International Journal of Innovative Research in Engineering
Volume 3, Issue 1 (January-February 2022), PP: 21-23
www.theijire.com **ISSN No: 2582-8746**

A Review on Anomaly Detection using PYOD Package

M. Guhanesvar [1], Dr. M. Marimuthu[2]
[1] *M.Sc. (integrated) Decision and computing Science, Coimbatore Institute of Technology, Coimbatore, India.*
[2] *Assistant Professor, Department of Computing, Coimbatore Institute of Technology, Coimbatore, India.*

How to cite this paper: M. Guhanesvar [1], Dr. M. Marimuthu[2], "A Review on Anomaly Detection using PYOD Package" IJIRE-V3I01-21-23.

***Abstract**: Anomaly detection (aka outlier analysis) is a step-in data mining that identifies data points, events, and/or observations that deviate from a dataset's normal behaviour. Anomalous data can show critical happenings, such as a technical difficulty, or potential opportunities, for instance a variation in consumer behaviour. Anomaly detection in high dimensional data is becoming a fundamental research problem that has various applications in the real world. However, many existing anomaly detection techniques fail to retain sufficient accuracy due to so-called "big data" characterised by high-volume, and high-velocity data generated by variety of sources. This phenomenon of having both problems together can be referred to the "curse of big dimensionality," that affect existing techniques in terms of both performance and accuracy. To address this gap and to understand the core problem, it is necessary to identify the unique challenges brought by the anomaly detection with both high dimensionality and big data problems.*

***Key Words:** Anomaly detection, PYOD, machine learning, outliers*

I. INTRODUCTION

One of the best ways to get going with anomaly detection in python is using the PyOD library package. PyOD includes more than 30 anomaly detection algorithms, from classical LOF (SIGMOD 2000) to the latest COPOD (ICDM 2020). The PyOD library follows the same syntax as scikit-learn packages. It consists of probabilistic, linear model, proximity-based, outlier ensemble and neural network-based models. There are different anomaly detection techniques in PyOD, and choosing the right one require some. Many models can be of used as experimentations in finding the right one. The authors of PyOD have created an excellent comparison of how different algorithms can react to the same problem.

II. LITERATURE REVIEW

Yue Zhao, Zain Nasrullah, and Zheng Li have produced a paper" PyOD: A Python Toolbox for Scalable Outlier Detection"[1]. In this paper they have aggregated various anomaly detection techniques used in the PyOD package. PyOD is a Python toolbox with open source for scalable outlier detection of multivariate data. Specifically, it provides access to a wide range of external detection algorithms, including outlier ensembles and the latest neural network-based methods, under a single, well-documented API designed for use by both physicians and researchers.

Archana Anandakrishnan, Senthil Kumar, Alexander Statnikov, Tanveer Faruquie and Di Xu authored "Anomaly Detection in Finance"[2] where they talk about multiple applications of financial service industry and how anomaly detection can be used in it. Anomalies often denote of something interesting, such as an unusually high demand, or something gone wrong, such as imminent failure and, as such, anomaly detection has received considerable attention in many different domains.

Irfan Ullah, Hameed Hussain, and Iftikhar Ali, Anum Liaquat author the paper" Churn Prediction in Banking System using K-Means, LOF, and CBLOF"[3] which helps in identifying those customers who are probable to stop a subscriptions, products or services, and is therefore very essential for any business. Churn predictions can be very valuable for customers retentions, as it helps in predicting customer that are at risks of send-off. The churn prediction techniques however; K-Means, Local Outlier Factors (LOF) and Cluster-Based Local Outlier Factors (CBLOF) are not used so far for this.

Raghavendra Chalapathy , Aditya Krishna Menon, and Sanjay Chawla proposed the paper "Anomaly Detection using One-Class Neural Networks"[4] where one-class neural network (OC-NN) model to detect anomalies in complex data sets. OC-NN combines the ability of deep networks to extract a continuous rich representation of data for the purpose of a single phase to create a solid envelope near normal data. The OC-NN method highlights a new foundation for the following important reason: the representation of the data in the hidden layer is driven by the OC-NN goal and thus made for ambiguous identification.

CarlosEiras-Franco,DavidMartínez-Rego,BerthaGuijarro-Berdiñas,AmparoAlonso-Betanzos, and Antonio Bahamonde authored "Large scale anomaly detection in mixed numerical and categorical input spaces"[5] where a new scalable method for anomaly detection capable of handling large datasets and high dimensionality scenarios and of dealing with data having both categorical and continuous variables. It constitutes an useful tool for an emerging problem that currently lacks capable algorithms.

Xudong Wang, Luis Miranda-Moreno, and Lijun authored "Hankel-structured Tensor Robust PCA for Multivariate Traffic Time Series Anomaly Detection"[6] which describes analysing Spatiotemporal traffic data is to identify and detect anomalous observations. The events from data with complex local and time dependencies. Robust Principal Component Analysis (RPCA) is a widely used tool for finding anomalies.

Samer Nofal, Abdullah Alfarrarjeh and Amani Abu Jabal authored "A use case of anomaly detection for identifying unusual water consumption in Jordan"[7] in which they presented a case of using unconventional discovery to identify unusual water use by consumers. Irregular water use may be due to improper water meter, water meter leakage, or leaking water pipes within the consumer's premises. They tested their hypothesis using known anomaly detection methods, namely: z-score (ZS), local outlier factor (LOF), density-based spatial clustering of applications with noise (DBSCAN), minimum covariance determinant (MCD), one - class vector support machine (OCSVM), and isolation forest(iforest)

Xiaoling Tao, Yang Peng ,Feng Zhao, SuFang Wang, and Ziyi Li authored the paper "An Improved Parallel Network Traffic Anomaly Detection Method Based on Bagging and GRU"[8] which describes network traffic has gone up and its access has become even more difficult. It poses the characteristics of the high-dimensional multivariable structure, which makes network traffic anomaly detection more and more difficult. Therefore, the paper proposes an improved approach to the acquisition of the same network traffic anomaly detection method based on Bagging and GRU (PB-GRU).

Osama Abdelrahman and Pantea Keikhosrokian authored "Assembly Line Anomaly Detection and Root Cause Analysis Using Machine Learning" [9]. The purpose of this paper is to identify anomalies and potentially the possible attributes that caused these anomalies on historical integration data for two product series. Several anomaly detection models were used; HBOS, IForest, KNN, CBLOF, OCSVM, LOF, and ABOD.

C Li, L Guo, H Gao, and Y Li authored "Similarity-Measured Isolation Forest: Anomaly Detection Method for Machine Monitoring Data"[10] where the paper describes about A rough environment or unexpected accident of data acquisition instrument can introduce some anomalies in monitoring data. Therefore, a robust method of obtaining an anomaly called similarity-measured isolation forest (SM-iForest) is proposed to obtain abnormal components and available data. The flexibility and instability of iForest were reduced during operation MMD benefits from the sliding window processing features.

Mohsin Munir, Steffen Erkel, Andreas Dengel, and Sheraz Ahmed published "Pattern-Based Contextual Anomaly Detection in HVAC Systems"[11]. This paper presents detailed anomaly detection evaluation on operational time-series data of Internet of Things (IoT) operating time series based on common home devices as well as the Heat, Air Temperature and Air (HVAC) systems directly. Due to the number of problems identified during the evaluation of widely used grade-based, mathematical, and group-based confusing diagnostic techniques, it also present a method based on confusing detection patterns in the HVAC timeline series.

Tommaso Zoppi, Andrea Ceccarelli, and Andrea Bondavalli authored "Exploring anomaly detection in systems of systems" [12].This paper talks about The loosely coupled interoperability of heterogeneous existing systems, together with the ongoing replacement of monolithic systems design with Off-The-Shelf (OTS) approaches, promotes a new architectural paradigm that is called System of Systems (SoS). In SoS's, independent and autonomous constituent systems (CSs) cooperate to achieve higher-level goals.

Zhaomin Chen, Chai Kiat Yeo, Bu Sung Lee, and Chiew Tong Lau presented "Autoencoder-based network anomaly detection"[13] which states that Anomaly detection is critical given the raft of cyber attacks in the wireless communications these days. It is thus a challenging task to determine network anomaly more accurately. In this paper,an Autoencoder-based network anomaly detection method. Autoencoder is able to capture the non-linear correlations between features so as to increase the detection accuracy. We also apply the Convolutional Autoencoder (CAE) here to perform the dimensionality reduction.

Miryam Elizabeth Villa-Pérez, Miguel Á.Álvarez-Carmona, Octavio Loyola-González, Miguel Angel Medina-Pérez, Juan Carlos Velazco-Rossell , and Kim-Kwang Raymond Choo authored "Semi-supervised anomaly detection algorithms: A comparative summary and future research directions"[14]. In this paper, the focus is on existing anomaly detection approaches, by empirically studying the performance of 29 semi-supervised anomaly detection algorithms on 95 benchmark imbalanced databases from the KEEL repository. These include well-established and commonly used classifiers (e.g., One-Class Support Vector Machine (OCSVM) and Isolation Forest) and recent proposals (e.g., BRM and XGBOD).

Nabila Ounasser, Maryem Rhanoui, Mounia Mikram, and Bouchra El Asri authored "Generative and Autoencoder Models for Large-Scale Multivariate Unsupervised Anomaly Detection"[15] were several methods that can be built on existing deep learning solutions for unsupervised anomaly detection, so that the outliers can be separated from the normal data in an efficient way. The focus is on approaches that use generative adversarial networks (GAN) and autoencoders for anomaly detection. By using these deep anomaly detection techniques, we can overcome the problem that we need to have a large-scale anomaly data in the learning phase of a detection system.

Raghavendra Chalapathy and Sanjay Chawla authored "Deep Learning for Anomaly Detection: A Survey"[16]. The aim of this survey is two-fold, firstly we present a structured and comprehensive overview of research methods in deep learning-based anomaly detection. Furthermore, we review the adoption of these methods for anomaly across various application domains and assess their effectiveness. We have grouped state-of-the-art research techniques into different categories based on the underlying assumptions and approach adopted. Within each category we outline the basic anomaly detection technique, along with its variants and present key assumptions, to differentiate between normal and anomalous behaviour.

Renjie Wu and Eamonn Keogh authored "Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress"[17] where they talk about the improvement in time-series anomaly detection and its flaws. The believe that many published comparisons of anomaly detection algorithms may be unreliable, and more importantly, much of the apparent progress in recent years may be illusionary. In addition to demonstrating these claims, with this paper introduces the UCR Time Series Anomaly Archive. We believe that this resource will perform a similar role as the UCR Time Series Classification Archive, by providing the community with a benchmark that allows meaningful comparisons between approaches and a meaningful gauge of overall progress.

Wunjun Huo, Wei Wang, and Wen Li published "Anomaly Detect: An Online Distance-Based Anomaly Detection Algorithm"[18]. Anomaly detection is a key challenge in data mining, which refers to finding patterns in data that do not conform to expected behaviour. It has a wide range of applications in many fields as diverse as finance, medicine, industry, and the Internet. In particular, intelligent operation has made great progress in recent years and has an urgent need for this technology. In this paper, we study the problem of anomaly discovery in the context of intelligent performance and discover the real need for high accuracy, outlier detection algorithms online and everywhere present in the time series database

III. DISCUSSION

These models that are used in the PYOD package. All of these models have there own significance in detection of outliers. So, all these models can be combined and used. The final result can be decided by using a voting classifier. The processing time of all the models combined seems to be less as per the data used.

IV.CONCLUSION

As anomalies play a crucial role in determining correctness of any measure. Most of the industries use try to reduce it all ways possible. These papers have not only shown us how the Anomaly detection techniques have improved so far but so finds their flaws and ways to improve upon them. There is no one such Anomaly detection model that works for everything. Every Anomaly detection model works in its own way in different applications. Anomaly detection models like twitter anomaly detection and LinkedIn Luminal are especially designed to capture anomalies in time series based data. Overall, summing it up the PYOD package consist of many anomaly detection models from Linear models to neural network based models. Anomaly detection models like twitter anomaly detection and LinkedIn Luminal are especially designed to capture anomalies in time series based data. So, model uses many models from the PYOD package in combination with twitter anomaly detection and LinkedIn Luminal to capture anomalies by using a voting classifier.

Reference

1. *"PyOD: A Python Toolbox for Scalable Outlier Detection" by Yue Zhao, Zain Nasrullah and Zheng Li in Journal of Machine Learning Research (2019)*
2. *"Anomaly Detection in Finance" by Archana Anandakrishnan, Senthil Kumar, Alexander Statnikov, Tanveer Faruquie and Di Xu in Proceedings of Machine Learning Research(2017)*
3. *"Churn Prediction in Banking System using K-Means, LOF, and CBLOF" by Irfan Ullah, Hameed Hussain, Iftikhar Ali, and Anum Liaquat in 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*
4. *"Anomaly Detection using One-Class Neural Networks" by Raghavendra Chalapathy , Aditya Krishna Menon, and Sanjay Chawla in arXiv:1802.06360 [cs.LG](2018)*
5. *"Large scale anomaly detection in mixed numerical and categorical input spaces" by CarlosEiras-Franco,DavidMartínez-Rego,BerthaGuijarro-Berdiñas,AmparoAlonso-Betanzos,and AntonioBahamonde in Information Sciences Volume 487 June(2019)*
6. *"Hankel-structured Tensor Robust PCA for Multivariate Traffic Time Series Anomaly Detection" by Xudong Wang, Luis Miranda-Moreno, and Lijun Sun in arXiv:2110.04352 [cs.LG].(2021)*
7. *"A use case of anomaly detection for identifying unusual water consumption in Jordan" by Samer Nofal, Abdullah Alfarrarjeh, and Amani Abu Jabal in Water Supply Vol 00 No 0, 1 doi: 10.2166/ws.2021.210(2021)*
8. *"An Improved Parallel Network Traffic Anomaly Detection Method Based on Bagging and GRU" by Xiaoling Tao, Yang Peng ,Feng Zhao, SuFang Wang, and Ziyi Liu in International Conference on Wireless Algorithms, Systems, and Applications(2020)*
9. *"Assembly Line Anomaly Detection and Root Cause Analysis Using Machine Learning" by Osama Abdelrahman, and Pantea Keikhosrokiani in IEEE Volume 8 (2020).*
10. *"Similarity-Measured Isolation Forest: Anomaly Detection Method for Machine Monitoring Data" by C Li, L Guo, H Gao, and Y Li in IEEE Transactions on Instrumentation Volume 70 (2021)*
11. *"Pattern-Based Contextual Anomaly Detection in HVAC Systems" by Mohsin Munir,Steffen Erkel, Andreas Dengel and Sheraz Ahmed in IEEE International Conference on Data Mining Workshops(2017)*
12. *"Exploring anomaly detection in systems of systems" by Tommaso Zoppi, Andrea Ceccarelli, and Andrea Bondavalli in SAC '17: Proceedings of the Symposium on Applied Computing (2017)*
13. *"Autoencoder-based network anomaly detection" by Zhaomin Chen, Chai Kiat Yeo, Bu Sung Lee, and Chiew Tong Lau in Wireless Telecommunications Symposium (WTS) IEEE(2018)*
14. *"Semi-supervised anomaly detection algorithms: A comparative summary and future research directions" by Miryam Elizabeth Villa-Pérez, Miguel Á.Álvarez-Carmona, Octavio Loyola-González, Miguel Angel Medina-Pérez, Juan Carlos Velazco-Rossell, and Kim-Kwang Raymond Choo in Knowledge-Based Systems, Volume 218,22 April (2021)*
15. *"Generative and Autoencoder Models for Large-Scale Multivariate Unsupervised Anomaly Detection" by Nabila Ounasser, Maryem Rhanoui, Mounia Mikram, and Bouchra El Asri in Networking, Intelligent Systems and Security pp 45-58(2021)*
16. *"Deep Learning for Anomaly Detection: A Survey" by Raghavendra Chalapathy and Sanjay Chawla in arXiv:1901.03407 (cs)(2019)*
17. *"Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress" by Renjie Wu and Eamonn Keogh in IEEE Transactions on Knowledge and Data Engineering(2021)*
18. *"Anomaly Detect: An Online Distance-Based Anomaly Detection Algorithm" by Wunjun Huo, Wei Wang and Wen Li in ICWS(2019)*