Anomaly Perception Data and Computation for Further Utilization

V. Rama Krishna Reddy[1], *Abhinav Reddy Vatte*[2], *A. Siva Kiran*[3], *K.Indrasena Reddy* [4], *A. Arul Prakash* [5*]

Department Of Computer Science and Engineering, Bharath Institute of Science & Technology affiliated to Bharath Institute of Higher Education and Research, Chennai, Tamilnadu, India.

*Corresponding authors mail id: arulpraksh.cse@bharathuniv.ac.in

ABSTRACT:

The outlier detection technique has recently received a lot of attention, and it has a lot of applications. Outlier detection methods have been requested for use in a variety of applications, including credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, athlete performance analysis, and other data-mining tasks. Outlier detection is an essential task in many safety-critical contexts because an outlier indicates aberrant running conditions that may result in considerable performance loss, such as an aviation engine rotation fault or a flow problem in a pipeline. An outlier is a strange object in a photograph, such as a land mine. An outlier may identify a malicious intrusion within a system, therefore early detection is critical. This study examines the research on recognizing outliers in the automobile sector, which reveals anomalous data and can be used effectively for production purposes. Developers should select an outlier detection technique that is appropriate for their data collection in terms of the right distribution model, the right attribute types, scalability, speed, and any incremental capabilities to allow additional exemplars to be saved. Our project proposal can cost-effectively identify the outliers in large-scale datasets from numerous data views with less computational complexity. Experiments conducted various data from various reporters' and a synthesis a processed data from the deep analysis and predicting the acquired data for further acknowledgement.

Keywords: Automotive Industry, Inspection team, supervised learning, LOF (Local outlier factor), Nodes, and Data.

1. INTRODUCTION

Anomaly detection, also called outlier detection, is the identification of unexpected events, observations, or items that differ significantly from the norm [1]. Often applied to unlabeled data by data scientists in a process called unsupervised anomaly detection, any type of anomaly detection rests upon two basic assumptions: 1)Anomalies in data occur only very rarely. 2) The features of data anomalies are significantly different from those of normal instances. Typically, anomalous data is linked to some sort of problem or rare event such as hacking, bank fraud, malfunctioning equipment, structural defects / infrastructure failures, or textual errors [2]. For this reason, identifying actual anomalies rather than false positives or data noise is essential from a business perspective.

Anomaly detection is the identification of rare events, items, or observations which are suspicious because they differ significantly from standard behaviors or patterns [3]. Anomalies in data are also called standard deviations, outliers, noise, novelties, and exceptions.

In the network anomaly detection/network intrusion and abuse detection context, interesting events are often not rare just unusual [4]. For example, unexpected jumps in activity are typically notable, although such a spurt in activity may fall outside many traditional statistical anomaly detection techniques [5].

Many outlier detection methods, especially unsupervised techniques, do not detect this kind of sudden jump in activity as an outlier or rare object [6]. However, these types of micro clusters can often be identified more readily by a cluster analysis algorithm [7].

There are three main classes of anomaly detection techniques: unsupervised, semi-supervised, and supervised [8]. Essentially, the correct anomaly detection method depends on the available labels in the dataset.

Supervised anomaly detection techniques demand a data set with a complete set of “normal” and “abnormal” labels for a classification algorithm to work with. This kind of technique also involves training the classifier [9]. This is similar to traditional pattern recognition, except that with outlier detection there is a naturally strong imbalance between the classes. Not all statistical classification algorithms are well-suited for the inherently unbalanced nature of anomaly detection [10].

Semi-supervised anomaly detection techniques use a normal, labeled training data set to construct a model representing normal behavior [11]. They then use that model to detect anomalies by testing how likely the model is to generate any one instance encountered.

Unsupervised methods of anomaly detection detect anomalies in an unlabeled test set of data based solely on the intrinsic properties of that data [12]. The working assumption is that, as in most cases, the large majority of the instances in the data set will be normal [13]. The anomaly detection algorithm will then detect instances that appear to fit with the rest of the data set least congruently [14].

The purpose of this project is to:

- ✓ Finding the fault with in outlier through statistical method.
- ✓ Increased productivity.
- ✓ Graphical view of monitoring system which is interactive to find outlier.
- ✓ Analyse& predict recommended result.

2. METHODOLOGY

2.1. Modules:

2.1.1. Production Team
2.1.2. Manager
2.1.3. Analyse Team
2.1.4. Inspection Team

2.1.1. Production Team:

The development of an automotive industry requires the constant involvement of a production team, which is made up of highly trained and skilled individuals. The production team in the automotive industry is essential to increasing industry productivity [15]. In our project, the production team register with details and redirect to the application which the production team can send raw material details which consist of process like (Stamping, Welding, Assembly, and painting) team confirms the details of the raw materials and the production team details that’s include (Production manager name, production team email, number of team fields and number of employers currently working in respective fields and the information will be forwarded to the manager, then receives an data of the analysis of the production report from the manager, from the use of that report begins the next process of production.

2.1.2 Manager:

The manager has a crucial role to perform in an organisation managing the process inside the firm, or organisation. Initially manager logs into the application by entering the specified user name and password then initially receive the raw material information from production team and then verify the raw material details that was uploaded by production team that was the preliminary role of manager, and then manager also check the production team details for smooth purposes and then a manager also receive an inspection team details to continue the process that totally depend upon the detail report uploaded by inspection team, and then receives the final production detail from the analysis team, verifies it, and sends it to the production team [16]. According to inspection team report the manager decided to assign work to inspection team.

2.1.3. Analyse Team:

The role of analyse team in industry to collect the dataset from various tier cities and by using the collected data the analyse team will using analyse method to provide the recommended solution to production team, after the manager receives the product information from production team, the analyse team from industry collect data set from the users for the further process, following completion of the analysis process by the analysis team, the data consist of (data's from various people), various location ,and various condition , which was collected by analyse team and then it will be forwarded to the inspection team for finding and reducing outliers that is presented in collected data [17]. Once the analyse team's work has been reviewed by the inspection team, the data will then continue through the analysis process before being forwarded to the manager and then manager verifies further it move onto to the production team.

2.1.4. Inspection Team:

The role of Inspection team in the industry was to provide the accurate data, which the data was received from the analyse team the role of inspection team check whether the data was accurate or else the data should be processed correctly and then move on to the further process in that process the inspection team will find the range of outliers that are presented in the collected data. In the beginning information data received from the manager, the inspection team will start their process and, receive data from the analysis team which was collected, and evaluate the collected data using the one of the supervised learning algorithm decision tree algorithm to identify outliers and its range [18]. If outliers are identified in the data, the inspection team will identify the outlier range to classify them, and the data will accurate data will be sent to the analysis team for the data should be in accurate so, it will easy to process for the analyse team [19].

2.2. System Architecture:

Production team can send raw material details which consist of process like (Stamping, Welding, Assembly, and painting) team confirms the details of the raw materials and the production team details that are include (Production manager name, production team email, number of team fields. Manager also check the production team details for smooth purposes and then a manager also receive an inspection team details to continue the process that totally depend upon the detail report uploaded by inspection team.

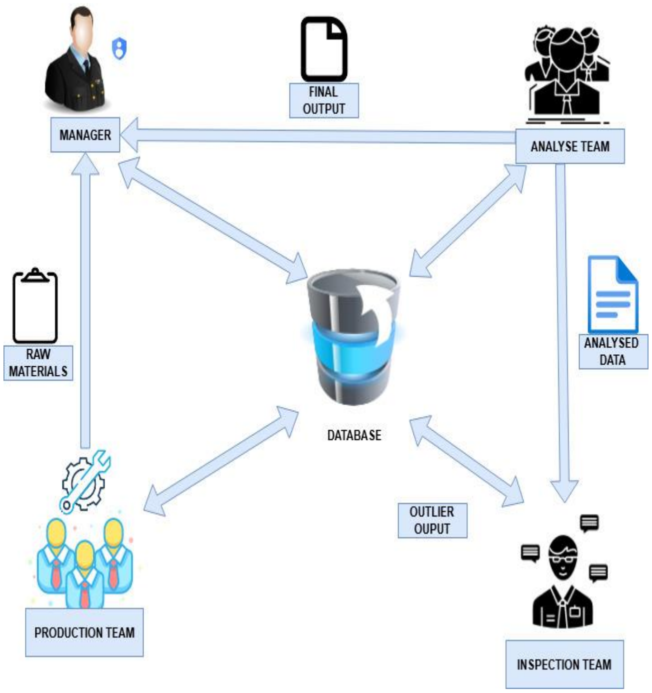

Figure 2.1: To Analyse Team's Work - Reviewed By the Inspection Team

According to inspection team report the manager decided to assign work to inspection team. Once the analyse team's work has been reviewed by the Inspection team as shown in figure 2.1. The inspection team will start their process and, receive data from the analysis team which was collected, and evaluate the collected data using the one of the supervised learning algorithm decision tree algorithm to identify outliers and its range. If outliers are identified in the data, the inspection team will identify the outlier range to classify them, and the data will accurate data will be sent to the analysis team for the data should be in accurate so, it will easy to process for the analyse team. The data will then continue through the analysis process before being forwarded to the manager and then manager verifies further it move onto to the production team.

2.3 Decision Tree Algorithm

A broad predictive modeling tool called decision tree analysis has a wide range of applications. Decision trees are often built using an algorithmic method that finds ways to divide a data set depending on several criteria. It is among the most popular and useful techniques for supervised learning. A non-parametric supervised learning technique called decision trees is utilized for both classification and regression applications. The objective is to learn straight forward decision rules derived from the data features in order to build a model that predicts the value of a target variable.

A decision tree is a graph in the shape of a tree where the nodes stand in for the places where, we choose an attribute and pose a question, the edges for the replies, and the leaves for the actual output or class label. They work with a straightforward linear decision surface when making non-linear decisions. Decision trees categorize examples by arranging them in a tree from the root to a leaf node, with the leaf node assigning the example to one of its classifications. The edges descending from each node in the tree represent one of the potential responses to the test case, and each node serves as a test case for a particular property. Recursive in nature, this procedure is repeated for each sub-tree rooted at the new nodes.

The ID3 (by Quinlan) method is the fundamental formula used in decision trees. Top-down, greedy construction of decision trees is the method used by the ID3 algorithm. The algorithm's steps are, in brief, as follows:

- Choose the best attribute, which is A.
- Designate A as the NODE's decision attribute (test case).
- Produce a new descendant of the NODE for each value of A.
- Assign the proper descendent node leaf to the training examples.
- Stop iterating over the new leaf nodes if instances are perfectly categorized; else, continue.

How to select the best attribute is now the key issue. The best attribute for ID3 is the one that has the largest information gain, which is a metric that demonstrates how well an attribute splits that data into groups based on classification.

3. RESULT AND DISCUSSION

Since ancient times, the existing outlier identification has been used to identify and, when necessary, eliminate abnormal occurrences from data. Mechanical flaws, shifts in system behavior, dishonest behavior, human error, instrument error, or merely population variances naturally occurring cause outliers. System flaws and fraud can be found via their identification before they worsen and have potentially disastrous effects. Equity or commodity traders can monitor certain shares or markets and identify fresh trends that could point to buying or selling opportunities. In previous cases Isolation Forest has been widely used after detecting the outliers for further prediction hence it consist of lot drawback as detecting local anomaly point, which affects the accuracy of algorithm we embedded with new techniques to find the outlier presents and using prediction algorithm to find exact type . How to examine outlier systems thoroughly in order to identify the algorithms that best fit a particular domain is a significant difficulty in the field of outlier systems.

In the proposed, the straightforward but crucial step of outlier analysis in data analysis. Stronger inferences can be drawn from your datasets by eliminating atypical observations, which are frequently false or wrong. Human, instrument, and natural population variations, fraud, changes in system behavior, and flaws in system design are some of the causes of outliers, Anomaly detecting algorithm like isolation forest, LOF (Local outlier factor) found in low data accuracy in our project, we used to identify outliers that were displayed in the data we had gathered. If outliers were discovered, our team would investigate and eliminate them before moving on to the next step. one of best performing clustering algorithm like Decision Tree known as supervised algorithm work by recursively partitioning the data are used then utilized to prediction after identifying the outliers from team data analysis, and supplying the range of outliers increases accuracy and aids the manufacture of automobiles, which will increase the industry's efficiency.

Advantage of Proposed System:

- Removing outliers- outliers increase the variability in your data, which decrease the Statistical power, removing outliers becomes statistically significant.
- Increase accuracy- removing outliers will increases the data accuracy improves the quality of datasets.

- Time efficient- the provided result helps solve the problem quick and efficient helps in less time consuming and provide more efficient time.
- Uncomplicated- Simple to understand and to interpret data.
- Measurement errors- will help in improve measurement errors, data entry and processing errors.

4. CONCLUSION

In this project, a general study of the performance of facing outliers is conducted one of the clustering algorithm hence many know substantial algorithm like Z-core method, isolation forest, graphical approach method are some popular method for detecting the outliers we proposed statis Since ancient times, outlier identification has been used to identify and, when necessary, eliminate abnormal occurrences from data. Mechanical flaws, shifts in system behaviour, dishonest behaviour, human error, instrument error, or merely population variances naturally occurring cause outliers. System flaws and fraud can be found via their identification before they worsen and have potentially disastrous effects. Equity or commodity traders can monitor certain shares or markets and identify fresh trends that could point to buying or selling opportunities. In previous cases Isolation Forest has been widely used after detecting the outliers for further prediction hence it consist of lot drawback as detecting local anomaly point, which affects the accuracy of algorithm we embedded with new techniques to find the outlier presents and using prediction algorithm to find exact type . How to examine outlier systems thoroughly in order to identify the algorithms that best fit a particular domain is a significant difficulty in the field of outlier systems.tical analysis method for to detect the outliers from the retrieved data likewise there are many different algorithms proposed to meet the requirement of discovering the outlier items in a large information space. Removing outliers is not our main intent the data will be more accumulate and accuracy after removed outliers will be processes for predicting the specification for that we use decision tree algorithm helps us to find the specification helps and helps to finding them providing solutions, May the data inaccuracy the result may Vary, So it is important to enhancement the data analysis and improves the finding outlier system by using various algorithm, methods and techniques and can lead to usage of detecting outlier system in various industries and for various purposes .

REFERENCES

[1] Thudumu, S., Branch, P., Jin, J. et al. A comprehensive survey of anomaly detection techniques for high dimensional big data. J Big Data 7, 42 (2020). https://doi.org/10.1186/s40537-020-00320-x.

[2] P. V. Haripriya and J. S. Anju, "An AIS based anomaly detection system," 2017 International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2017, pp. 708-711, doi: 10.1109/ICCMC.2017.8282557.

[3] S. Fan and F. Meng, "Video Prediction and Anomaly Detection Algorithm Based On Dual Discriminator," 2020 5th International Conference on Computational Intelligence and Applications (ICCIA), Beijing, China, 2020, pp. 123-127, doi: 10.1109/ICCIA49625.2020.00031.

[4] Costa, Kelton & Papa, João & Passos Júnior, Leandro & Colombo, Danilo & Del Ser, Javier & Muhammad, Khan & Albuquerque, Victor. (2020). A Critical Literature Survey and Prospects on Tampering and Anomaly Detection in Image Data. Applied Soft Computing. 10.1016/j.asoc.2020.106727.

[5] Sarker, I.H. Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. SN COMPUT. SCI. 2, 377 (2021). https://doi.org/10.1007/s42979-021-00765-8

[6] Costa, Kelton & Papa, João & Passos Júnior, Leandro & Colombo, Danilo & Del Ser, Javier & Muhammad, Khan & Albuquerque, Victor. (2020). A Critical Literature Survey and Prospects on Tampering and Anomaly Detection in Image Data. Applied Soft Computing. 10.1016/j.asoc.2020.106727.

[7] R, Rengarajan and Shekar Babu. "Anomaly Detection using User Entity Behavior Analytics and Data Visualization." 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom) (2021): 842-847.

[8] Alabe, L. W., Kea, K., Han, Y., Min, Y. J., & Kim, T. (2022). A Deep Learning Approach to Detect Anomalies in an Electric Power Steering System. Sensors (Basel, Switzerland), 22(22). https://doi.org/10.3390/s22228981

[9] DeMedeiros, K.; Hendawi, A.; Alvarez, M. A Survey of AI-Based Anomaly Detection in IoT and Sensor Networks. Sensors 2023, 23, 1352. https://doi.org/10.3390/s23031352

[10] Yasami Y., Mozaffari S. P., A novel unsupervised classification approach for network anomaly detection by k-Means clustering and ID3 decision tree learning methods; The Journal of Supercomputing; 53(1); 2010; p. 231-245.

[11] Tang D. H., Cao Z., Machine Learning-based Intrusion Detection Algorithms; Journal of Computational Information Systems; 5(6); 2009; p. 1825-1831.

[12] Chitrakar R., Chuanhe H., Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-Medoids Clustering and Naïve Bayes

[13] Classification, In Proceedings of 8th IEEE International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM); 2012; p

[14] Chitrakar R., . Chuanhe,H., Anomaly detection using Support Vector Machine classification with k-Medoids clustering; In Proceedings of IEEE Third Asian

[15] Himalayas International Conference on Internet (AH-ICI); 2012; p. 1-5.