Outlier Detection on Clinical Data using Deep Learning Algorithms

Nafisa Ali[1], S Praveen[2]

[1]Student, Department of Electronics and Communication, RV College of Engineering, Bangalore, India
[2]Assistant Professor, Department of Electronics and Communication, RV College of Engineering, Bangalore, India

---***---

Abstract - *Outliers, which were formerly thought to be only noisy data in statistics, have evolved into a significant subject that is being studied in a variety of subjects and application areas. Many outlier detection algorithms have been created specifically for certain application areas, while others have been developed more broadly. Some application fields, such as research on medical data, crime data and even healthcare sector data are being studied under tight confidence. The majority of existing data mining research focuses on detecting patterns in massive datasets and then utilizing that information to make organizational decisions. In this work, we seek to synthesize a variety of outlier identification strategies into a systematic and generalized description for a given set of data along with data ingestion techniques, the whole of which is followed by data transformation using a machine learning technique. Various algorithms specifically for outlier detection such as autoencoders, KNN, MCD etc. have been used with proper parameter setting. The parameters while using different algorithms for anomaly/ outlier detection are hyper-tuned according the given data set and even compared with respect to its performance in the form of scores. Two important algorithms namely, Isolation forest and Auto encoders have also been compared and inferences have been drawn out of it. Apart from the algorithms, data ingestion part has also been taken care of and that data preprocessing part is also included.*

Key Words: Data ingestion, univariate analysis, multivariate analysis, outlier detection algorithms

1. INTRODUCTION

Exceptions are outrageous qualities that digress from different perceptions on information, they may show an inconstancy in an estimation, exploratory mistakes or an oddity. As such, an anomaly is a perception that separates from a general example on an example. Exceptions or outliers can be of two sorts: univariate and multivariate. Univariate exceptions can be discovered when taking a gander at a circulation of qualities in a solitary component space. Multivariate anomalies can be found in a n-dimensional space (of n-highlights). Taking a gander at circulations in n-dimensional spaces can be exceptionally hard for the human cerebrum, that is the reason we need to prepare a model to do it for us. Anomalies can likewise come in various flavors, contingent upon the climate: point exceptions, logical anomalies, or aggregate anomalies. Aggregate anomalies can be subsets of oddities in information, for example, a sign that may show the disclosure of new wonders. During the time spent delivering, gathering, handling and examining information, exceptions can emerge out of numerous sources and cover up in numerous measurements. Those that are not a result of a blunder are called curiosities or novelties.

2. LITERATURE REVIEW

In paper [1], different information digging procedures are depicted for the abnormality recognition that had been proposed in the previous few years. This survey will be useful to analysts for acquiring an essential understanding of different methodologies for the inconsistency identification. Although much work had been finished utilizing autonomous calculations, cross breed approaches are as a rule immensely utilized as they give better outcomes also, conquer the disadvantage of one methodology over the other. Consistently new obscure assaults are seen, and, in this way, there is a need of those methodologies that can recognize the obscure conduct in the informational collection put away, moved or adjusted. In this exploration work combination or mix of previously existing calculations are referenced that have been proposed.

In paper [2], the authors have examined various manners by which the issue of peculiarity identification has been detailed in writing and have endeavored to give an outline of the tremendous writing on different methods. For every classification of irregularity recognition procedure, they have recognized an exceptional suspicion with respect to the idea of typical and bizarre information. While applying an offered strategy to a specific area, these presumptions can be utilized as rules to evaluate the viability of the procedure around there. Preferably, an exhaustive review on irregularity location ought to permit a peruse to not just comprehend the inspiration driving utilizing a specific oddity recognition method, yet additionally give a relative investigation of different procedures. Yet, the momentum research has been done in an unstructured design, without depending on a bound together thought of oddities, which makes the work of giving a hypothetical comprehension of the oddity recognition issue troublesome.

In paper [3], a DNN-based model is proposed for distinguishing abnormalities in reconnaissance recordings. The model predicts the future video dependent on past video and yields a mistake contrasting the two, which is then kept as a limit to recognize abnormalities. A critical benefit of the proposed approach is that the element extraction measure doesn't contain any hand-made highlights. Another benefit is that it doesn't have any dataset explicit boundaries. These benefits permit the proposed way to deal with sum up well. Contrasted with other DNN based strategies, the intricacy of the proposed approach is a lot of lower, permitting it to be prepared and tried a lot quicker. Testing the proposed on the freely accessible dataset showed that it performs equivalently to different techniques in the writing, approving its exhibition.

In paper [4], the authors exhibited the achievability of managed AI procedures for oddity recognition and classification of assaults. The UNSW dataset was picked as it is the most recent and the most far reaching freely accessible dataset. Results exhibit that arbitrary woodland (RF) strategy alongside highlight choice plan can accomplish 99 percent precision with abnormality location.

In paper [5], a total of seven different categories of anomaly detection are described, each of which defines how an anomaly in a dataset is discovered. In the unlabeled dataset, all ADTs can detect anomalies. Although the classification methodology is effective in distinguishing between distinct classes, performance suffers when class labels are incorrect. Nearest neighbor is adaptable to various data formats, although having fewer neighbors leads to incorrect classification and slow testing computation. Unlike nearest neighbor clustering, which provides a faster testing calculation, false labelling occurs when there is no significant cluster for anomalies. The statistical methodology is based on data distribution and hypothesis testing. Unlike statistical techniques, information�theoretic techniques do not rely on the dataset's distribution, although the measure used has an impact on performance. Only if data is separable in high dimension is spectral approaches applicable for high dimensional datasets. The graph-based method is good for collecting and visualizing data, but it only provides pairwise connections. To identify unusual events, the AD dataset must contain the bulk of nominal instances.

3. Types of anomalies

The incidence of anomalies varies significantly. Anomalies may be classified into three categories:

- Point/Global Anomalies
- Collective Anomalies
- Contextual Anomalies

Point anomaly: A Point or Global Anomaly is defined as the ability to characterize a single instance of data as abnormal in comparison to the rest of the data. This is the most basic sort of outlier. The Point Anomaly appears as a solitary outlier in two-dimensional space in the figure below. The data may be separated into three distinct clusters. Because the highlighted data point cannot be allocated to any of the clusters, it must be considered that it is an Anomaly.

Contextual anomaly: A Contextual Anomaly occurs when an instance looks anomalous exclusively in a given context. The structure of the provided dataset determines the context.

Collective Anomaly: A Collective Anomaly is defined as a group of linked cases that may be detected as abnormal in comparison to the rest of the dataset. Individually, these cases may not be identified as outliers when compared to the remainder of the dataset, but their presence in the aggregate warrants a categorization as an outlier.

4. Approaches to identify anomalies

Anomaly Detection is a discipline that aims to find instances of a dataset that are exceptional or different from the bulk of the data. This generally refers to data that does not fit into a predetermined distribution model. The normal distribution is the most well-known distribution function, and it may be used to accurately explain the distribution of observed values in a wide range of economic and technical processes. There are also decision trees, distance/density methods, and reconstruction techniques that may be used in addition to these probabilistic methods. Methods that are supervised may also be quite useful. Model-based approaches may potentially be a viable option, especially for labelled training data. Because most technological processes are cyclical, they are represented by repeated signal patterns that may be studied using a Regression or Time Series Analysis. As a result, even little deviations from the" normal" procedure may be identified. As an example, we look at the graph below. It depicts an airborne sound level recording near to a milling machine that is in operation. A cyclic signal may be identified in many automated procedures (in this case it repeats approximately every 40 seconds). Time series models can replicate this repeated signal with great precision. This model-based technique outperforms unsupervised techniques by simply comparing prediction and recording. If the training dataset has no, a small, or a known number of outliers, this is quite beneficial. The created signal model is shown in the figure. The aberrations are replicated in this basic scenario by a hammer strike on the machine's proximity. When comparing the model with the real signal, this divergence is very visible. This method may be enhanced with other characteristics such as frequency analysis.

5. Approaches of the algorithms

The following list attempts to categories the many ways. However, this should not be a rigorous categorization, since diverse strategies use methods from several fields.

Probabilistic method: These methods are based on a set of probabilistic assumptions regarding event occurrence. The probability distribution of data points is used to assess them. Outliers are rare occurrences having an extremely low likelihood. (For example, the Robust Covariance Estimator)

Distance and density method: Methods with no parameters think about and analyze data items in the context of their surroundings. The data is judged as normal if there are enough comparable data points in the region surrounding one data point. The distance between the data points is often used to show the data's similarity. This idea is followed by the k-Nearest-Neighbor Algorithm.

Clustering method: These methods seek for related items and structures to group together. The instances are partitioned into groups in such a manner that the data within each group is as comparable as feasible, while the data of separate partitions is as distinct as feasible. Outliers are instances that cannot be allocated to any of the groups.

Reconstruction method: These approaches aim to find patterns in data in order to recreate the signal without noise. Principal Component Analysis (PCA) and Replicator Neural Networks are two well-known techniques that fall within this category (RNN).

Most techniques seek to represent the areas in the feature space that represent the process's usual behavior. Anomaly is a term used to describe data that falls outside of a predetermined range. However, numerous variables complicate this rather straightforward strategy. In fact, it is sometimes difficult to define the normal range, and the distinction between normal and abnormal behavior is not always evident. Associated probability distributions may be used to characterize the set of all occurrences of normal and abnormal behavior.

Anomalous detection systems provide an anomaly score to each data point, which serves as the foundation for future decision-making. Instances are classed as normal or anomalous based on their anomaly score and a predetermined threshold. The threshold value is a hyper-parameter that affects how sensitive the system is to abnormal situations. While modest deviations from the normal state may have a big effect on medical and must be identified as anomalies, in the production environment, too much sensitivity is unsuitable owing to the large number of random disturbances. Large noise components in the data, which frequently resemble abnormalities and contribute to misclassifications, are caused by stochastic impacts in the environment of the examined region.

The basic aim is to find the most cost-effective threshold. An undiscovered abnormality, for example, might result in equipment failure and maintenance effort. On the other hand, a system with a high False-Positive Rate requires a significant amount of operator control effort. The literature often utilizes a healthcare example to demonstrate this trade-off. A low threshold is useful for cancer detection testing. Failure to identify early-stage illness lowers the patient's chances of survival, but failure to detect the test in a healthy patient leads to more testing. As a result, a high TPR is given more weight than a low FPR.

6. Design and implementation of outlier detection model

6.1 Description of the input data and pre processing

The data provided is of a clinical study which contains multiple domains. The domains have been classified as adverse effects, lab test, medical history, comorbidities, demographics, vital signs, tumor response etc. Clinical data usually contains the patient history and his/her related test values which can be useful in many applications as such. Some of the examples are:

- Data set Adverse effects (AE) lets us know what the adverse effects are the patient have undergone after intake of any medicine or after the lab test.
- Data set Lab test (LB) lets us know about various lab results for a patient and for a test.
- Data set vital signs (VS) will be useful when we want to know about the patient's crucial symptoms which lead him/her to the disease suffering.
- Data set tumor response (RS) will give information about types of responses registered over a period of cycles for a patient.

Clinical data doesn't capture names and any personal information of a patient and each patient is given a USUBJID and along with his/ her medical details for analysis. An XML sheet will contain the details of each column of a related domain. The details also include the data type of the column and hence we go for XML parser to fix any issue in the domains related to mismatch of data.

6.2 Data Labelling

The labels attached to a data instance indicate whether it is normal or abnormal. It should be mentioned that getting reliable and representative labelled data for all sorts of actions is frequently prohibitively costly. Because labelling is generally done manually by a human expert, obtaining the labelled training data set necessitates a significant amount of work. Obtaining a labelled collection of anomalous data examples that covers all types of anomalous behavior is often more challenging than obtaining labels for regular behavior.

We replace the categorical value with a numeric value between 0 and the number of classes minus 1 in Python label encoding. We utilize (0, 1, 2, 3, and 4) if the categorical variable value has five unique classes uncommon. In one hot encoding, we generate a new column (also referred to as a dummy variable) for each category of a feature to indicate whether a given row corresponds to that category.

6.3 Analysis

After the pre analysis step, we look into the data manually to figure out various types of correlation between columns in a particular domain and even with other domains. The major findings after manual analysis were:

Univariate Analysis:

1. Variance based analysis - This is subject based anomaly where we filter a subject and its one of the tests (if we are working in lab test (Lab Results (LB)) domain) and find the variance between the lab results values. The highest variance for each subject will be termed as an anomaly.
2. Missing data analysis - This is record based anomaly where if the subject's corresponding results are missing will termed as anomaly.
3. Out of range anomaly analysis - This is record based anomaly where if the subject's results are not in a normal range specified will be termed as anomaly.
4. Percentage change-based analysis - This is subject level anomaly which is correlated with the variance-based analysis, difference being that now the test is being detected as being anomaly.

Multivariate Analysis:

For each subject (Unique Subject ID (USUBJID) - USUBJID), each VISIT, there are four - RSTEST: New Lesion, Non-Target Response, Target Response and an overall response. A general rule to come up with an overall response:

- If New Lesion=Y or Target Lesion = Progressive Disease or Non-Target Lesion = Progressive disease, then overall response should be progressive disease.
- If none of the above criteria meets, then overall response should not be progressive disease.

Similarly, there are other rules that determine whether Overall response will be Stable disease or Partial Response.

We formulate our manual analysis in the form of code.

1. Mark the data points which are null in nature. These can be termed as anomalous as they can be termed as missing data.
2. Develop a code to find the variance of each lab test (or any other test) of a subject and make a separate column for it on which the outlier detector algorithm will work.

3. Develop a code to find the results value lying under a certain specified range.
4. Find the percentage change in the values of each test of a subject i.e. records and the highest one should be anomalous.

6.4 Application of outlier detection algorithms

We have applied various algorithms where we run the algorithms on all the manually stipulated columns which were calculated with respect to variance, missing data etc.

The algorithms are:

- K- Nearest Neighbors (KNN) algorithm with contamination varying from 1% to 10% and n_neighbours = 5
- Histogram-based outlier detection (HBOS) algorithm with contamination varying from 1% to 10%, n_bins = 10
- Maximum covariance determinant (MCD) algorithm with contamination varying from 1% to 10%
- Isolation forest (iForest) algorithm with contamination varying from 1% to 15% and n estimators=100
- Autoencoder algorithm with contamination varying from 1% to 15% and hidden neurons changing from [3,1,3] to [8,3,8]

For multivariate analysis, we go for the domains which have multiple attributes which can affect each other or are incomplete without each other.

These algorithms would generate anomaly score and anomaly flag as an output and a general note is that the higher the anomaly score, the more chances of the data point to be anomalous.
Outliers are identified by large differences between input and reconstructed data. These algorithms would generate anomaly score and anomaly flag as an output and a general note is that the higher the anomaly score, the more chances of the data point to be anomalous.

7. Results & Discussions

7.1 Experimental results for univariate analysis

The univariate analysis has been performed on data sets in which certain anomaly occurrence have been notified by manual analysis.

Below figure depicts the results when the variance calculation was done for Lab Results (LB) data sets.

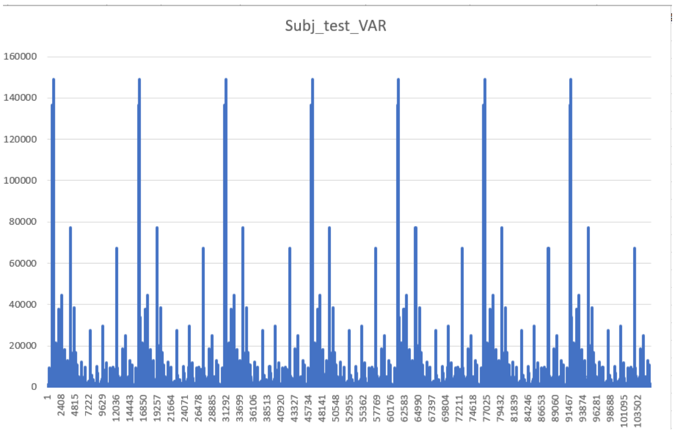

Fig -1: Variance calculation

The spikes for each subject ID indicate there are abnormalities or a huge variation between two visits/ cycles/ lab results of a subject. The higher the standard deviation, the higher the score of it being the anomaly. Threshold is a certain level above which the certain subject corresponding a test is decided by the algorithms which we use on them.

The column - Algorithm used, when filtered, will be having the options of all the algorithms used and hence we can get an insight on the different algorithms predicting the anomaly on the same dataset.

Below figure depicts the results when the variance calculation was done for Vital Signs (VS) data sets

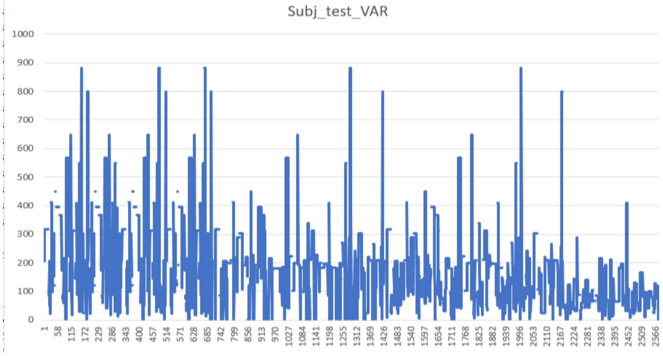

Fig-2: Variance calculation for Vital Sign (VS) data sets

The spikes for each subject ID indicate there are abnormalities or a huge variation between two visits/ cycles/ lab results of a subject. The higher the standard deviation, the higher the score of it being the anomaly.

Similar process was also conducted for all the datasets.

7.2 Experimental results for multivariate analysis

The data set used for multivariate analysis is Tumor response (RS) as it had multiple columns related to each other and feature engineering was done on top of that.

We get the output of tumor response using the Isolation Forest (Iforest) by keeping the contamination as 0.15 i.e. 15%.

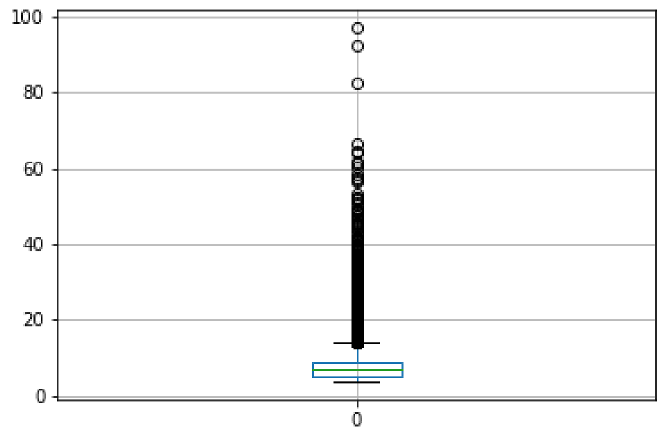

Fig-3: Box plot for Model anomaly scores

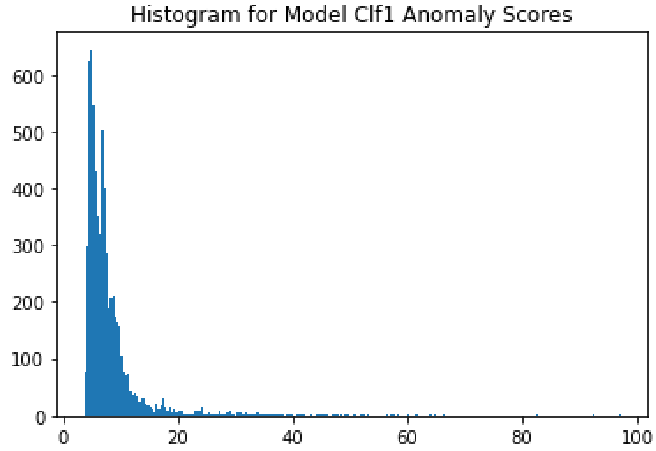

Fig-4: Histogram for Model Anomaly Scores

8. Performance Comparison between Auto Encoders and Isolation forest and inferences drawn

The tumor response and adverse effects was also tested with multiple variations in the columns, by taking few combinations from the base data and few from the featured engineered outputs. By taking into account the cross variations in the combinations, auto encoder algorithm was being compared with that of the isolation forest. The hidden neurons were the only factor which was supposed to be changed apart from the contamination factor as we are in search of the correct quantity of anomalies which is compatible with the quantity of data too.

Here are some of the histograms and box plots depicting the performance of Auto encoder predicting the threshold.

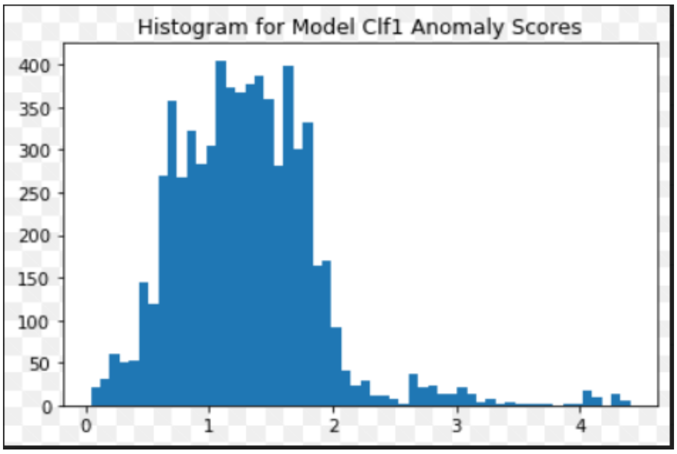

Fig-5: Histogram depiction for Autoencoder model architecture with hidden neurons [8,3,8] and contamination = 10%

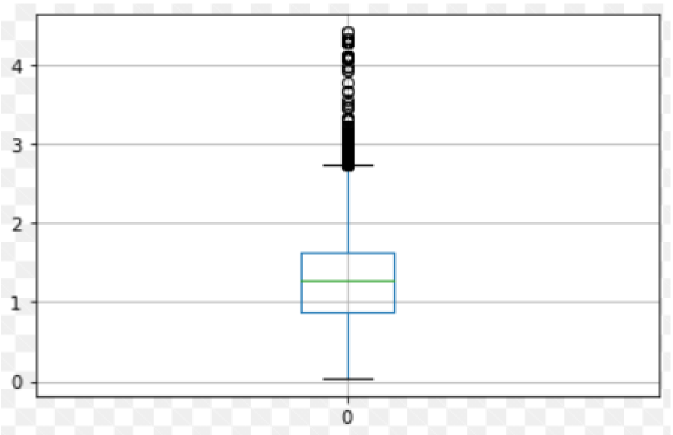

Fig-6: Box plot depiction for Autoencoder model architecture with hidden neurons [8,3,8] and contamination = 10%

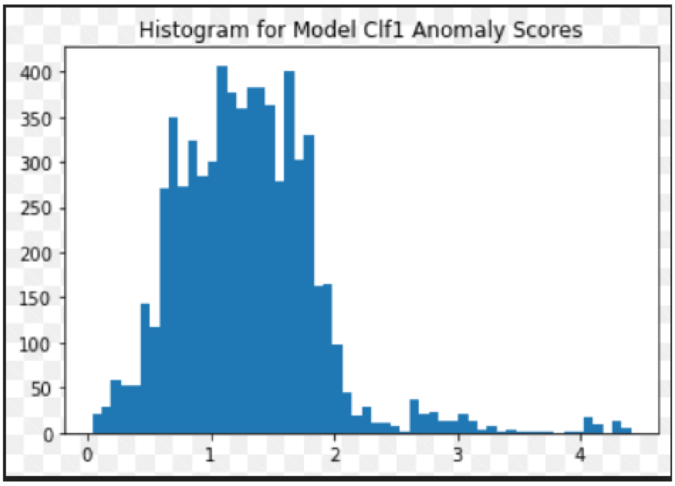

Fig-7: Histogram depiction for Autoencoder model architecture with hidden neurons [8,3,8] and contamination = 15%

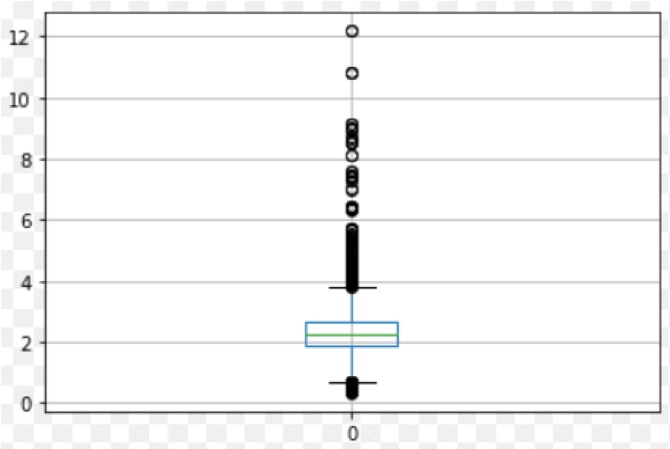

Fig-8: Box plot depiction for Autoencoder model architecture with hidden neurons [8,3,8] and contamination = 15%

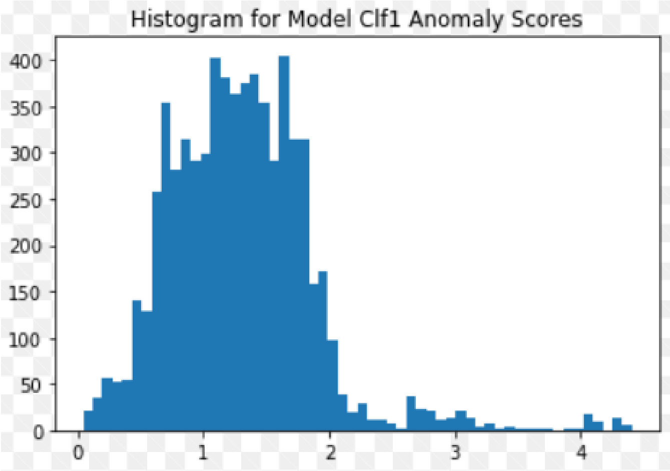

Fig-9: Histogram depiction for Autoencoder model architecture with hidden neurons [3,1,3] and contamination = 10

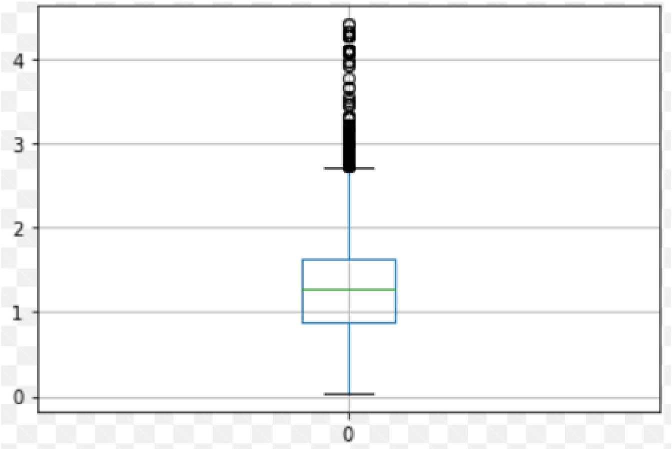

Fig-10: Box plot depiction for Autoencoder model architecture with hidden neurons [8,3,8] and contamination = 15%

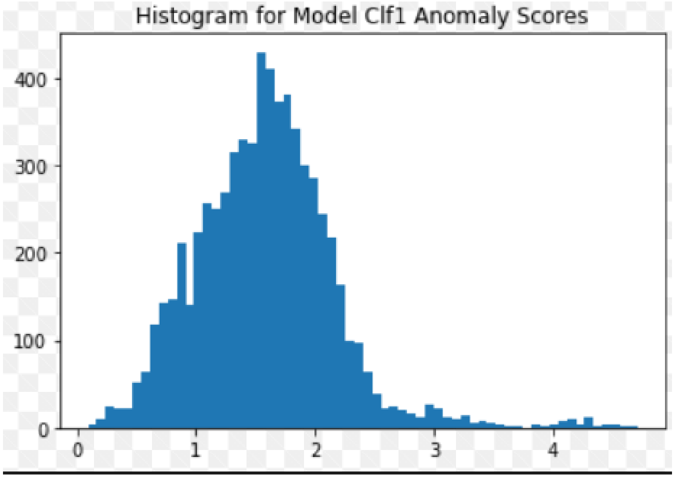

Fig-11: Histogram depiction for Isolation forest model architecture with contamination = 10%

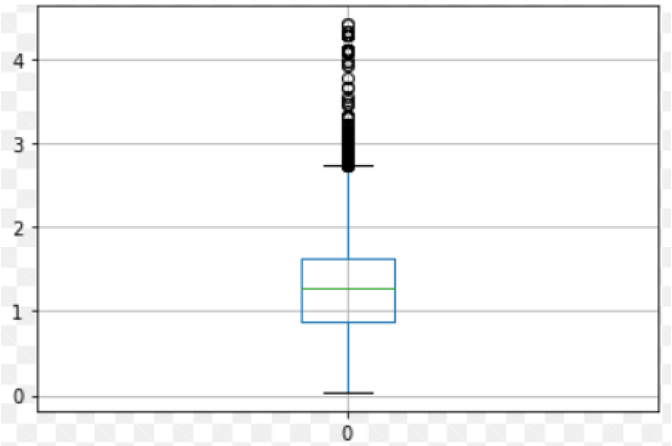

Fig-12: Box plot depiction for Isolation forest model architecture with contamination = 10%

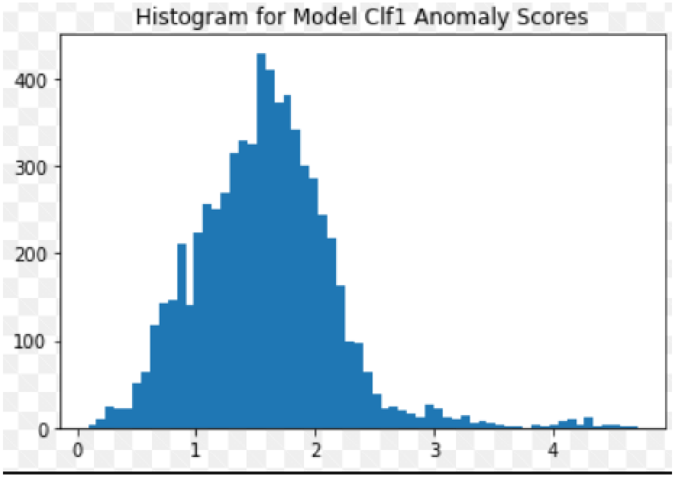

Fig-17: Histogram depiction for Isolation forest model architecture with contamination = 15%

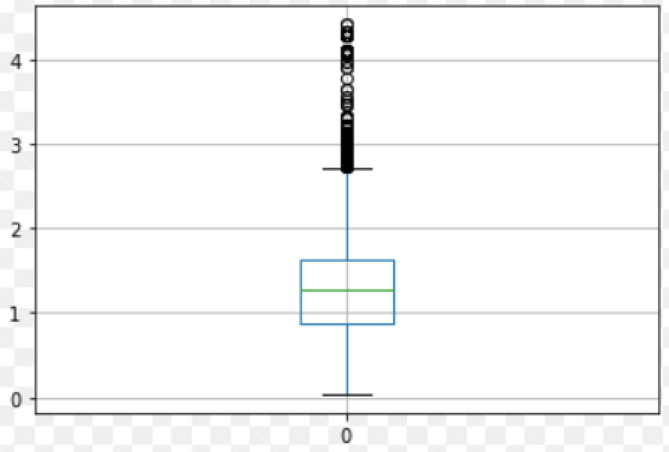

Fig-13: Box plot depiction for Isolation forest model architecture with contamination = 15%

With the variations in the contamination factor in iForest and variations of contamination factor and hidden neurons, we notice that with the same contamination factor, auto encoder performs with accuracy greater than 90% if working on large dataset. Isolation forest on the other hand needs more features and less base columns to find a pattern to establish and therefore have a greater control over finding the anomalies with the developed pattern.

10. Conclusion

In this work, various types of outlier detection algorithms have been tested on the preprocessed data-set and hence, using the score and by the histogram, we conclude the anomalous points in the data. The results of an experimental examination of various standard outlier identification approaches were provided in this work. To begin, we evaluate outlier identification strategies used in statistical approaches: linear regression and deep learning methods. The experimental findings show that when it comes to detecting outlier data, the deep learning methodology outperforms the liner regression methodology. We tried our hands on with hyper tuning different parameters with the help of grid search methods and the best parameters were taken into consideration to find the right set of outliers. Also, in this work, we provide a systematic and generalized explanation of multiple outlier identification strategies. We have a better awareness of the numerous paths of study on outlier analysis as a result of this work.

11. Future scope

Despite advancements in outlier identification research, there are still many outstanding research topics and challenges to be addressed, according to our review. In most outlier detection-based systems, further research is required. As a result, in addition to the previously mentioned future work in each category, the following are still needed to fill research gaps:

- More research is needed to completely describe and link some of these approaches to real-world data, especially in extremely large and high-dimensional databases, where first-hand approaches for predicting data densities are useful. The curse of dimensionality and distance concentration are still unresolved issues to be solved in high-dimensional data sets.
- We discovered that the parameter K is sensitive in distance-based approaches based on KNN, therefore establishing the value k is highly important. For high dimensional data, addressing the equidistance problem and creating effective distance measures is required. The closest neighbors picked for the models are sensitive to the neighbor-based OD methods. As a result, further research may be done to identify the exact number of neighbors required.
- Apart from building more robust algorithms for finding outliers more effectively, we noted that no study has been done to examine the effects of parametric and non-parametric techniques in the outlier identification process, to the best of our knowledge.

REFERENCES

[1] F. Angiulli, S. Basta, and C. Pizzuti, “Distance-based detection and prediction of outliers,” IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 2, pp. 145–160, 2006. doi: 10.1109/TKDE.2006.29.

[2] H. Wang, M. J. Bah, and M. Hammad, “Progress in outlier detection techniques: A survey,” IEEE Access, vol. 7, pp. 107 964–108 000, 2019. doi: 10.1109/ACCESS. 2019.2932769.

[3] S. B. Wankhede, “Anomaly detection using machine learning techniques,” in 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1–3. doi: 10.1109/I2CT45611.2019.9033532.

[4] G. R. Jidiga and P. Sammulal, “Anomaly detection using machine learning with a case study,” in 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, 2014, pp. 1060–1065. doi: 10.1109/ICACCCT. 2014.7019260.

[5] Y. Wang, B. Xue, L. Wang, H.-C. Li, L.-C. Lee, C. Yu, M. Song, S. Li, and C.-I. Chang, “Iterative anomaly detection,” in 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2017, pp. 586–589. doi: 10.1109/IGARSS. 2017.8127021.

[6] K. Zhang, X. Kang, and S. Li, “Isolation forest for anomaly detection in hyperspectral images,” in IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, 2019, pp. 437–440. doi: 10.1109/IGARSS.2019.8899812.

[7] Y. Qin and Y. Lou, “Hydrological time series anomaly pattern detection based on isolation forest,” in 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2019, pp. 1706–1710. doi: 10.1109/ ITNEC.2019.8729405.

[8] X. Chun-Hui, S. Chen, B. Cong-Xiao, and L. Xing, “Anomaly detection in network management system based on isolation forest,” in 2018 4th Annual International Conference on Network and Information Systems for Computers (ICNISC), 2018, pp. 56–60. doi: 10.1109/ICNISC.2018.00019. 74RV College of Engineering®, Bengaluru - 560059

[9] H. Ma, B. Ghojogh, M. N. Samad, D. Zheng, and M. Crowley, “Isolation mon□drian forest for batch and online anomaly detection,” in 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020, pp. 3051–3058. doi: 10.1109/SMC42975.2020.9283073.

[10] S. F. Yilmaz and S. S. Kozat, “Robust anomaly detection via sequential ensemble learning,” in 2020 28th Signal Processing and Communications Applications Conference (SIU), 2020, pp. 1–4. doi: 10.1109/SIU49456.2020.9302327.

[11] S. Buschj¨ager, P.-J. Honysz, and K. Morik, “Generalized isolation forest: Some theory and more applications extended abstract,” in 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020, pp. 793–794. doi: 10.1109/DSAA49011.2020.00120.

[12] D. Xu, Y. Wang, Y. Meng, and Z. Zhang, “An improved data anomaly detection method based on isolation forest,” in 2017 10th International Symposium on Computational Intelligence and Design (ISCID), vol. 2, 2017, pp. 287–291. doi: 10.1109/ISCID.2017.202.

[13] A. Vikram and Mohana, “Anomaly detection in network traffic using unsupervised machine learning approach,” in 2020 5th International Conference on Com□munication and Electronics Systems (ICCES), 2020, pp. 476–479. doi: 10.1109/ ICCES48766.2020.9137987.

[14] H. Qu, Z. Li, and J. Wu, “Integrated learning method for anomaly detection combining klsh and isolation principles,” in 2020 IEEE Congress on Evolutionary Com□putation (CEC), 2020, pp. 1–6. doi: 10.1109/CEC48606.2020.9185626.

[15] S. Hariri, M. C. Kind, and R. J. Brunner, “Extended isolation forest,” IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, pp. 1479–1489, 2021. doi: 10.1109/TKDE.2019.2947676.

[16] Y. Huang, Y. Xue, Y. Su, and S. Han, “Hyperspectral anomaly detection based on isolation forest with band clustering,” in IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, 2020, pp. 2416–2419. doi: 10.1109/ IGARSS39084.2020.9323988. [17] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, “Anomalynet: An anomaly detection network for video surveillance,” IEEE Transactions on Informa□ion Forensics and Security, vol. 14, no. 10, pp. 2537–2550, 2019. doi: 10.1109/ TIFS.2019.2900907.

[18] A. Toshniwal, K. Mahesh, and R. Jayashree, “Overview of anomaly detection techniques in machine learning,” in 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020, pp. 808–815. doi: 10.1109/I-SMAC49090.2020.9243329.

[19] Z. Ding, Y. Mo, and Z. Pan, “A novel software defect prediction method based on isolation forest,” in 2019 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE), 2019, pp. 882–887. doi: 10. 1109/QR2MSE46217.2019.9021215.

[20] A. Toshniwal, K. Mahesh, and R. Jayashree, “Overview of anomaly detection tech￾niques in machine learning,” in 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020, pp. 808–815. doi: 10.1109/I-SMAC49090.2020.9243329.

[21] A. Alazizi, A. Habrard, F. Jacquenet, L. He-Guelton, F. Obl´e, and W. Siblini, “Anomaly detection, consider your dataset first an illustration on fraud detection,” in 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), 2019, pp. 1351–1355. doi: 10.1109/ICTAI.2019.00188.