Review paper on Outlier Detection using Machine Learning Technique

Siddarth Pagariya
M. Tech. Scholar
University Institute of Technology, BU, Bhopal

Dr. Rachna Kulhare
Asst. Professor
University Institute of Technology, BU, Bhopal

Abstract: - OD is a Data Mining Application. Anomaly contains boisterous information which is explored in different areas. The different strategies are now being explored that is more conventional. We reviewed on different procedures and uses of OD that gives an original methodology that is more helpful for the novices. Machine learning methods are widely used for prediction and classification tasks in medical diagnosis. The classification of a disease with greater precision and efficiency for disease diagnosis are the goals of ML methods. The life support equipment and systems for patients are expanding gradually. Human life expectancy rises as a result of this growth. Yet, these medical care frameworks face the few difficulties and issues like deceiving patients' data, protection of information, absence of exact information, absence of medico data, classifiers for expectation and some more. Numerous disease diagnosis and prediction systems, including expert systems, clinical prediction systems, decision support systems, and personal health record systems, have been developed to address these issues. The objective of the proposed system is to assist physicians in making accurate diagnoses of heart and diabetes conditions.

Keywords: - Outlier Detection (OD), Data Mining, Machine Learning (ML)

I. INTRODUCTION

Finding the information in data is an important task in many applications in discovery of criminal activities, in electronic business, credit card fraud detection and network intrusion detection. Outlier detection approaches focus on discovering patterns that occur infrequently in the data, as opposed to traditional data mining techniques that attempt to find patterns that occur frequently in the data. One of the most widely accepted definitions of an outlier pattern is provided by Hawkins [84] is: “An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism”, and is frequently treated as noise that is required to be removed from a dataset in order to build an accurate model. However, outlier detection techniques can also be used to discover important information in the data, “one person‟s noise is another person’s signal” [65]. Outlier detection strategy can also be used to clean the data before applying any algorithm. Examples of the outlier discovery are irregular credit card transactions, credit card fraud and abnormal symptoms of patients due to suffering from a specific disease or ailment. Most of the research efforts in outlier detection strategies have focused on datasets that are comprised of numerical attributes or ordinal attributes that can be directly mapped into numerical values.

Quite often, when the data with categorical attributes, it is assumed that the categorical attributes could be easily mapped into numerical values. However, there are cases of categorical attributes where this mapping to numerical attributes is not a straightforward process, and the results greatly depend on the mapping that is used. Recently there has been some focus on data with categorical or mixed attributes [1-4]. Yet, these efforts have not been contrasted to each other and they have been evaluated using different datasets. Another issue that has only recently gained focus in the literature is related to the large and distributed nature of the datasets available today. With the explosion of technology, the size of data for a particular application has grown and will continue to grow. In addition, most of the data is distributed among different sites belonging to the same or different organizations. Transferring the data to a central location and then detecting is usually impractical because of the size of the data and the expense of constantly moving it, without accounting for data ownership and control issues. Hence, successful outlier detection strategies must perform well and be scalable as the size and dimensionality of the dataset grows. Furthermore, in order to deal with the distributed nature of the data, the communication overhead and synchronization between the different sites in which the data resides should be minimized; consequently, the passes over the data should be minimal.

II. LITERATURE REVIEW

Abdallah Abdellatif et al. [1], suggested a anomaly detection was an essential task in data mining that have an intention to found components to show the different behavior compared to the other forms and such were known as outliers. One amongst the broadly utilized measures for determining whether a component was an exception depends on the sum of neighboring components within a distance and a limit. That types of outliers were removed based on the exceptions. It states both an extendible structure for anomaly identification

calculations and exact exception location calculations were mentioned, where the anomaly recognition was persistently completed over an information stream. These algorithms were realized in MOA framework, for extend its process with outlier detection proficiencies. One more significant problem for additional exploration was the capability for visualizing the data streams in any metric space. The multidimensional data sets were reinforced (specifically, 2D data sets) for visualization limitations. This challenge was incorporated into MOA operative perception of common metric spaces.

H. B. Kibria et al. [2], recommended an Improved Genetic KMeans (IGK) algorithm to proficiently discover the outliers. The main concern of this clustering-based outlier detection algorithm was used to identify the outliers and data clustering concurrently. In this, it regularly observed the noise, which should be detached in order to provide more reliable clustering method. During the process of clustering and outlier discovery they estimate the centroids of the generative distribution process. The IGK was an efficient clustering technique that handles the large amount data with the help of Genetic Algorithm (GA). The findings of this technique were as used to: avoid the disserving big clusters. In some degree, it overwhelmed the deflection of data and it reduced the sensitivity to isolated point. Here, the outlier detection could be accomplished only on numeric dataset. When the clustering method was utilized in outlier detection, then they focused mainly on those elements as outliers which was deceitful outside the clusters.

G. Magesh et al. [3], described the basic process of outlier detection in data mining tasks were well examined because of its various applications. In that most applications occur in high-dimensional spaces. A blockage of prevailing methodologies was certain or, on the other hand unambiguous evaluations on ideas of separation or closest neighbor were decayed in high dimensional data. The utilization of angle-based outlier aspect was considered in mining high-dimensional outliers. That technique run in cubic time with a quadratic time heuristic, they suggest a new irregular projection-based strategy that can assess the edge-based outlier aspect for the data which were focused on time linear in the data size. Additionally, their method was reasonable to perform in corresponding condition to accomplish an equivalent speedup. They offered a hypothetical study of the quality estimation to ensure the fixed quality of this proposed assessment system. The observational tests on manufactured and real-world datasets exhibit the scalability, efficacy and competence to identify the outliers in large high-dimensional data sets.

Kangqing Yu. et al. [4], introduces an data mining (DM) based way to deal with creating outfit models for anticipating following day vitality utilization and pinnacle control request, with the point of enhancing the forecast precision. In addition, an outlier detection method was also offered to detect the abnormal building operative patterns. It was more implemented for analyzing the huge energy consumption data of the highest building. Three different stages were involved in this approach, they were: Right off the bat, exception location, which consolidates highlight extraction, grouping examination, and the summed up extraordinary studentized veer off (GESD), was performed to evacuate the anomalous day by day vitality utilization profiles. Also, the recursive element end (RFE), an installed variable determination strategy, was connected to choose the ideal contributions to the base expectation models grown independently utilizing eight famous prescient calculations. The outcomes additionally demonstrate that the exception identification strategy has powerful in recognizing the rare day by day energy utilization profiles. The RFE procedure can fundamentally diminish the calculation stack while upgrading the model execution. The gathering models were important for creating methodologies of fault identification and determination in advance operation. The multiple linear regression (MLR) and ARIMA models, do not execute splendidly, since the building associated processes were usually nonlinear and difficult.

Yu. K. et al. [5], discussed about the exposure of distance-based outliers from huge dimensions of data stream was dangerous for current applications extending from credit card fraud detection to moving object monitoring. It considered a framework to tackle three different classes of distance-based outliers in the streaming atmospheres. They were: Minimal Probing standard used a lightweight analytical operation together least yet adequate proof for outlier detection. Lifespan-aware prioritization rule use the fleeting connections among stream information focuses to organize the handling request among them amid the testing process. Guided by these two standards, we plan an exception identification methodology which was ended up being ideal in the costs of CPU expected to decide the exception status of any information point amid its whole life. Scalability was needed to improve in modern distributed multi-core clusters of machines for outlier detection.

N. Khateeb et al. [6], depict a probabilistic, nonparametric strategy for irregularity identification, in light of a squared-misfortune objective capacity which has a straightforward logical arrangement. The technique rises out of expanding ongoing work in nonparametric least squares order to incorporate a "nothing from what was just mentioned" class which models oddities as far as non-anamalous preparing information. The strategy shares the adaptability of other piece based abnormality identification techniques, yet is regularly a lot quicker to prepare and test. It can likewise be utilized to recognize different inlier classes and peculiarities. The probabilistic nature of the result makes it direct to apply in any event, when test information has underlying

conditions; we show how a secret Markov model structure can be consolidated to recognize irregular aftereffects in a test grouping. Exact outcomes on datasets from a few areas show the technique to have similar discriminative execution to well-known other options, however with an unmistakable speed advantage.

Nonso Nnamoko et al. [7], persistent anomaly location in information streams has significant applications in misrepresentation location, network security, and general wellbeing. The appearance and flight of information objects in a streaming way force new difficulties for exception location calculations, particularly in reality productivity. In the previous ten years, a few investigations have been performed to resolve the issue of distance-based anomaly location in information streams (DODDS), which takes on a solo definition furthermore, has no distributional presumptions on information values. Our work is spurred by the absence of near assessment among the best in class calculations utilizing the same datasets on a similar stage. We methodically assess the latest calculations for DODDS under different stream settings and anomaly rates. Our broad outcomes show that in many settings, the MCOD calculation offers the unrivaled execution among every one of the calculations, including the most late calculation Thresh LEAP.

Muhammad et al. [8], information stream is a recently arising information model for applications like climate checking, Web click stream, network traffic observing, and so forth. It comprises of a limitless succession of information focuses went with timestamp coming from outer information source. Normally information sources are found nearby and truly defenseless against outside assaults and regular catastrophes, in this manner exceptions are extremely normal in the datasets. Existing strategies for exception recognition are deficient for information streams in light of its transformative information conveyance and vulnerability. In this paper we propose an exception location strategy, called Distance-Based Outline Detection for Data Streams (DBOD-DS) in light of an original constantly versatile likelihood thickness work that resolves every one of the new issues of information streams. Broad investigations on a genuine dataset for meteorology applications show the matchless quality of DBOD-DS over existing procedures with regards to exactness.

Wang Q. et al. [9], anomaly identification is a deep rooted area of measurements however the vast majority of the current exception discovery strategies are intended for applications where the whole dataset is accessible for irregular access. A common anomaly location procedure builds a standard information dissemination or model and recognizes the strayed elements from the model as exceptions. Obviously these strategies are not appropriate for online information streams where the whole dataset, because of its unbounded volume, isn't accessible for arbitrary access. Additionally, the information circulation in information streams change over the long haul which challenges the current anomaly identification strategies that expect a steady standard information dissemination for the whole dataset. Likewise, information streams are portrayed by vulnerability which forces further intricacy. In this paper we propose a versatile, online anomaly recognition method tending to the previously mentioned attributes of information streams, called Adaptive Outlier Detection for Data Streams (A-ODDS), which distinguishes exceptions regarding every one of the got relevant items along with transiently close elements. The transiently close information focuses are chosen in view of time and change of information dissemination. We likewise present a productive and online execution of the method and a presentation concentrate on showing the prevalence of A-ODDS over existing procedures as far as precision and execution time on a genuine dataset gathered from meteorological applications.

Hanifah et al. [10], have fact that deal with information streams makes among various enormous Data applications those. An information stream is an arrangement of items with timestamps that has the properties of transiency, boundlessness, vulnerability, idea float, and multi-dimensionality. In this paper we propose an exception identification method called Orion that tends to every one of the attributes of information streams. Orion searches for an extended component of multi-layered elements with the assistance of a transformative calculation, and recognizes an informative item as an anomaly assuming it lives in a low-thickness area in that aspect. Tests contrasting Orion and existing procedures utilizing both genuine and engineered datasets show that Orion accomplishes a normal of 7X the accuracy, 5X the review, and a serious execution time contrasted with existing methods.

Problem Formulation:-

In the semi-administered strategy, the marked and unlabeled information are utilized to recognize the exceptions. The semi-regulated methodologies are trailed by the analysts as announced. The creators introduced a fluffy unpleasant c-implies grouping to identify the anomalies. In this framework, the ordinary occasions are utilized to assemble the outfit component to distinguish the oddity from the got examples. It utilized the entropy measure to distinguish the anomalies. At first, the unfaltering negative examples are taken from unlabeled and positive information, and afterward the exceptions are recognized dependent on the entropy score to eliminate the anomalies. Likewise, introduced a score base anomaly discovery utilizing stochastic organization technique. A semi-supervised cluster was also proposed in the literature to detect the outliers from the digital mammograms. The number of False-positives is quite high in some specific cases, which can be further reduced. The result is undesirable when processing high-dimensional data. An algorithm

for detecting Type III outliers has much to be researched.

III. OUTLIER DETECTION IN DATA MINING

In the dataset, the outlier is an anomaly that differs from the different data points. In data mining it is also termed as abnormalities, deviants, abnormalities and anomalies. In data mining, the major problems occur are the outlier detection and future prediction techniques. It is defined as the process of finding outliers that depending up on the behavior and distribution of data. The discoveries of abnormal features with inconsistent characteristics are one of the intension of outlier detection. The regression modelling, removes the outlier method and it considers separately to improve the accuracy [11, 12]. There are various categories of outliers namely, Point outliers, Context outliers, Collaborative outlier, vector outlier, sequence outlier and graph outlier. The primary step of data mining applications is the outlier detection There are numerous methods associates with outlier detection, such as differentiating amongst the univariate vs multivariate techniques and parametric vs non-parametric measures. If outliers carry some information, it may consider as error or noise. Outlier detection methods suggests for various applications such as detection of fraudulent in credit card, voting irregularity examination, medical trials, data cleaning, network interruption, severe weather forecast, environmental information and other data mining tasks. There are various difficulties in analyzing an outlier in its present form [13]. Encompassing every possible normal behavior in the region. The normal and outlier have an imprecise boundary. It is difficult to determine and remove, due to noise in the data. Due to the contrary notion of outliers, it is difficult to apply the technique in one domain to another domain [14, 15].

Outlier Detection Methodologies

- Statistical-Based OD
- Deviation-Based OD
- Distance-Based OD

Challenges of OD

- Displaying common substances and exceptions suitably
- Extreme to appraise all probable commonplace exercises in an application The limit among the average and exception substances is normally an defined situation
- Application-explicit exception discovery Selection of distance measure in the midst of substances and the model of association between the substances are frequently application-dependent.
- Tackling noise in outlier detection
- Noise may falsify the typical substances and distort the difference amongst the typical substances and outliers. It helps to hide the outliers and minimizes the efficacy of outlier detection
- Understandability Understanding the outliers: Validation of the detection Identify the grade of an outlier: the impossibility of the item being made by a standard strategy.

IV. TYPES OF OUTLIERS

A vital part of an exception recognition method is the idea of the ideal anomaly. Exception Classification is done based on their event; for the most part there are three sorts of anomalies which are counted as follows [16, 17]:

- Point Outliers
- Context oriented Outliers
- Aggregate Outliers

Point Outlier: when an information occurrence is not the same as set of information then, at that point, occasion is named as point anomaly. It is the most straightforward type of exception and utilized in different explores. For instance Visa misrepresentation Detection, the anomaly can be identified concerning sum spent assuming consumption is higher contrasted with ordinary exchanges then it is an exception [18].

Context oriented Outlier: when an information occurrence is abnormal regarding some unique situation (condition), then, at that point, case said to be Contextual Outlier. Context oriented exceptions for the most part investigated on time series information. For instance, in setting old enough a six feet grown-up might be a typical individual while six feet kid is an exception [19].

Aggregate Outlier: When an assortment of related information is irregular from rest of the whole informational index, then, at that point, it is an aggregate Outlier. They can happen just in informational indexes where information occurrences are connected. Aggregate anomaly has been investigated on graphical information, consecutive information and spatial information.

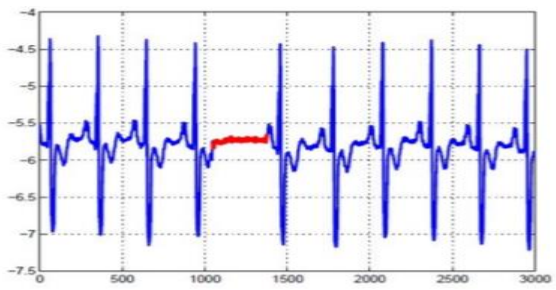

Fig. 2: Collective Outlier

For instance Human Electrocardiogram yield showed in Figure 2. The featured district is an anomaly since same worth exists for the unusually lengthy timespan.

Aggregate anomalies can be applied for chart information, grouping information and spatial information [20, 21].

V. MACHINE LEARNING

The terms "machine learning" and "artificial intelligence" have a big influence on the figures. The increasing use of AI in our daily lives has led to an increase in its popularity. It's now found in a wide range of large and small devices and machines. Regardless of the situation, everyone is on high alert. “We should look at the short history of machine learning in order to dispel these myths. ML is a cycle of providing information to PC frameworks so that the PC can learn to measure and perform the action in the future without having to be modified or cared for with similar or additional information.” The ability to think is being added to PCs to make them more intelligent and user-friendly [22].

Their worth will skyrocket, and they will become an essential resource for humanity. The application of ML can benefit practically every area of epistemology. Computational life structures, gaming, portable sites, regular language preparation and robot development and motion, clinical finding, arrangement mining, conduct examination, etymology and interpretation, misrepresentation identification, and so on are some of the current applications. The list continues to shrink. The acoustic signals of rotating machines were obtained and used with wavelets to help determine the results. The multiracial highlights of the wavelet chiefs were shown to be a viable contender for the rotating system's problem discovery [23, 24]. The wavelet highlights were separated from the vibration signals, which were then arranged using a choice tree algorithm. When combined with wavelet highlights, the J48 calculation proved to be the most effective. In order to better understand how to find flaws in software systems, fuzzy reasoning and unpleasant sets have also been studied. To lay out some fundamental principles for this project, a fluffy motor and harsh sets were used. Thankfully, the outcomes were favorable, and the strategies were well-received [25].

VI. CONCLUSION

We presume that basic examination on uses of exception recognition will help in additional exploration draws near. Exception data is exceptionally helpful when information is contrasted and the first information. The above basic audit will help in the further examination. Exception recognition approaches gives a basic and substantial result for the given information. Our exploration work remembers the basic examination for the different application areas and methods of the exception identification. It has been an extraordinary work for the individuals who need to begin the exploration on anomaly discovery and its space. The whole work comprises various stages and loads of hypothetical ideas in regards to the Anomalies.

We aim to propose new solutions that overcome aforementioned challenges in streaming context and adopt the sliding window technique, but efficiently store in memory a statistical summary of obsolete data, which contributes to the prediction of future data.

REFERENCES

[1] Abdallah Abdellatif, Hamdan Abdellatef, Jeevan Kanesan, Chee- Onn Chow, Joon Huang Chuah and Hassan Muwafaq Gheni, “An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyper parameter Optimization Methods”, IEEE Access, Vol. 10, 2022.

[2] H. B. Kibria and A. Matin, “The severity prediction of the binary and multi-class cardiovascular disease machine learning- based fusion approach”, Comput. Biol. Chem., vol. 98, Art. no. 107672, 2022.

[3] G. Magesh and P. Swarnalatha, “Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction”, Evol. Intell., vol. 14, no. 2, pp. 583_593, Jun. 2021.

[4] Kangqing Yu, Wei Shi and Nicola Santoro, “Designing a Streaming Algorithm for Outlier Detection in Data Mining—An Incrementa Approach”, Sensor, MDPI 2020.

[5] Yu, K.; Shi, W.; Santoro, N.; Ma, X., “Real-time Outlier Detection over Streaming Data”, In Proceedings of the IEEE Smart World Congress, Leicester, UK, 19–23 August 2019.

[6] N. Khateeb and M. Usman, “Efficient heart disease prediction system using K-nearest neighbor classification technique”, in Proc. Int. Conf. Big Data Internet Thing (BDIOT), pp. 21_26, 2017.

[7] Nonso Nnamoko and Ioannis Korkontzelos, “Efficient treatment of outliers and class imbalance for diabetes prediction”, Artificial Intelligence in Medicine, Volune 104, 01-12, 2020.

[8] Muhammad Fazal Ijaz, Ganjar Alfian, Muhammad Syafrudin and Jongtae Rhee, “Hybrid Prediction Model for Type 2 Diabetes and Hypertension Using DBSCAN-Based Outlier Detection, Synthetic Minority Over Sampling Technique (SMOTE), and Random Forest”, Applied Science, Vol. 8, 01-22, 2018.

[9] Wang, Q.; Luo, Z.; Huang, J.; Feng, Y.; Liu, Z. A novel ensemble method for imbalanced data learning: Bagging of extrapolation-SMOTE SVM. Comput. Intell. Neurosci. 2017.

[10] Hanifah, F.; Wijayanto, H.; Kurnia, A., “SMOTE Bagging Algorithm for Imbalanced Dataset in Logistic Regression Analysis”, Appl. Math. Sci., 9, 6857–6865, 2015.

[11] Tantithamthavorn, C.; Hassan, A.; Matsumoto, K., “The impact of class rebalancing techniques on the performance and interpretation of defect prediction models”, arXiv, 2018.

[12] Sadik, S.; Gruenwald, L., “DBOD-DS: Distance based outlier detection for data streams”, In Proceedings of the 21st International Conference on Database and Expert Systems Applications (DEXA), Bilbao, Spain, 30 August–3 September pp. 122–136, 2010.

[13] Sadik, S.; Gruenwald, L., “Online Outlier Detection

for Data Streams", In Proceedings of the 15th Symposium on International Database Engineering & Applications Symposium (IDEAS), Lisboa, Portugal, 21–23, pp. 88–96, 2011.

[14] Sadik, S.; Gruenwald, L.; Leal, E., "In pursuit of outliers in multi-dimensional data streams", In Proceedings of the 4th IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 512–521.

[15] Forrest, S., Esponda, F., and Helman, P., "Aformal framework for positive and negative detection schemes", In IEEE Transactions on Systems, Man and Cybernetics, Part B. IEEE, 357 – 373, 2004.

[16] Gonzalez, F. A. and Dasgupta, D., "Outlier detection using real-valued negative selection", Genetic Programming and Evolvable Machines, Vol. 4, No. 4, pp. 383- 403, 2003.

[17] Lee, W., Stolfo, S. J., and Mok, K. W., "Adaptive intrusion detection: A data mining approach", Artificial Intelligence Review, Vol. 14, No. 6, pp. 533 – 567, 2000.

[18] Gwadera, R., Atallah, M. J., and Szpankowski, W., "Reliable detection of episodes in event sequences", Knowledge and Information Systems, Vol. 7, No. 4, pp. 415 – 437, 2005.

[19] Chow, C. and Yeung, D.-Y., "Parzenwin down network intrusion detectors", In Proceedings of the 16th International Conference on Pattern Recognition, Vol. 4, IEEE Computer Society, Washington, DC, USA, 40385, 2002.

[20] Lin, J., Keogh, E., Fu, A., and Herle, H. V., "Approximations to magic: Finding unusual medical time series", In Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems, IEEE Computer Society, Washington, DC, USA, 329 – 334, 2005.

[21] Adeli, E, Thung, KH, An, L, Wu, G, Shi, F, Wang, T & Shen, D., "Semi-supervised discriminative classification robust to sample-outliers and feature-noises", IEEE Transactions on Pattern Analysis and Machine Intelligence (Early Access), pp. 1-1, 2018.

[22] Ahmad, S, Lavin, A, Purdy, S & Agha, Z., "Unsupervised real-time anomaly detection for streaming data", Neurocomputing, vol. 262, pp. 134-147, 2017.

[23] Alcalá-Fdez, J, Alcalá, R, Gacto, MJ & Herrera, F., "Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms", Fuzzy Sets and Systems, vol. 160, no. 7, pp. 905-921, 2009.

[24] Au, WH & Chan, KC., "Mining fuzzy association rules in a bank account database", IEEE Transactions on Fuzzy Systems, vol. 11, no. 2, pp. 238-248, 2003.

[25] Cai, R, Liu, M, Hu, Y, Melton, BL, Matheny, ME, Xu, H, Duan, L & Waitman, LR., "Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports", Artificial Intelligence in Medicine, vol. 76, pp. 7-15, 2017.