A Review Outlier Detection Types, Causes and Techniques

Maninder Kaur[1], Mandeep Singh Saini[2]
[1]M.Tech (Scholar), [2]Assistant Professor
Department of Information Technology, Chandigarh engineering college, Landran, Mohali, (Punjab)

Abstract – Outlier detection has received considerable attention in the field of data mining because of the need to detect unusual events in a variety of applications, such as fraud detection, human gait analysis and intrusion detection. This paper aims to provide a comprehensive overview of outlier detection, types and techniques of outliers. In addition, we identified the causes and impact of outliers and how to detect and remove the outliers. We've also discussed several techniques of outliers based on different approaches and applications of outliers. Several previous paper are reviewed to get better idea of outliers and its techniques and implementations.

I. INTRODUCTION

An outlier is defined as an information point which is altogether different from remaining information based few measures. Such a point frequently contains helpful data on the irregular conduct of the framework depicted by the information. The exception location strategy finds applications in Visa misrepresentation, arrange interruption identification, financial applications and promoting. This issue commonly emerges with regards to high dimensional informational indexes. A significant part of the current work on finding anomalies utilize strategies which make verifiable presumptions of the generally low dimensionality of the information. These strategies don't perform when dimensionality is high and information winds up noticeably meagre [1]. The values deviated from other observations on data, which indicate a variance in measurement, experimental errors or novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample. In numerous information investigation undertakings countless are being recorded or then again inspected. One of the initial moves towards getting a reasonable examination is the discovery of outlaying perceptions. Despite the fact that anomalies are frequently considered as a mistake or commotion, they may convey critical data. Recognized exceptions are possibility for deviant information that may some way or another antagonistically prompt model misspecification, one-sided parameter estimation and off base outcomes. It is thusly essential to distinguish them before displaying and investigation [2].

II. TYPES OF OUTLIERS

Outliers are of two types:

- ***Univariate outliers*** found while looking at distribution of values in single feature space.
- ***Multivariate outliers*** found in n-dimensional space. Tracing at distributions in n-dimensional spaces can be very difficult for the human brain that is why we need to train a model to do it for us.

Outliers are available in several flavours based on environment:

- ***Point outliers*** are single data points that lay far from the remaining all distribution.
- ***Contextual outliers*** are data noise, like punctuation symbols realizing background noise signal or text analysis while speech recognition.
- ***Collective outliers*** are the subsets of novelties in data e.g. a signal indicates locating new phenomena as shown in figure.

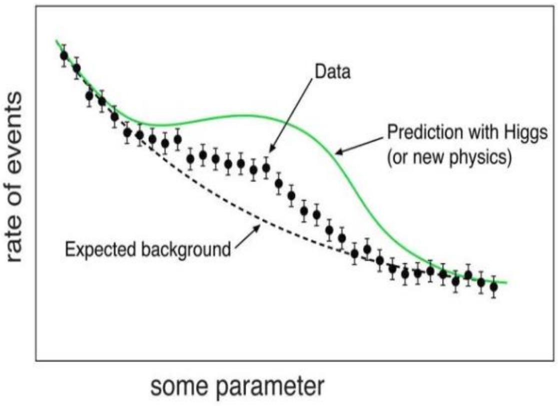

Fig: Collective Outliers

III. CAUSES OF OUTLIERS

A perfect way to handle the outlier is to find out the reasons of having them and dealing with them based on the reason of their existence. Causes of Outliers are categorized as: **Natural and Artificial (error)/Unnatural.** These are further classified as:

- ***Data entry errors:*** Human errors or mistakes caused amid information accumulation, recording, or section can cause exceptions in information.

- ***Measurement errors:*** These are general source of outliers, caused when the measurement instrument used turns out to be faulty.

- ***Experimental errors:*** data extraction or experiment planning/executing errors

- ***Intentional errors:*** dummy outliers that are self-reported to test the detection methods.

- ***Data processing errors:*** data extracted from several datasets while performing data mining, due to which manipulative or extraction errors leads to outliers.

- ***Sampling errors:*** extracting or mixing information from wrong or multiple sources leads to outliers.

- ***Natural errors:*** these aren't an error, but novelties in data.

Outliers can come from many sources and hide in many dimensions in process of producing, collecting, processing and analysing data,.

Outlier detection is significant for almost any quantitative discipline (i.e. Physics, Machine Learning, Economy, Finance, and Cyber Security). The quality of data and quality of prediction model are equally important in machine learning [3]. Outlier Detection includes parts of an expansive range of systems. Numerous methods utilized for recognizing exceptions are essentially indistinguishable yet with various names picked by the creators. In case, researchers portray their different methodologies as anomaly discovery, noise identification, peculiarity discovery, clamour recognition, deviation identification or special case mining.

Impact of Outliers

Outliers can definitely change the consequences of the information examination and factual demonstrating. There are various horrible effects of anomalies in the informational collection:

- It expands the blunder change and lessens the energy of factual tests
- In the event that the exceptions are non-arbitrarily dispersed, they can diminish ordinariness
- They can inclination or impact gauges that might be of substantive intrigue
- They can likewise affect the fundamental suspicion of Regression, ANOVA and other measurable model presumptions.

How to detect Outliers

Generally used method to detect the outlier is visualization. Several visualization methods such as Box plot, Scatter Plot, Histogram. Various researchers use multiple thumb rules to detect outliers [4]. Few of them are:

- Any Value that is past the scope of - 1.5 x IQR to 1.5 x IQR.
- Utilize capping strategies. Any value that stands out of scope of fifth and 95th percentile can be accounted as anomaly.
- Information focuses, at least three standard deviation far from mean are thought about exception.
- Outlier discovery is simply an uncommon instance of examining data for persuasive information focuses and it likewise relies upon the business understanding.
- In SAS, PROC Univariate, PROC SGPLOT are utilized. To distinguish anomalies and powerful perception, factual measure like STUDENT, COOKD, RSTUDENT, etc. were considered additionally.

How to remove Outliers

Various methods used to deal with outliers are identical to techniques such as:

Deleting Observations: Erasing anomaly esteems due to information processing mistake, exception perceptions are less in numbers. We can likewise utilize trimming at the two closures to evacuate exceptions.

Transforming and Binning values: Transforming factors can likewise wipe out exceptions. Normal log of an esteem diminishes the variety caused by extraordinary esteems. Binning is likewise a type of variable change.

Imputing: We can utilize mean, middle, mode ascription strategies. Before attributing esteems, we ought to dissect on the off chance that it is common exception or manufactured. On the off chance that it is counterfeit, we can run with attributing esteems.

Treat separately: If there are noteworthy number of exceptions, we should treat them independently in the measurable model. One approach is to regard the two gatherings as two unique gatherings and fabricate singular model for the two gatherings and after that consolidate the yield.

IV. TECHNIQUES OF OUTLIER DETECTION

Several common methods of outlier detection are:

- ***Extreme Value Analysis*** (parametric approach)***:*** it's most general form of outlier detection useful in 1-D data. It is considered that extremely big or small values are the outliers. Z-test and Student's t-test are examples of these statistical methods. These are good heuristics for initial analysis of data. They can be used as final steps for interpreting outputs of other outlier detection methods [5].

  Z-Score

  The z-score or standard score of a perception is a metric that demonstrates what number of standard deviations an information point is from the example's mean, accepting a Gaussian appropriation. This influences z-to score a parametric technique. Frequently information indicates are not depicted by a gaussian dissemination, this issue can be fathomed by applying changes to information i.e. scaling it.

Some Python libraries like Scipy and Sci-unit Learn have simple to utilize capacities and classes for a simple execution alongside Pandas and Numpy. In the wake of influencing the fitting changes to the chose highlight to space of the dataset, the z-score of any information point can be ascertained with the accompanying expression:

$$z = \frac{x - \mu}{\sigma}$$

When figuring the z-score for each example on the informational index a limit must be indicated. Some great 'thumb-control' limits can be: 2.5, 3, 3.5 or more standard deviations.

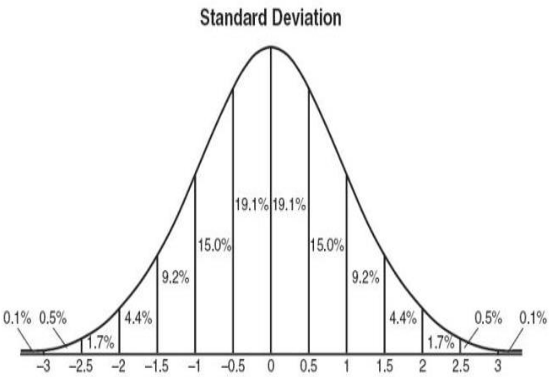

Fig: Normal Curve

It is a simple and robust method to avoid outliers in data if you are dealing with parametric distributions in a low dimensional feature space.

- ***Probabilistic and Statistical Modelling*** (parametric approach)***:*** These models assume specific distributions for data. Then using the expectation-maximization (EM) methods they estimate the parameters of the model. Lastly, computed probability of membership of each data point to calculated distribution. The points with low probability of membership are marked as outliers.

- ***Linear Regression Models (PCA, LMS):*** These methods model the data into a lower dimensional sub-spaces with the use of linear correlations. This distance is used to find outliers. PCA (Principal Component Analysis) is an example of linear models for anomaly detection.

- ***Proximity Based Models*** (non-parametric approach)***:*** The idea with these methods is to model outliers as points which are isolated from rest of observations. Cluster analysis, density based analysis and nearest neighbourhood are main approaches of this kind. Proximity based methods can be classified in 3 categories: 1) Cluster based methods 2) Distance based methods 3) Density based methods.

  Group based strategies order information to various bunches and check focuses which are not individuals from any of referred to bunches as exceptions. While distance based approaches on other hand are more granular and use the distance between individual points to find outliers.

  Dbscan (Density Based Spatial Clustering of Applications with Noise)

  In machine learning and data analytics clustering methods are useful tools that help us visualize and understand data better. Relationships between features, trends and populations in a data set can be graphically represented via clustering methods like dbscan, and can also be applied to detect outliers in nonparametric distributions in many dimensions.

  Dbscan is a density based clustering algorithm, it is focused on finding neighbours by density (Min Pts) on an ‘n-dimensional sphere’ with radius ε. In other words, a cluster is a maximal set of ‘density connected points’ in the feature space.

  Dbscan then defines different classes of points:

  - **C*ore point*:** **A** is a core point if its neighbourhood (defined by ε) contains at least the same number or more points than the parameter Min Pts.
  - ***Border point***: **C** is a border point that lies in a cluster and its neighbourhood does not contain more points than Min Pts, but it is still ‘*density reachable’* by other points in the cluster.
  - ***Outlier***: **N** is an outlier point that lies in no cluster and it is not ‘*density reachable’* nor ‘*density connected’* to any other point. Thus this point will have “his own cluster”.

- ***Statistics Based Models:*** The idea of such methods proves that outliers maximize the minimum code length to describe data set.

V. APPLICATIONS OF OUTLIERS

Various implementations that utilise outlier detection are:

- ***Fraud detection:*** shopping behaviour generally changes in case of stolen credit card. Thus the abnormal pattern of purchase highlights the card abuse.
- ***Loan application processing:*** detects the fraud or problem creating customers [6].
- ***Intrusion detection:*** detects uncertified access in computer network.
- ***Activity monitoring –*** monitors the unusual activities for credit cards, state benefits or mobile phone for suspicious trades in equity markets, performance of computer networks, etc.
- ***Fault diagnosis:*** screening processes to identify issues in motors, pipelines, generators, or space instruments like: space shuttles

- ***Medicine and Public Health Issues:*** unusual symptoms test results show the serious health problem of patient depending on the features of patients like gender age. Diagnoses the presence of particular diseases e.g. tetanus, etc. different aspects like frequency or correlation, etc.
- ***Detecting novelties in images:*** for robot neo-taxis or surveillance systems. Detecting image features moving independently of background.
- ***Sports Statistics:*** several parameters recorded as per the player's performance. Outstanding (positive or negative) identified as abnormal values.
- ***Detecting measurement errors:*** derived experiment values wrongly measured and removing errors in data mining and analysis tasks [7].

VI. LITERATURE REVIEW

Jeremiah D. Deng, (2016) [8] presented a set of outliers detection algorithms available online based on PCA. Outlier detection or anomaly detection is an important and challenging issue in data mining, even so in the domain of energy data mining where data are often collected in large amounts but with little labeled information. Novel algorithmic treatments are introduced to build incremental PCA and kernel PCA algorithms with online learning abilities. Some preliminary experimental results obtained from a real-world household consumption dataset have produced some promising performance for the proposed algorithms.

Aya Ayadi et al., (2017) [9] attempted to give structured review on present work relative to outlier detection techniques that classification based and could be implemented to wireless network. These days, many wireless sensor networks have been distributed in the real world to collect valuable raw sensed data. The challenge is to extract high-level knowledge from this huge amount of data. However, the identification of outliers can lead to the discovery of useful and meaningful knowledge. In the field of wireless sensor networks, an outlier is defined as a measurement that deviates from the normal behavior of sensed data. Many detection techniques of outliers in WSNs have been extensively studied in the past decade and have focused on classic based algorithms. These techniques identify outlier in the real transaction dataset**.** Thus, we have identified key hypotheses, which are used by these approaches to differentiate between normal and outlier behavior. Additionally, they tried to provide an easier and brief understanding of classification based techniques. Furthermore, they identified the advantages and disadvantages of different classification based techniques and we presented a comparative guide with useful paradigms for promoting outliers detection research in various WSN applications and suggested further opportunities for future research.

Mahsa Salehi et al., (2016) [10] proposed an organized memory incremental local outlier detection algorithm (MiLOF) and a flexible version having accuracy nearing iLOF with fixed memory bound. Outlier detection is an important task in data mining. Due to increased demand of high speed analyzed data streams, outlier detection becomes more challenging as compared to existing detection techniques. The popular Local Outlier Factor (LOF) algorithm has an incremental version (called iLOF), which assumes unbounded memory to keep all previous data points. In addition MiLOF F is robust to changes in the number of data points, underlying clusters and dimensions in the data stream. Experimented results proves that both proposed approaches have better memory and time complexity than Incremental LOF. Additionally, they showed that MiLOF_F is robust to changes in number of data points, underlying clusters and dimensions in data stream. MiLOF / MiLOF_F are suitable to implementation surroundings with limited memory and also applicable to high volume data streams.

Ke Yan et al., (2016) [11] proposed a hybrid outlier detection method namely Pruning-based K-Nearest Neighbour (PB-KNN), which unites cluster based and density-based methods with KNN algorithm for efficient performance. Technology advancements in health care informatics, digitalizing health records, and telemedicine has resulted in rapid growth of health care data. One challenge is how to effectively discover useful and important information out of such massive amount of data through techniques such as data mining. Outlier detection is a typical technique used in many fields to analyse big data. However, for the large scale and high dimensional heath care data, the conventional outlier detection methods are not efficient. The proposed PB-KNN adopts the case classification quality character (CCQC) as the medical quality evaluation model and uses the attribute overlapping rate (AOR) algorithm for data classification and dimensionality reduction. To evaluate the performance of the pruning operations in PB-KNN, we conduct extensive experiments. The experiment results show that the PB-KNN method outperforms the k-nearest neighbour (KNN) and local outlier factor (LOF) in terms of the accuracy and efficiency.

Jia Liu et al., (2016) [12] proposed and algorithm for outlier detection to overcome the issue of traditional methods i.e., performance is sensitive to parameter k, and interpretability is not strong. Outlier objects have lower density than their neighbours and relatively large distance from objects with higher density. Outlier is great concern in machine learning task. They've compared the proposed method with other existing methods based on various types of synthetic datasets. They also applied the proposed method in real water quality data. The results of numerical experiments indicated that the proposed method has better effectiveness, stability, and interpretability on detection of low-density outlier detection.

Haizhou Du, (2015) [13] proposed a robust method for outlier detection with statistical parameters, which incorporates clustering based ideas with big data. With the rapid expansion of data scale, big data mining and analysis has attracted increasing attention. Outlier detection as an important task of data mining is widely used in many applications. However, conventional outlier detection methods have difficulty handling large-scale datasets. In addition, most of them typically can only identify global outliers and are over sensitive to parameters variation. Firstly, this method finds some density peaks of dataset by 3σ standard. Secondly, each remaining data object in dataset is assigned to same cluster as its nearest neighbour of higher density. Finally, they use Chebyshev's inequality and density peak reachability to identify local outliers of each group. The experimental results demonstrate the efficiency and accuracy of the proposed method in identifying both global and local outliers, Moreover, the method also proved more robust analysis than typical outlier detection methods, such as LOF and DBSCAN.

Madhu Shukla et al., (2015) [14] reviewed different techniques of outlier detection to stream data and their issues in detail and presented results of the same. Data mining is one of the most exciting fields of research for the researcher. As data is getting digitized, systems are getting connected and integrated, scope of data generation and analytics has increased exponentially. Today, most of the systems generate non-stationary data of huge, size, volume, occurrence speed, fast changing etc. these kinds of data are called data streams. One of the most recent trend i.e. IOT (Internet Of Things) is also promising lots of expectation of people which will ease the use of day to day activities and it could also connect systems and people together. This situation will also lead to generation of data streams, thus present and future scope of data stream mining is highly promising. Characteristics of data stream possess many challenges for the researcher; this makes analytics of such data difficult and also acts as source of inspiration for researcher. Outlier detection plays important role in any application.

VII. CONCLUSION

This paper gives a brief overview about the classification strategies for outlier detection techniques. Outlier Detection (otherwise called uncommon occasion or irregularity discovery) has gotten impressive consideration in the field of information mining on account of the need to identify surprising occasions in an assortment of uses, for example, misrepresentation location, human step investigation and interruption identification. This paper expects to give a far reaching review of anomaly discovery, sorts and methods of exceptions. What's more, we recognized the causes and effect of anomalies and how to distinguish and evacuate the exceptions. We've additionally talked about a few procedures of exceptions in view of various methodologies and utilizations of anomalies. A few past paper are surveyed to show signs of improvement thought of anomalies and its procedures and executions.

VIII. REFERENCES

[1]. Aggarwal, Charu C., and Philip S. Yu. "Outlier detection for high dimensional data." In *ACM Sigmod Record*, vol. 30, no. 2, pp. 37-46. ACM, 2001.

[2]. Williams G. J., Baxter R. A., He H. X., Hawkins S., Gu L., "A Comparative Study of RNN for Outlier Detection in Data Mining," IEEE International Conference on Data-mining (ICDM'02), Maebashi City, Japan, CSIRO Technical Report CMIS-02/102, 2002

[3]. "A Brief Overview Of Outlier Detection Techniques – Towards Data Science". 2018. Towards Data Science. https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561.

[4]. Exploration, A, A Exploration, and Sunil Ray. 2018. "A Complete Tutorial Which Teaches Data Exploration In Detail". Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/.

[5]. "Introduction To Outlier Detection Methods". 2018. *Datasciencecentral.Com*. https://www.datasciencecentral.com/profiles/blogs/introduction-to-outlier-detection-methods.

[6]. Hodge, Victoria, and Jim Austin. "A survey of outlier detection methodologies." *Artificial intelligence review* 22, no. 2 (2004): 85-126.

[7]. "Cite A Website - Cite This For Me". 2018. Dbs.Ifi.Lmu.De. http://www.dbs.ifi.lmu.de/~zimek/publications/SDM2010/sdm10-outlier-tutorial.pdf.

[8]. Deng, Jeremiah D. "Online Outlier Detection of Energy Data Streams Using Incremental and Kernel PCA Algorithms." In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, pp. 390-397. IEEE, 2016.

[9]. Ayadi, Aya, Oussama Ghorbel, M. S. Bensaleh, Abdelfateh Obeid, and Mohamed Abid. "Performance of outlier detection techniques based classification in Wireless sensor networks." In *Wireless Communications and Mobile Computing Conference (IWCMC), 2017 13th International*, pp. 687-692. IEEE, 2017.

[10]. Salehi, Mahsa, Christopher Leckie, James C. Bezdek, Tharshan Vaithianathan, and Xuyun Zhang. "Fast memory efficient local outlier detection in data streams." *IEEE Transactions on Knowledge and Data Engineering* 28, no. 12 (2016): 3246-3260.

[11]. Yan, Ke, Xiaoming You, Xiaobo Ji, Guangqiang Yin, and Fan Yang. "A Hybrid Outlier Detection Method for Health Care Big Data." In *Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom), 2016 IEEE International Conferences on*, pp. 157-162. IEEE, 2016.

[12]. Liu, Jia, and Guoyin Wang. "Outlier detection based on local minima density." In *Information Technology, Networking, Electronic and Automation Control Conference, IEEE*, pp. 718-723. IEEE, 2016.

[13]. Du, Haizhou. "Robust local outlier detection." In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, pp. 116-123. IEEE, 2015.

[14]. Shukla, Madhu, Y. P. Kosta, and Prashant Chauhan. "Analysis and evaluation of outlier detection algorithms in data streams." In *Computer, Communication and Control (IC4), 2015 International Conference on*, pp. 1-8. IEEE, 2015.