Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

A Comprehensive Study of Outliers

Dr. Namita Srivastava
HOD, Department of Statistics, St. Johns's college
Dr. B.R. Ambedkar University
Agra, India

Ruchi Trivedi
Research Scholar, Department of Statistics
Dr. B.R. Ambedkar University
Agra, India

***Abstract*- - The research is based on the study of outliers, which are defined as a data point that deviates from the rest of the data. Detection of Anomaly is very emergent issue in many multiple fields like machine learning, deep learning, image processing and statistics which have been researched in the various application area from different domain such as healthcare, agriculture, banking and fraud detection. We shall briefly address outliers with data mining and statistics techniques, application and methods in this paper. Another part of paper is consisting of pros and cons of distinct outlier algorithms and also provides a concise familiarity of discrete types of tools for detecting outliers and reviewed about different databases which we can easily use for outlier analysis. Outlier detection is very crucial aspect in area of data mining because it creates very adverse influence on data set. In today era mostly research becomes data mining oriented, so familiarity with data mining is also essential. The objective of this research is to furnish an understanding of outliers and its methods, application areas to detect anomalies for research.**

Keywords:- Outliers, statistics, image-processing, anomaly detection, machine-learning.

I. INTRODUCTION

Outlier detection are those observations that do not coincide towards the expected behavior. Sometimes it is also known as Anomaly, Discordancy, unusual events and many more. Study of Outlier is totally about detecting the deviant objects, unusual events and exceptionally different from the other objects. Outliers are data points that cannot be grouped into any form of clusters because they differ from the other items in some way. Inspite of all this; there is no standard techniques for detecting Anomalies so it is difficult to describe characteristics on which outliers can be selected. Detecting outliers yields valuable actionable information in a variety of applications like fraud detection [1], [2], intrusion detection in cybersecurity [3], and health diagnosis [4]. Data Mining is highly concerned with outlier detection is an important part for better research. Outliers Detecting from a collection of datasets is a well-known Data Mining process. Outliers major goal is to recover objects from enormous datasets that behaves differently than the rest of the data. This paper includes all important aspects about outlier detection, its methods applications and role of data mining along with tools which helps to detect outliers.

II. DATA MINING

In today's era, there is a tremendous growth in data. Currently, the increment in data is very rapidly, for this reason data mining becomes essential for the analysis of data. Manual data analysis and data retrieval is highly time consuming. Data mining is an emergent and inspiring field for research. The practice of obtaining hidden and useful information from massive datasets is known as Data Mining. Achieving knowledge and making decision is a crucial aspect of data mining. Rise in data dimensionality is considered as a major challenge in the process of Data Mining. Dimensionality refers to the number of features or variables. Due to rising bulk of data, several issues such as data redundancy and missing data develops. Data mining involves various techniques and methods to extract knowledge. Data Mining is multidisciplinary area of computer science which can be with artificial intelligence and statistics.

III. OUTLIERS

Outlier Detection is a crucial activity in many emerging application areas such as intrusion detection, fraud detection and health fraud detection. Outliers are different patterns or observations in a dataset that avert from the other observations. Detection of outliers is an outlook to determine the patterns from given dataset whose nature is not normal. These unexpected natures are detected as the Outliers. Outliers are also known as Anomaly, Discordancy, etc. Outlier Detection is an essential part of today's research because outliers help to detect the unusual behavior is very fruitful for our research. Because of its widespread use in variety of applications, Outlier's detection becomes a crucial and vast study branch in data mining. After recognizing outliers, you can gain useful information that can help you make better data decisions.

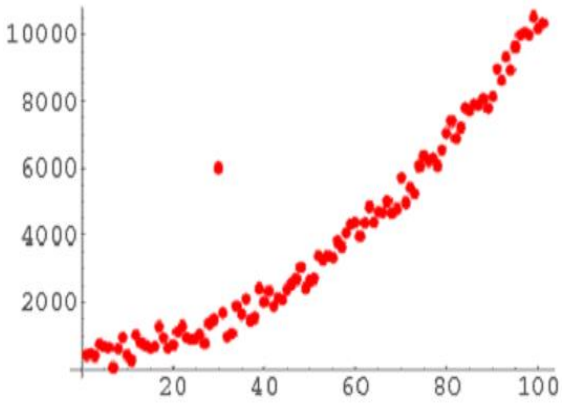

Fig 1. Example Of Outlier

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

1. CAUSES OF OUTLIERS:

Outlier arises due to malicious activity (frauds in criminal activity), human error (mistake in data entry), sampling error (data collection from wrong source), instrumental fault (faults in machine or wire)

Reasons for Handling Outliers: outliers generally recorded as an error; but sometimes it gives fruitful results for our research. Outliers have very remarkable impact on results. So, it is necessary to study outliers.

2. TYPES OF OUTLIERS:

A data point is known as a *GLOBAL OUTLIER* if any data point is distinct from the complete dataset. It is one of the simplest forms of outlier. For example: credit card fraud detection. When a data datapoint is anomalous with respect to some context(condition) then data point is said to be *CONTEXTUAL OUTLIER*. For example: weight of adult is 60kg may be normal while weight of a child is 60kg is an outlier. *COLLECTIVE OUTLIER* when a group of data point is anomalous from rest of the entire dataset. For example: Human Electrocardiogram output.

3. APPLICATION OF OUTLIERS*:*

Outlier detection have immense role in various areas. It is not possible to cover all of the application area, so few of them are discussed according to the recent research and need.

- *Intrusion Detection*: It identifies the suspicious activities that may indicates a system attack or unauthorized access and extracting some useful information.
- *Credit Card Fraud Detection*: Fraud detection becomes a great threat for the institution and banks using a credit card transaction. For example, if a card is lost or stolen; then the behavior of purchasing limit suddenly fluctuates, and this abnormal fluctuation rise to outliers.
- *Fake News and Information:* Although social media has become a venue for disseminating news, it has increasingly become difficult to distinguish between fake news and authentic news. Fake news can be spotted as an Outlier when using credible sources.
- *Medical and Public Health Outlier Detection*: Many methods of outlier detection are used in medical diagnosis which helps to detect critical diseases at early stage and it can be prevent to become a severe disease.
- *Image Detection*: Outlier detection is to detect the abnormal patterns of image that includes color, texture and brightness.
- *Data-Stream and time-series Detection:* Detecting abnormal pattern in data-stream and in time-series is essential part because these abnormal patterns will impact the fast computation and reduce the efficiency of results.

IV. CHARACTERISTICS OF OUTLIERS

Depends on multiple features like type of data, size of data is taken into considerations for identifying outliers. Degree for contemplate outlier scalar (binary) means that data objects is actually an outlier or not. Another is “outlierness” which refers to fact that a data point is regarded an outlier in comparison to other data points. It can be considered as outlier score. Determine outliers on basis of dimensions a univariate data it is classified is an outlier on the basis of single attribute in data. In contrast, multivariate data it is classified an outlier on the basis of numerous attributes.

Different types of approaches for outlier detection:

A plenty of research have been already developed for detecting outliers. Some of existing outlier detection method has been discussed.

1. STATISTICAL BASED APPROACH

Statistical approach is one of the prior techniques to detect outliers. (Tukey,1977) proposed a graphical tool; “BOX PLOT” to visualize data [5]. It is a graphical representation of quartiles and interquartiles of distribution about data. A data point is said to be outliers if it lies outside the whiskers of box plot. Depth-Based Outlier detection identification is a statistical method that is one of the versions of statistical outlier detection. Outlier data points that are based on depth that are represented by n-d space with assigned depth, data points which have smaller depth is considered as an outlier. Another renowned statistical approach is GRUBB’S statistics; it works on univariate data; Z value is calculated as the difference between mean value and dubious value is divided by standard deviation to get the Z value. It is also known as extreme studentized deviate test or maximum normalized residual test. Outlier detection using regression models is one of common method for detecting outliers. This model works in two phase principle. Firstly, there is a training phase in which regression model is construct to fit the data and this model might be a linear or non-linear in nature. In Next there is a test stage, in which regression model a test is perform by comparing the data instance to the model. When a significant difference exists between the actual value and expected value derived by regression model, a data point is labelled as an outlier. Dengel [6] suggest a Histogram-Based outlier detection method that calculate outlier score by using static and dynamic bin width histograms. Hido *et al.* [7] proposed a Density Ratio Estimation as a new statistical approach for inlier-based outlier detection.

Advantages and Disadvantages of Statistical Outlier detection approach:

Statistical outlier methods are mathematically admissible and it gives best efficient result if model is probabilistic. For quantitative real valued data or some quantitative ordinal data sets, statistical models are often prone to it. Most of statistical models are apply on univariate space but it does not applicable to multivariate space. While facing with this problems of increasing dimensionality, statistical model

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

adopts different types of technique resulting it increased time complexity, computational cost and mis-presentation of the distribution of data.

2. DENSITY BASED APPROACH

It is one of the earliest approaches to detect outliers. Density based approaches assumed that non-outliers are assumed to be in high density, while outliers are located in low density. Local outlier factor (LOF) it is first basic approach related to density-based clustering approach. After introduction of LOF different diversity of LOF has been developed such as COF, LOCI and LOCI. Vaquez *et al.* [8], suggest an approach for detecting outliers in data with low density which is known as Sparse Data Observers. It is an easy learning and it reduces the computational cost comparison to other rank-based outlier detection method. Another algorithm proposed by Su *et al*. [9], for scattered data called E2DLOS.

Advantages and Disadvantages of Density-based Approach:

Outlier detection based on density is generally more successful and efficient than the distance-based outlier detection method. This approach is non-parametric, meaning it does not use any kind of distribution to fit the data. Most density-based method are considered as best approach for outlier detection, but they are very complicated and computational cost are also expensive. They are sensitive to determine the size of neighbors as parameter. Many algorithms such as INFLO and MDEF, are unable to robustly tackle the data streams due to their implicit complexity and lack of updation of outlierness score.

3. DISTANCE BASED APPROACH

The distance-based technique is based on the distance between the data points. Anomalies are those observations who do not have sufficient neighbors and calculation of neighbors are done by using Euclidean distance or Manhattan distance. This approach mainly applies on low dimensional dataset to detect outliers. This method is non-parametric and can be used to analyses huge datasets with moderate to high density. In comparison to statistical methods, distance-based outlier approach has a more solid foundation and more efficient. Hunag *et al.* [10], suggest a method to rank the neighbors called rank-based outlier detection. Radavanovi's *et al*. [11], proposed a reverse nearest neighbor strategy to deal with curse of dimensionality a fundamental challenge in computing outliers in high- dimensional data sets this approach work on both type of data low as well as high-dimensional data sets.

Advantages and Disadvantages of Distance based approach:

The major and important advantages of this approach is they do not depend on any type of assumed distribution to fit the data because it is non-parametric approach. In scalability aspect, they scale better in multidimensional space and they are computational effective comparative to statistical approaches. There is major drawback of this approach. Due to the dimensionality curse, their performance suffers in high-dimensional space. Using KNN neighborhood along with distance-based approach in high dimensional is being too expensive.

4. CLUSTERING BASED APPROACH

This approach mainly counts on using clustering methods to define the nature of data. Outliers are generally considered as smaller cluster with data points in comparison to another cluster. Clustering is unsupervised method it does not need any prior knowledge or trained data set. Various clustering algorithms are used to detect outliers. K-means is a traditional outlier detection approach which is done with the help of distance measure and centroid. Similarly, k-medoids, hierarchical clustering and Agglomerative clustering. Clustering approach is robust to different types of data set. Several research studies have used clustering-based techniques to reduce the impact of outliers. Zhang [12], present the use of several algorithms and categories them into groups. Ren *et al*. [13], proposed a SD stream algorithm, which employes the sliding window concept. Assent *et al*. [14], suggest AnyOut to detect outliers anytime in streaming data. AnyOut uses ClusTree to construct a precise tree structure. Some questions are taken into consideration while construct a clustering-based approach for outlier detection.

1. Whether the data points are part of a cluster or not, and whether or not an outlier can be identified.
2. Whether cluster and the data points are closer or farther apart. Is it possible to classify is an outlier if it is that far away?
3. Is there a small cluster of data points, and if so, how should the data points within the cluster be labelled?

Advantages and Disadvantages of cluster-based outlier detection approach:

Cluster based approach are robust to different types of data. It is unsupervised method and very acceptable for outlier detection. Clustering is very entrenched area for research and it includes various algorithm to perform according to our need. The most fascinating part of clustering while detecting outliers, after analysing the cluster, new points can be easily insert in cluster and then analyse for outlier detection. This makes a better variation and no prior knowledge required for the distribution of data. Types of clusters like hierarchical and partitioning cluster they are very versatile and maintains a good performance on different types of data sets. Generally clustering techniques rely on the users how they divide the clusters in advance, which creates a difficulty to extract a good result. In clustering method, mostly clusters are identified by visualizing the shape of clusters it can cause some problems to find exact clusters of the data. Outliers in clustering are generally binary there is no quantitative exhibitive of the object outlierness. They also referred as lack of back-tracking capability.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

V. TOOLS FOR OUTLIER DETECTION

Nowadays, outlier detection is done with the help of various tools which reduce time complexity and enhance the efficiency of results. Many tools have been used for outlier detection. some of them we introduce.

- MATLAB

MATLAB is an easy-to-use programme that has a variety of outlier detection methods and function.

- Python Outlier Detection (PyOD)

PyOD is mostly used for multivariate datasets to detect outliers. It is a scalable tool of python which is generally used in commercial projects.

- R programming

R Programming provides many statistical outlier detection packages like abod (angle-based outlier detection) and dbscan (density-based outlier detection) to detect outliers. It is very useful statistical software and most fascinating part of this tool it is freely accessible.

- Scikit-learn Outlier Detection

This tool provides some machine learning tool. It includes algorithm like LOF and isolation forest.

VI. DATASETS FOR OUTLIER DETECTION

Outlier detection algorithm have been used on a variety of data types, including regular and high-dimensional data sets like streaming datasets, network data, uncertain data and time series data. Two types of data are mostly taken into consideration which is real-world and synthetic datasets. Few of them most popular databases that includes a lot of datasets are discussed below:

- The UCI Repository

The UCI repository have more than hundred freely accessible datasets to use for the outlier identification methods. Mostly datasets are related to classification methods, but in outlier scenarios mostly approach is to preprocessing the datasets.

- Outlier detection datasets (ODDS)

ODDS provides free of collection of datasets designed especially for outlier discovery. which is freely accessible. Multi-variate, univariate, time series, time series graph are among the numerous types of datasets.

- ELKI Outlier Datasets:

ELKI provides a various dataset which is suitable for outlier algorithms.

- Unsupervised Anomaly Detection Datasets

Unsupervised outlier are commonly techniques are commonly used in these datasets to discover outliers by comparing them to standards.

Some of the major concerns while handling with datasets for the analysis includes how to tackle with the problem of down sampling of data, normalization, missing values, duplicate data, data cleaning, data authentication. In future it is critical to research how to evaluate the data for outlier detection and what attributes should be considered. When selecting a data set for outlier detection methods, consider the data in terms of useful information that may be applied to the problem specification. For an OD approach involving a data stream, for example, streaming data is preferable to the other types of data.

VII. CONCLUSION

In this paper, we have provided a detailed literature related to outlier detection in a very structured manner by grouping them into different categories. In this survey we have discussed the most important aspects which is advantages and disadvantages of different outlier approach. Furthermore, we have also reviewed about the data tools for outlier detection and also about the databases repository. There is no single standard approach for outlier detection method. Many scholars have used a wide range of approaches covering the complete spectrum of statistical, neural, machine learning as discussed in earlier reviews. We may deduce from the review that the majority of the strategies used are focused on algorithm. These necessitates specialised expertise and the background for recognising outliers differ by domain. It has been discovered that the effectiveness of outlier detection method strongly reliant on data distribution and type. Some of the strategies discussed in this paper, necessitate prior knowledge. We tried to give a wide overview of current strategies, but it is clear that we would not able to cover all of them under one roof.

ACKNOWLEDGEMENT

I would like to offer my heartfelt appreciation to everyone who helped me produce this research report. This paper and research behind it would not be possible without the extreme support of my supervisor. She has taught me the research methodology and to carry out my research. Throughout the process, I am grateful for their diligent direction, constructive criticism and friendly valuable suggestions. I am grateful to every one of them for sharing their accurate and enlightening knowledge with me.

REFERNCES

[1] E. L. Paula, M. Ladeira, R. N. Carvalho, and T. Marzagáo, ‘‘Deep learning anomaly detection as support fraud investigation in Brazilian exports and anti-money laundering’’ in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Anaheim, CA, USA, Oct. 2016, pp. 954–960.

[2] U. Porwal and S. Mukund, ‘‘Credit card fraud detection in e-commerce: An outlier detection approach,’’ 2018, *arXiv:1811.02196*. [Online]. Available: https://arxiv.org/abs/1811.02196

[3] K. Alrawashdeh and C. Purdy, ‘‘Toward an online anomaly intrusion detection system based on deep learning,’’ in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Anaheim, CA, USA, Dec. 2016, pp. 195–200.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

[4] G. Gebremeskel, C. Yi, Z. He, and D. Haile, ''Combined data mining techniques-based patient data outlier detection for healthcare safety,'' *Int. J. Intell. Comput. Cybern.*, vol. 9, no. 1, pp. 42–68, 2016.

[5] John Tukey, "An efficient method for displaying a five number data summary", 1997.

[6] M. Goldstein and A. Dengel, ''Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm,'' in *Proc. Poster Demo Track*, Sep. 2012, pp. 59–63.

[7] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, ''Statistical outlier detection using direct density ratio estimation,'' *Knowl. Inf. Syst.*, vol. 26, no. 2, pp. 309–336, 2011.

[8] F. I. Vázquez, T. Zseby, and A. Zimek, ''Outlier detection based on low density models,'' *Proc. ICDM Workshops*, 2018, pp. 970–979.

[9] S. Su, L. Xiao, L. Ruan, F. Gu, S. Li, Z. Wang, and R. Xu, ''An efficient density-based local outlier detection approach for scattered data,'' *IEEE Access*, vol. 7, pp. 1006–1020, 2019.

[10] H. Huang, K. Mehrotra, and C. K. Mohan, ''Rank-based outlier detection,'' *J. Stat. Comput. Simul.*, vol. 83, no. 3, pp. 518–531, Oct. 2013.

[11] M. Radovanović, A. Nanopoulos, and M. Ivanović, ''Reverse nearest neighbors in unsupervised distance-based outlier detection,'' *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1369–1382, May 2015.

[12] J. Zhang, ''Advancement of outlier detection: A survey,'' *ICST Trans. Scalable Inf. Syst.*, vol. 13, pp. 1–26, Feb. 2013.

[13] J. Ren and R. Ma, ''Density-based data streams clustering over sliding windows,'' in *Proc. Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD)*, Jul. 2009, pp. 248–252.

[14] I. Assent, P. Kranen, C. Baldauf, and T. Seidl, ''AnyOut: Anytime outlier detection on streaming data,'' in *Proc. 17th Int. Conf. Database Syst. Adv. Appl.*, 2012, pp: 228-242.

[15] V. Barnett and T. Lewis, "Outliers in statistical data".

[16] Charu.C.Agarwal, "Outlier Analysis", 2013.