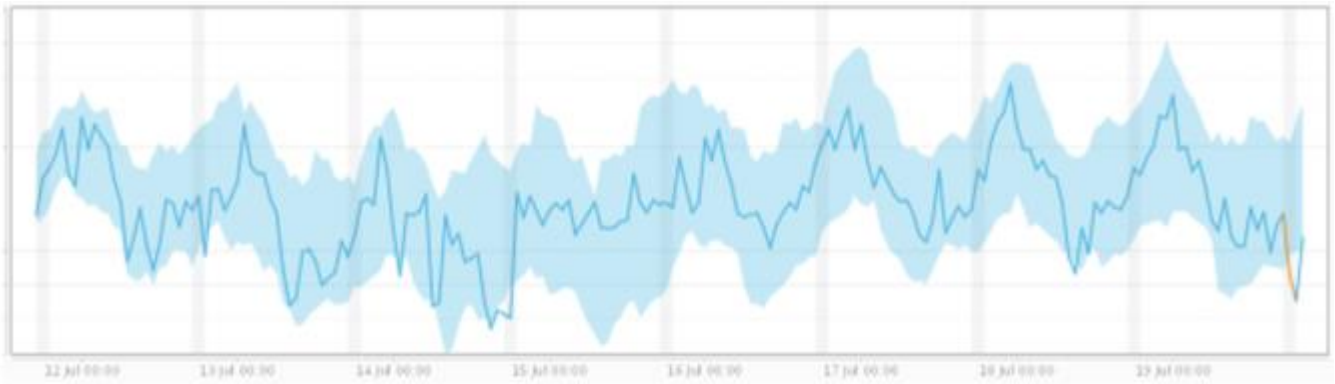


3.3 Collective Outliers

Given a data set, a subset of data objects forms a collective outlier if the objects as a whole deviate significantly from the entire data set. Individual data objects may not be outliers but as a whole they are outliers. Unlike global and contextual outlier detection, in collective detection, we have to consider the behaviour of the individual object as well as the group of objects. So we need the background knowledge to model the relationship among the objects to find groups o outliers.

In the example, two-time series that were discovered to be related to each other are combined into a single anomaly. For each time series, the individual behaviour does not deviate significantly from the normal range, but the combined anomaly indicated a

bigger issue.



IV. APPLICATIONS

- Quality control applications
- Financial applications
- Web log analytics
- Intrusion detection applications
- Medical applications
- Text and social media applications
- Earth science applications
- Detection of fraud values in credit cards, debit cards, use of mobile phones or insurance claims, financial transactions etc.
- Functioning of Network- Nursing and checking of computer networks to find blockages
- Environmental monitoring example Cyclone , Tsunami , Floods, Drought, Fire, Typhoon etc
- Public health- for detecting unusual symptoms or abnormal test results to identify instrumentation / clerical errors or real health problems.
- Medical conditioning monitoring example heart rate monitors
- Pharmaceutical research- recognizing novel molecular structure.
- Localization and tracking
- Logistics and transportations
- Noticing unforeseen entries in databases
- Detecting wrongly labelled data in a training data module.

V. CONCLUSION

Data Mining is the process of extracting use of information and patterns from a large amount of data. It is very useful in many applications in a variety of ways. But there are many issues in data mining that need study and research. And outliers are one of them. They may be dissimilar from complete data sets or maybe difficult from its neighbourhood only. The work presents a review of the outlier and its applications. The study comprises an analysis of various outliers. Outlier information is very useful when data is compared with the original data. According to the type of application and data sets, it is to be decided by the programmer which outlier detection technique is the most suitable and beneficial for the application.

VI. ACKNOWLEDGEMENT

Gratitude is a feeling which is more than words. It gives me immense of pleasure and satisfaction in submitting this paper **OUTLIERS IN DATA MINING .**In under taking this paper publication I need the direction, Assistance, cooperation of various Individuals. I would like to express my sincere gratitude to my college management, parents and my friends for their support.

REFERENCES

[1] http://www.wikipedia.com
[2] www.sciencepubco.com
[3] Jiawei Han & Micheline Kamber, Data Mining, concepts and techniques.. 3rd
[4]https://www.researchgate.net/publication/269802647_Comparative_Analysis_of_Outlier_Detection_Techniques
[5]https://ieeexplore.ieee.org/document/7508146
[6]https://www.mathworks.com/matlabcentral/answers/54798-what-is-the-difference-between-noise-and-outlier#:~:text=Noise%20is%20anything%20that%20is,signal%20can%20be%20an%20outlier.
[7]https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561
[8] https://towardsdatascience.com/outliers-analysis-a-quick-guide-to-the-different-types-of-outliers-e41de37e6bf6
[9]https://www.anodot.com/blog/quick-guide-different-types-outliers/
[10] https://link.springer.com/chapter/10.1007/978-3-319-47578-3_13
[11] https://www.ijert.org/research/outlier-detection-for-different-applications-review-IJERTV2IS3508.pdf