IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006204|

Detection of Anomalies using Local Outlier Factor and Isolation Forest algorithm

Sunny Nimani[1], Mahesh Khairnar[2], Prathamesh Khele[3], Vishal Patil[4], Prof. S.S. Banait[5]

Students, Dept. of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, India[1-4]

Assistant Professor, Dept. Dept. of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, India [5]

ABSTRACT: Data generated from smart devices or applications are in time-series format, in which information is recorded for each specific time. Anomalies in log data refer to certain patterns or points in data that deviate from average data. Anomaly detection is concerned with identifying data patterns that deviate remarkably from the expected behavior. This is an important research problem, due to its broad set of application domains, from data analysis to e-health, cybersecurity, predictive maintenance, financial fault prevention, and industrial automation. Efficiency of Local Outlier Factor Algorithm, Isolation Forest Algorithm is compared. Testing dataset is obtained from Indian Council of Medical Research (ICMR) and credit card company transactional data.

KEYWORDS: Anomaly Detection, LOF Algorithm, Isolation Forest Algorithm.

I. INTRODUCTION

Anomaly detection is the process of identifying those patterns in data which do not conform to expected behaviour. These patterns are called as anomalies or outliers. According to the definition of an outlier is: “An outlier is an observation that deviates so much from other observations (considered normal) as to arouse suspicion that it was generated by a different mechanism”. For example, an abnormal traffic pattern in a computer network could mean that an unauthorizeduser is trying to compromise a system in the network. Anomaly detection finds application in a wide variety of domains such as intrusion detection in the field of networks and security; fraud detection in the field of banking and insurance, fault detection. More generally, anomaly detection is concerned with identifying data patterns that deviate remarkably from the expected behaviour. This is crucial in the process of finding out important information about the system functioning, detecting abnormalities that are often rare or difficult to model or, otherwise, to predict. A timely identification of anomalies is crucial to tackling a number of underlying problems that, if undetected, may lead to costly consequences.Examples are: spotting stolen credit cards; preventing systems failure; or anticipating cancer occurrence.

II. LITERATURE SURVEY

A Review of Current Machine Learning Approaches for Anomaly Detection Wasim A. Ali, ManasaK. N, MalikaBendechache, Mohammed FadhelAljunaid, P. Sandhya

Easily embed time-series anomaly detection capabilities into your apps to help users identify problems quickly using artificial intelligence. Due to the advance in network technologies, the number of network users is growing rapidly, which leads to the generation of large network traffic data. This large network traffic data is prone to attacks and intrusions. Therefore, the network needs to be secured and protected bydetecting anomalies as well as to prevent intrusions into networks. Network security has gained attentionfrom researchers and network laboratories. In this

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006204|

paper, a comprehensive survey was completedto give a broad perspective of what recently has been done in the area of anomaly detection. Newly published studies in the last five years have been investigated to explore modern techniques with future opportunities. In this regard, the related literature on anomaly detection systems in network traffic has been discussed,with a variety of typical applications such as WSNs, IoT, high-performance computing, industrial control systems (ICS), and software-defined network (SDN) environments. Finally, we underlined diverse open issues to improve the detection of anomaly systems. With our lives becoming more and more digitalized, computer networks are becoming more critical and dependable services. At the same time, they become more prone to anomalies and worse—malicious attacks. This motivates researchers to propose different solutions to the overarching issue of anomaly detection in network traffic, particularly machine learning techniques, whether supervised,unsupervised or semi supervised.In this paper, we surveyedworks in the field of anomaly detection using machine learning in the last five years. First, we defined the background related to our work: (i) types of network anomalies;(ii) categories of machine learning approaches;and (iii) types of network attacks. Then, we reviewed, categorized, and discussed the papers that used machine learning techniques for anomaly detection. Furthermore, we underlined some of the open issues to improve the detection of anomalies systems.

III. METHODOLOGY

- Data will be collected securely in accordance with requirement from several sources. This process outcomes the raw data which is dependent on the type and quantity of data available and how it is stored.
- At this point the raw data will be analyzed to test initial hypotheses.
- Once collected the data needs to be 'cleaned' to prepare it for processing. This involves identifying gaps in the data, making data compatible and fixing errors in storage systems.Exploratory Data Analysis will be performed to discover meaningful attributes.
- Based on patterns and features, models will be created to fit on the dataset which is created after the raw data is preprocessed after preforming exploratory data analysis and removing unnecessary attributes.
- Machine learning models will be trained and evaluated using historical data. These will then be applied to fresh data. The algorithms contained in the models will be updated to improve performance as new data becomes available.

IV. SOFTWARE REQUIREMENT SPECIFICATION

A. FUNCTIONAL REQUIREMENTS

- To identify and remove anomalies.
- To perform exploratory data analysis.
- To classify identified anomalies as point, contextual or collective.
- To compare the efficiency of all algorithms.
- To provide visualization of outcome.

B. NON-FUNCTIONAL REQUIREMENTS

1.Performance requirement: -

- Response time: The application should respond to the user quickly as possible.
- Processing time: The application should process and predict result quickly

2.Usability: -

- People with no training and little understanding of English shall be able to use the system.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006204|

C. CONSTRAINTS

1.Operational Constraints: -

- Input dataset format should be in .csv format

2.Hardware Constraints :-

- The system meets the minimum requirement specifications.
- Processor should be based on quad core architecture or better.

3.Software Constraints :-

- All the modules required are updated to the minimum required version.

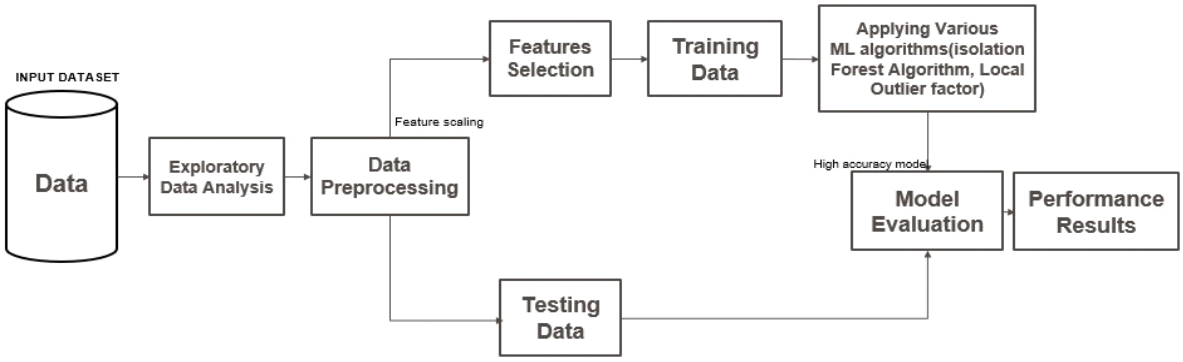
D.HARDWARE REQUIREMENTS

- 4GB Ram and minimum 2.2GHz base frequency of processor.

E.**SOFTWARE REQUIREMENTS**

- Operating System: Windows/Mac/Linux
- IDE: Jupyter Notebook.
- Programming Languages: Python3
- Libraries and Frameworks: PyCaret, Matplotlib,Pandas,Scikit-Learn

V. DETAILED DESIGN



In fig Data preprocessing is done to ensure that only needed attributes will be selected during data cleaning process. Initial analysis performed on raw data and model will be selected based on outlier detection accuracy with small sample of data and after that the process will be performed on entire dataset leading to final outcome. In above figure Initially exploratory data analysis will be performed on raw data and model will be selected based on outlier detection accuracy with small sample of data and after that the process will be performed on entire dataset leading to final outcome

Block Diagram:-

- Input Dataset: Dataset from the repository is uploaded
- Exploratory Data Analysis: PCA transformation is performed
- Data Preprocessing: Filtering or cleaning of the dataset is done

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006204|

- Feature selection: The required freatures are selected
- Training and Testing Data: Dataset is seperated in 2 parts
- Applying various ML algo.: Isolation Forest and Local Outlier Algorithm are applied on model
- Model Evaluation: Comparing the efficiency the algorithm

VI. IMPLEMENTATION

Module 1: Data Collection

Data will be collected securely in accordance with requirement from several sources.This process outcomes the raw data which is dependent on the type and quantity of data available and how it is stored.

Module 2: Examination
At this point the raw data will be analyzed to test initial hypotheses.

Module 3: Exploratory Data Analysis
Once collected the data needs to be 'cleaned' to prepare it for processing. This involves identifying gaps in the data, making data compatible and fixing errors in storage systems. Exploratory Data Analysis will be performed to discover meaningful attributes.

Module 4: Implementation Based on patterns and features, models will be created to fit on the dataset which is created after the raw data is preprocessed after preforming exploratory data analysis and removing unnecessary attributes.

Module 5: Testing Machine learning models will be trained and evaluated using historical data. These will then be applied to fresh data. The algorithms contained in the models will be updated to improve performance as new data becomes available.

- **Software Testing**

Test Case and Test Result:-

We have adopted the following strategies for evaluating Machine learning models on already collected fixed datasets and rarely explore more than accuracy, confusion matrices and F1 scores or recall, precision.25 is measured using precision, recall and F1 score. The project can also be implemented as a part of cloud services or to analyze big data for noise.

```
Isolation Forest: 73
Accuracy Score :
0.9974368877497279
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.26      0.27      0.26        49
```

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006204|

```
    accuracy                           1.00     28481
   macro avg       0.63      0.63      0.63     28481
weighted avg       1.00      1.00      1.00     28481

Local Outlier Factor: 97
Accuracy Score :
0.9965942207085425
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.02      0.02      0.02        49

    accuracy                           1.00     28481
   macro avg       0.51      0.51      0.51     28481
weighted avg       1.00      1.00      1.00     28481
```

VII. RESULTS AND DISCUSSION

• Isolation Forest detected 73 errors versus Local Outlier Factor detecting 97 errors.

• Isolation Forest has a 99.74% more accurate than LOF of 99.65%.

• When comparing error precision & recall for 2 models, the Isolation Forest performed

much better than the LOF as we can see that the detection of fraud cases is around 27 %

versus LOF detection rate of just 2 %.

• So overall Isolation Forest Method performed much better in determining the fraud cases

which is around 30%.

- **Efficiency Issue**

• Efficiency would be measured by Mean Absolute Percentage Error is the most widely used measure for checking forecast accuracy. It comes under percentage errors which are scale independent and can be used for comparing series on different scales.

• It might be difficult to estimate joint and pinpoint locations in case there is same background and costume color as well as any sorts of barrier betweenthe weightlifter and camera.

$$MAPE = mean(|e_i|/y_i) * 100$$

Where e_i is the error term and y_i is the actual data at time i.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006204|

VIII. CONCLUSION AND FUTURE WORK

With our lives becoming more and more digitalized, the value and the importance of accuracy of data becomes equally important. At the same time, they become more prone to anomalies and worse malicious attacks. This motivates researchers to propose different solutions to the overarching issue of anomaly dtection in several domains. Outliers or Anomalies in dataset can be categorized as point, contextual and cllective anomalies in all these categories across the key attributes of a dataset the performance of an algorithm can be monitored, assessed and improvements can be made.

Our project does the task of outlier detection on several datasets of several kinds using Outlier detection algorithms but it could be implemented further as a web application which would have a graphical user interface and have an option to upload the files and let the user select the attributes and the machine learning algorithm and view the results in more improved visual form. Also, it can be implemented as an API which one can easily embed for time-series anomaly detection capabilities into your apps to help users identify problems quickly. Detect spikes, dips, deviations from cyclic patterns, and trend changes through both univariate and multivariate APIs. Customize the service to detect any level of anomaly. Deploy the anomaly detection service where you need it—in the cloud or at the intelligent edge.

- **Acknowledgement**

We are overwhelmed in all humbleness and appreciation to recognize our profundity to everyone who have assisted us with putting together our preliminary project report on **'APPROACHES FOR DETECTION OF ANOMALIES'.** We want to communicate our special thanks of appreciation to our Project Guide **Prof**. **S. S. Banait** who offered us the brilliant chance to do this awesomeventure on the topic like Detection of anomalies and several algorithms for the same, which has additionally assisted us with exploration of countless things we were earlier unaware of, and also for giving us her knowledge in the domain and constantly supporting and pushing us to do better in our project. We are truly grateful **Prof. Dr. S.S Sane** for always being up for suggesting improving measures regarding our project.

REFERENCES

[1] Adeel Hashmi and Tanvir Ahmad "FAAD: A Self-Optimizing Algorithm for Anomaly Detection.", International Arab Journal of Information Technology 2020. [2] L. Erhan, M. Ndubuaku, M. Di Mauro, W. Song, M. Chen, G. Fortino, O. Bagdasar, A. Liotta "Smart anomaly detection in sensor systems.", Information Fusion October 2020. [3] Wasim A. Ali, ManasaK. N, MalikaBendechache, Mohammed FadhelAljunaid, P. Sandhya, "A review on current machine learning approaches for anomaly detection," International J. of computational science and engineering March 2020.

[4] Mahmood Safaei , Shahla Asadi , Maha Driss, Wadii Boulila, AbdullahAlsaeedi,Hassan Chizari , Rusli Abdullah and Mitra Safaei,"A systematic literature review on outlier detection in wireless sensor network.,"MPDI Journal February 2020.

[5] Sahabul Alam Debashis De Analysing security threats in wireless sensor network ` 2020

[6] L. Erhan, M. Ndubuaku, M. Di Mauro, W. Song, M. Chen, G. Fortino,O. Bagdasar, A. Liotta Smart anomaly detection in sensor systems: A multi perspective review.

[7] Mohiuddin Ahmed , Al-Sakib Khan Pathan Deep learning for collective anomaly detection Int. J. Computational Science and Engineering,2020.

[8] https://docs.microsoft.com/en-us/azure/cognitive-services/anomaly-detector/ Micrisoft azure cloud anomaly detector.

[9] Pelumi Oluwasanya Anomaly Detection in Wireless Sensor Networks,Cornell university,2017.

[10] Murad A. Rassam ,ORCID ,Anazida Zainal and Mohd Aizaini Maarof Advancements of Data Anomaly Detection Research in Wireless Sensor Networks: A Survey and Open Issue

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details