OUTLIER DETECTION USING IOT AND MACHINE LEARNING

[1]MS. DEORE PRIYANKA DEVIDAS, [2]MS. DEVHADRAO GAYATRI VILAS, [3]MS. KADAM MONALI BHAGWAN
[4]MS. BORNARE APOORVA RAGHUNATH, [5]PROF.CHANDGUDE A.S

DEPARTMENT OF COMPUTER ENGINEERING
S.N.D. COLLEGE OF ENGINEERING & RESEARCH CENTRE,
BABHULGAON YEOLA – 423401 2021 - 2022

***Abstract*: Outliers once upon a time regarded as noisy data in statistics, has turned out to be an important problem which is being researched in diverse fields of research and application domains. Many outlier detection techniques have been developed specific to certain application domains, while some techniques are more generic. Some application domains are being researched in strict confidentiality such as research on crime and terrorist activities. The techniques and results of such techniques are not readily forthcoming. A number of surveys, research and review articles and books cover outlier detection techniques in machine learning and statistical domains individually in great details. In this paper we make an attempt to bring together various outlier detection techniques, in a structured and generic description. With this exercise, we hope to attain a better understanding of the different directions of research on outlier analysis for ourselves as well as for beginners in this research field who could then pick up the links to different areas of applications in details.**

***Keywords*: Machine Learning, Processing, Dataset, Support Vector Machine, Database, IOT.**

INTRODUCTION

Outliers once upon a time regarded as noisy data in statistics, has turned out to be an important problem which is being researched in diverse fields of research and application domains. Many outlier detection techniques have been developed specific to certain application domains, while some techniques are more generic. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Examination of the data for unusual observations that are far removed from the mass of data. These points are often referred to as outliers. An outlier is an observation of data that does not fit the rest of the data. It is sometimes called an extreme value. When you graph an outlier, it will appear not to fit the pattern of the graph.

MOTIVATION

The term exception, otherwise called irregularity is initially taken from the field of insights. Exceptions can be raised on account of human blunder, machine mistake, mechanical blames and changes in the conduct of framework or might be because of regular abnormality in the climate. Overcome this all cause's data entry an experiment measurement errors, sampling problems, and natural variation. An error can occur while experimenting/entering data. During data entry, a typo can type the wrong value by mistake.

LITRATURE SURVEY

This chapter contains the existing and established theory and research in this report range. This will give a context for work which is to be done. This will explain the depth of the system. Review of literature gives a clearness and better understanding of the exploration/venture. A literature survey represents a study of previously existing material on the topic of the report. This literature survey will logically explain this system.

An Improved LOF Outlier Detection Algorithm LOF (Local Outliers Factor) algorithm is a very classic anomaly detection algorithm. In order to detect the outliers more accurately, avoid too much testing error, and ensure the detection can be implemented relatively accurately in the data set without professional knowledge, on the basis of traditional LOF algorithm, an improved detection algorithm LOF Outliers is proposed. According to the different distribution densities of logarithmic data points, all the data point sets A1 that are most likely to become outliers are found out. Then, the information entropy weighted LOF algorithm is used to detect the data set to get the result A2. The point set A1 is intersected with A2 to get the final point set A which is the final result. The experimental results show that the algorithm is feasible, and it is more accurate and contains fewer false detection points[1].

A k-Nearest Neighbor Medoid-Based Outlier Detection Algorithm Outlier detection techniques are well known for identifying a small amount of data objects named outliers that are far away from clusters and exist in sparse regions of data space. However, most outlier detection algorithms based on k nearest neighbors are sensitive to parameter k. The outlier detection algorithms based on clusteringrely on specific clustering algorithms, and outliers are by-products. To partially circumvent these problems, motivated by the medoid concept of Kmedoids clustering algorithms, in this paper, we propose a k-nearest neighbor medoid-based outlier detection method that is easy to implement and can provide competing performances with existing solutions. At the same time, a method to determine the parameter k is proposed in combination with the outlier detection algorithm proposed in this paper. Experiments performed on real datasets demonstrate the efficacy of our method[2]

UN-AVOIDS: Unsupervised and Nonparametric Approach for Visualizing Outliers and Invariant Detection Scoring he visualization and detection of anomalies (outliers) are of crucial importance to many fields, particularly cybersecurity. Several approaches have been proposed in these fields, yet to the best of our knowledge, none of them has fulfilled both objectives, simultaneously or cooperatively, in one coherent framework. Moreover, the visualization methods of these approaches were introduced for explaining the output of a detection algorithm, not for data exploration that facilitates a standalone visual detection. This is our point of departure in introducing UNAVOIDS, an unsupervised and nonparametric approach for both visualization (a human process) and detection (an algorithmic process) of outliers, that assigns invariant anomalous scores (normalized to [0, 1]), rather than hard binary-decision[3].

Research on Landing Vertical Acceleration Warning Mechanism based on Outlier Detection: Flight Operational Quality Assurance (FOQA) is one of the effective means of civil aviation warning and safety management. In the process of data analysis, whether it is correct outliers or wrong outliers will have an impact on the analysis results, in particular, it is necessary to eliminate the adverse effects of wrong outliers on the analysis results. The research focus of this paper is how to use the outlier detection method to achieve civil aviation warning. It mainly includes the following five parts: The first part mainly explains the related concepts and research status of civil aviation warning, and the data set used in this article, paves the way for the subsequent content; the second and third parts are the main parts of this article, mainly around " The standard deviation method and the quartile method are discussed, and the characteristics and applicable data set types of these two methods are discussed. The fourth part is the numerical experiment part, which demonstrates the shortcomings of the current civil aviation warning and the "quartile" from the numerical results. The superiority of"bit method" in civil aviation warning; the fifth part is the summary of this article and the prospect for the future[4]

LIMITATION OF EXISTING SYSTEM

• Internet: It is an important factor to which training the dataset.

EXPERIMENTAL SETUP

Python is a high-level, interpreted, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.[31][32] Guido van Rossum began working on Python in the late 1980s as a successor to the ABC programming language and first released it in 1991 as Python 0.9.0.[33] Python 2.0 was released in 2000 and introduced new features such as list comprehensions, cycle-detecting garbage collection, reference counting, and Unicode support. Python 3.0, released in 2008, was a major revision that is not completely backward-compatible with earlier versions. Python 2 was discontinued with version 2.7.18 in 2020

A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as for example height and weight of an object, for each member of the data set. Data sets can also consist of a collection of documents or files.[1] In the open data discipline, data set is the unit to measure the information released in a public open data repository. The European Open Data portal aggregates more than half a million data sets.[2] Some other issues (real-time data sources,[3] non-relational data sets, etc.) increases the difficulty to reach a consensus.

SCOPE:

This project is presenting Identification of potential outliers is important for the following reasons. An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. Outliers may be due to random variation or may indicate something scientifically interesting.

PROBLEM STATEMENT:

The term outlier, also known as anomaly is originally taken from the field of statistics. Outliers can be raised because of human error, machine error, mechanical faults and changes in the behavior of system or may be due to natural deviance in the environment.

SYSTEM ARCHITECTURE

While exception identification is the strategy for observing examples from a given arrangement of information that essentially contrasts or veers off definitely from the typical or normal of the informational collection. Exception discovery is characterized as observing examples in information that acts suddenly. Objective of anomaly discovery is to track down gadgets by their conduct that contrasts from the normal and already seen in the field of IoT. The assessment of utilized AI strategy furnishes high precision of 97.8 percent with proposed exception recognition techniques and right around 2 percent improvement in the exactness of confinement process in indoor climate subsequent to taking out anomalies.

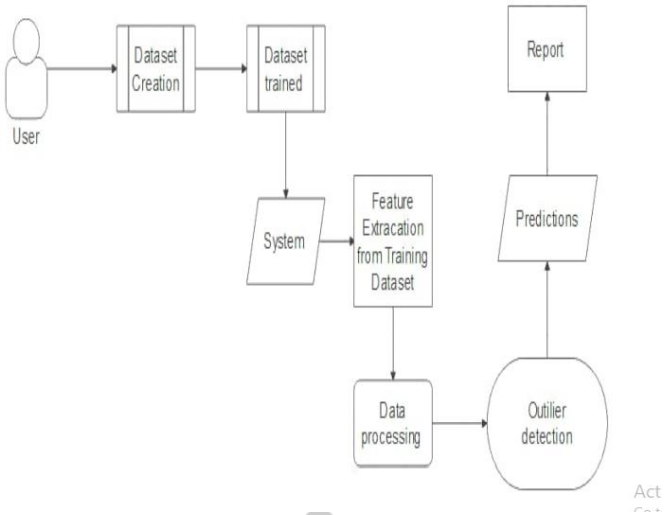

Fig -1: System Architecture Diagram

ADVANTAGES

1. To recognize patterns in dataset
2. To reduce the detection time
3. Training will provide proper accuracy
4. Detecting a Outlier automatically
5. Easy to system
6. Provide better solution in Low Cost

APPLICATION:

1. Personal
2. Forensic Department
3. Police Department
4. Investigation department.

MATHEMATICAL MODEL

S = (I,O,F)
Where,
S: System
I= {UL , DSU} are set of Input
Where,
• UL : User Login
• DSU : Data Set Upload
F = {A, PU} are set of Function
Where,
• A: Authentication
• PU: Processing
O = {N, OD, R} are set of Output
Where,
• N: Notification
• OD: Outlier Detection
• R: Report
Success Condition : Proper database, File upload perfect.
Failure Condition : No Database, No Internet Connection

METHODOLOGY

The algorithm in which every operation is uniquely defined is called deterministic algorithms. The algorithm in which every operation may not have unique result, rather there can be specified set of possibilities for every operation, such algorithms are called Non deterministic algorithms. Non deterministic means no particular rule is followed to make guess. Problem Solving Methods are concerned with efficient realization of functionality. This is an important characteristics of Problem Solving Methods and should be deal with it explicitly.

- P Class: This group consists of all algorithms whose computing times are polynomial time that is there computing time is bounded by polynomials of small degree. Eg. insertion sort, merge sort, quick sort have polynomial computing time

- NP Class: This group consists of all algorithms whose computing time are non- deterministic polynomial time. Eg. Traveling salesman problem.

  The NP class problem can be classified into two groups:

  (a) NP Hard Problems: Normally optimization problems are NP-Hard problems. All NP complete problems are NP hard but some NP hard are not NP complete. A problem is NP hard if and only if its at least as hard as NP complete problem.

  (b) NP complete problems: Normally decision problems are NP-Complete problems. Non deterministic polynomial time complete problems. Decision Problems: Any problem having the answer either zero or one is called decision problem.

5. CONCLUSION

Hence we are overcoming the drawback of existing system , we are providing the better solution than existing system in affordable cost. We proposed a system which is use to identify the Outlier detection using CNN algorithm algorithm , which is based deep learning.

REFERENCES

[1] K. K. Almuzaini and A. Gulliver, "Range-based localization in wireless networks using densitybased outlier detection," Wireless Sensor Network, vol. 2, no. 11, p. 807, Nov. 2010.
[2] H. Aly, A. Basalamah, and M. Youssef, "Accurate and energy-efficient GPS-less outdoor localization," ACM Trans. Spatial Algorithms Syst., vol, 3, no. 2, pp. 1–31, July 2017.
[3] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," IEEE Commun. Surveys Tuts., vol, 12. no. 2, pp. 159–170, 2010.
[4] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," J. Network Computer Applicat., vol. 60, pp. 19–31, 2016.
[5] I. Cramer et al., "Detecting anomalies in device event data in the IoT," in Proc. IoTBDS, 2018, pp. 52–62
[6] P. Yang and B. Huang, "A modified density based outlier mining algorithm for large dataset," in Proc. FITME, 2008.
[7] J. F. Liu and Z. Ning, "Localization anomaly detection for wireless sensor networks," in Proc. IEEE ICIS, 2010.
[8] Y. C. Chen and J. C. Juang, "Outlier-detection-based indoor localization system for wireless sensor networks," Int. J. Navigation Observation, vol. 2012, pp. 1687–5990, 2012.
[9] S. Capkun, S. Ganeriwal, F. Anjum, and M. Srivastava, "Secure RSSbased localization in sensor networks," Technical report/Swiss Federal Institute of Technology Zurich, Department of Computer Science, vol 529, 2011.
[10] Z. Yang et al., "Detecting outlier measurements based on graph rigidity for wireless sensor network localization," IEEE Trans. Veh. Technol., vol. 62, no. 1, pp. 374–383, 2012