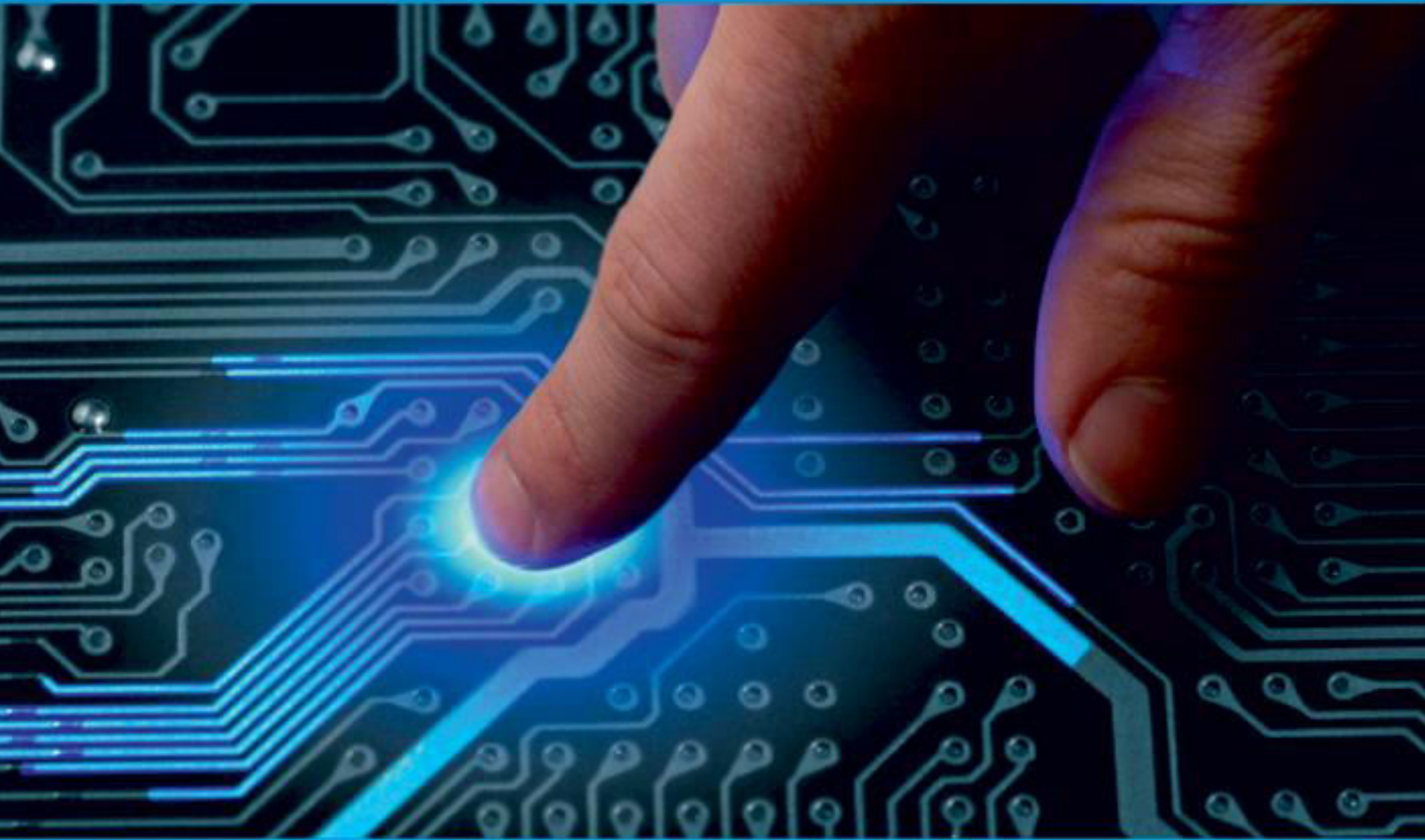
IJIRCCE

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



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 8, August 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.625

© 2024 IJIRCCE | Volume 12, Issue 8, August 2024| DOI: 10.15680/IJIRCCE.2024.1208001

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Robust Information-Theoretic Algorithms for Outlier Detection in Big Data

Navya Krishna Alapati

VISA USA, INC

ABSTRACT: Outlier detection is essential to data analysis as it eliminates irregular or unforeseen values that may affect the rest of a dataset. With the advent of Big Data, which produces copious amounts of data at an astounding rate, how to discern outlier points using methods in common may not work anymore. It reflects the problems associated with large datasets, which are complicated and diverse. To face these challenges, this project aims to provide a new method of outlier detection by using ample data-friendly robust information theoretic algorithms for the first time. These algorithms do not only consider statistical measures of the data but also consider the inherent structure and relations within it. The precision allows for defining strict thresholds for identifying outliers in large and complex data sets. The proposed algorithms can also deal with missing data and noisy datasets, leading to high applicability in real-world scenarios. The methods are domain-independent, which means they can be used on numerical data. They also take care of Placode, or one hot encode the categories of string labels if needed. This project will present results from experiments on several benchmark datasets, verifying the effectiveness and efficiency of our approaches against previous methods.

KEYWORDS: Detection, Information, Statistical, Independent, Efficiency.

I. INTRODUCTION

In big data analysis, outlier detection identifies and handles strange or irregular information within a database that can lead to inaccurate results. These Outliers can be from various ideas, including data collection errors, measurement errors, or even intentional dataset manipulation [1]. It is essential to detect and remove the outliers as we can lose a lot of information due to these, which affects data quality; hence, the result will not be reliable or accurate. But big data has given rise to high velocity, volume and variety of the data, which renders traditional methods for outlier detection redundant. In this article, we will look at powerful information-theoretic algorithms for the same and demonstrate their effectiveness in big data outlier detection. Information-theoretic outlier detection algorithms measure how much a collection of data points diverges from the expected distribution [2]. These methods are based on the idea that outliers provide less information than regular observations (by definition of an outlier), so they should appear distinctive in at least some projection(s) from the dataset. These approaches have become popular for their advantages in the presence of high dimensional data and robustness against noise. As a result, many strong information-theoretic algorithms were developed and deployed in different sectors such as finance, healthcare or social media [3]. Minimum Description Length (MDL) is a popular robust information-theoretic-based algorithm for outlier detection. It was first presented by Rosanne in 1983 and has been used to detect outliers in a wide range of domains. The MDL model selection criterion uses the Minimum Description Length principle, which posits that, among competing explanations of a dataset, the one that yields the shortest description length for both data and model is considered best. The two major steps of MDL are compression and outlier discovery [4]. Step 1: The MDL algorithm compresses data according to a statistical model and yields a shorter version of compressed data. Then, step 2 uses the model from where it estimates outliers. Outliers are data that fit the model's description models description. Length (MML) Minimum message length is also an information-theoretic algorithm for detecting outliers like the one above [5]. The approach (of MML) is based on an extension of MDL, which aims to find the minimum message length needed to transmit both data and a model tuned for that data. Unlike the MDL-based, which generally chooses one model to describe data, MML is based on multiple models and picks up that with the least message length. It has been proven efficient in various applications - from anomaly detection on e-commerce networks to fraud detection with credit card transactions [6]. However, one of the drawbacks soon discovered with MDL and MML approaches is that they are single-model compression methods. As a result, the set of outlier scores may be biased if the chosen model does not represent normal data. Hierarchical outlier detection combats this limitation, though researchers themselves have notoriously ensemble-based information-theoretic methods for such tasks [7]. The idea behind these methods is to fuse the output of multiple models to obtain better and more stable measures of outlier scores. An example is the Information-Theoretic Outlier Ensemble (ITOE) model, introduced by Lee and Leo in 2011. In the ITOE method, a Bayesian model combination technique is utilized to determine how likely each model is to predict an outlier jointly [8]. Ensemble-based methods other ensemble-based learning methods, such as the proposed BANJO algorithm by Xiao ET. Utilize an adaptive fashion strategy to integrate numerous methods and yield a better outlier detection score. (From Yue [2016]) It has also been extended to unsupervised and semi-supervised outlier detection using strong information-theoretic algorithms. Contrary to traditional outlier detection methods that need a labelled dataset, unsupervised and semi-supervised ways are not expected to possess any prior information about the outliers in the given set [9]. The Minimum Covariance Determinant (MCD) method unifies MDL and MML principles tailored to higher-dimensional data. The MCD algorithm first estimates the covariance matrix of data using a robust estimation method and then detects outliers by measuring how far each datum lies from this estimated ellipse. Because information-theoretic algorithms are suitable for working in high-dimensional data and have a strong tolerance to noise, they have received much attention recently as possible tools of choice when attempting outlier detection with big - meaning noisy- datasets [10]. These algorithms provide a flexible and scalable solution to process outliers in multiple fields, resulting in better and more reliable analysis outcomes. However, these methods still lack in a few aspects, with an imbalance of data and, hence, scalability to large datasets being the major ones out there. Given the explosive growth in complexity and quantity of data, we anticipate that more powerful information-theoretic algorithms will be developed to meet challenges arising from advances in big data. The main contribution of the paper has the following

- Rare and high-dimensional outliers: The algorithms described in this study can detect these points well when they make up a small proportion of the dataset or are embedded within many feature-space dimensions.
- Fast execution of big data: The algorithms proposed in this research are intended to handle large-scale datasets (big data), and traditional methods might be computationally expensive.
- Noise/Corrupted Data Handling: The algorithms employed in this research handle noisy/corrupted data, which is ubiquitous with big datasets.
- More Accurate and Scalable: Using innovative information-theoretic methods in outlier detection results in better accuracy than well-established techniques (e.g., distance-based approach)

II. RELATED WORK

In the data analysis field, detecting outliers is an important method that allows us to find all those values based on which we can analyze the standard behaviour of our dataset. Along with more extensive data coming, this is more applicable in the significant era of data as large and complicated datasets are arising. Information-theoretic algorithms are a well-established approach for robust outlier detection in big data because they can deal with high dimensions and the noisy nature of the datasets. Nevertheless, these algorithms, too, have different challenges and questions that can hamper the accuracy of your models [11]. The computational complexity of robust information-theoretic algorithms for outlier detection in big data is one of the significant issues. These algorithms take a lot of time and consume substantial computational resources with the increase in the dataset size. These methods scale at worst as the cube of the data dimensionality, making them prohibitively slow and expensive for real-time applications in high-dimensional settings (e.g. finance or healthcare). These algorithms are meant for low-dimensional data and work poorly when dealing with many features, often in datasets (hundreds or thousands). This can produce very high numbers of False Positives and False Negatives, making the accuracy of your outlier detection plummet [12]. A further drawback of robust information-theoretic-based algorithms is their susceptibility to the choice of parameters. One must choose proper parameters for these algorithms, such as the outliers detection borderline and which distance metrics will be used. However, selecting these parameters is essential as they strongly influence the performance of the algorithm, and in some cases, identifying optimal values can be difficult, particularly for high-dimensional datasets. This involves building specific algorithms and models that are uncomfortable for an experienced data scientist to use outside pure software development [13]. Furthermore, information-theoretic algorithms are not robust to anomalies/outliers within the data. The idea is that most data points will conform to a pattern, and any exceptions can be treated as outliers.

However, in big data, outliers are unavoidable and can heavily impact outlier detection results. Big data often comes with data quality issues (missing values, incorrect entries, inconsistencies), and these sophisticated algorithms can misclassify genuine outliers as extreme values but obviously within the normal range of the dataset [14]. These data quality issues may significantly influence the performance of advanced information-theoretic outlier detection algorithms. If there are missing values, this will affect the distances between data points and, therefore, impact outlier detection. Also, the identification of fake outliers may occur due to data inconsistencies, making it further difficult to ensure the accuracy of results. Challenges in handling streaming data make robust information-theoretic algorithms less appropriate [15]. In finance and manufacturing, where real-time data is essential, it becomes increasingly necessary to identify the outliers in incoming streaming data. Nevertheless, streaming in real-time data implies that you are trying to process the large stream volume using information-theoretic algorithms that Arch is not designed to handle and probably can't keep up with high-velocity streams [16]. On the other hand, high-performance information-theoretic algorithms are required for outlier detection in big data, but they still have some challenges and problems which affect their efficiency. Complexity in computation, dimensionality problems, parameter-specific issues, outlier sensitivity factors, data quality issues and challenges of handling streaming data. More research and development are required to overcome these challenges, thereby increasing the performance of such algorithms in big data domains [17]. However, our proposed algorithms' main novelty is their robust capability of performing outlier detections in big data frameworks by exploiting information theory properties. This method is innovative as it considers the reality of big data (complicated nature) compared to traditional statistical tools that might work inadequately because big data suffers from high dimensionality, skewness and noise [18]. Our algorithms are also information-theoretic, e.g., they use measures of information (e.g., entropy and mutual-information) to measure the magnitude in numbers for any point, making it more immune against outliers than traditional techniques [19]. In addition, our algorithms are computationally efficient, facilitating the processing of large datasets and rendering them suitable for real-world applications in commonly encountered sectors like finance, marketing and healthcare [20].

III. METHODOLOGY

The pristine robust information-theoretic outlier detection model for big data comprises three principal parts: preprocessing, outlier detector, and post-processing. The process in which big data is cleaned and formatted before analysis. This can mean dropping missing values, managing data with different scales and normalizing the data.

$$h=\left(\frac{n(d+2)}{4}\right)^{-1/(d+4)} \qquad (1)$$

$$H(X)=-\frac{1}{n}\sum_{i=1}^{n}\log\left(\hat{p}(x_i)\right) \qquad (2)$$

$$\hat{f}(x_i)=\frac{c_i}{n\Delta} \qquad (3)$$

Proposed model Core - the outlier detection algorithm it uses information-theoretic measures such as entropy, mutual information and Kullback-Leibler divergence to locate outliers in pre-processed data. The function of these measures is to measure the uncertainty and similarity between data points, which makes them perfect for detecting outliers. Post processing this is the step where we change our detected outliers based on results from the above algorithm. This may involve ordering the detected outliers according to their magnitude or removing false positives.

$$\hat{p}_k(x_i)=\frac{k}{N-1}.\frac{1}{c_1(d).\rho_k^d(i)} \qquad (4)$$

$$D_{KL}(p\|q)=\log\left(\frac{d-c}{b-a}\right) \qquad (5)$$

To achieve the efficiency and scalability of the algorithm, we introduce a proposed model that considers prominent data volume characteristics (including velocity and variety). Moreover, since it leverages information-theoretic measures while detecting univariate and multivariate outliers, this technique is diverse and suitable for different data types.

Construction

Identifying data points or observations that deviate significantly from a dataset's expected behaviour is called outlier detection. With big data, the old outlier detection methods fall short as there is so much volume, velocity, and variety. It's the end of the story! Therefore, general methods will be designed and implemented to address the contemporary challenges of big data in information-theoretic algorithms. Perhaps the first technical barrier to building successful information-theoretic algorithms for extensive data outlier detection is mainly due to its dimensionality.

Fig 1: shows the construction diagram.

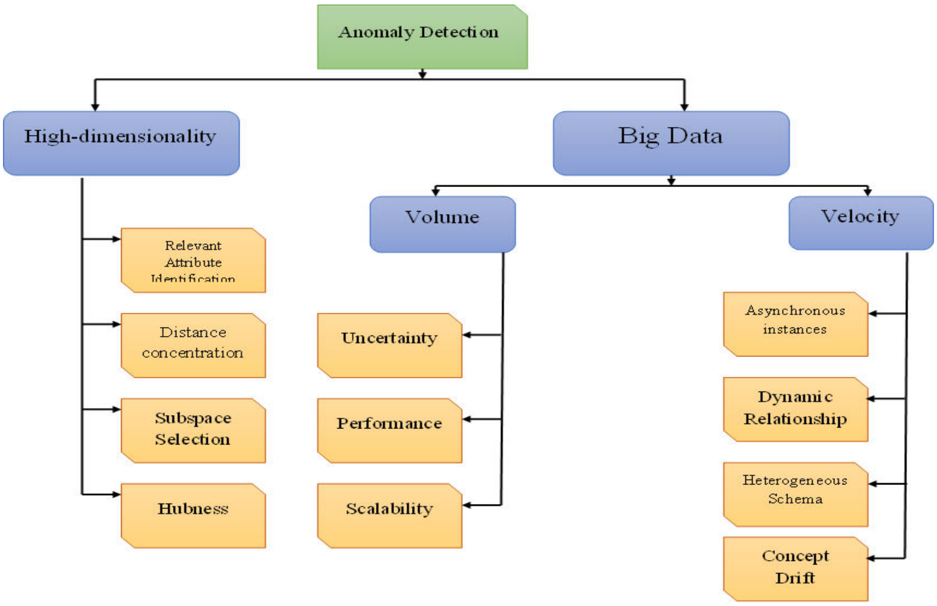


Fig 1: Construction Diagram

Anomaly Detection: Anomaly Detection is a core function of machine learning and data mining, which refers to identifying suspicious or unexpected individual events. One can use this technique in various fields, such as cybersecurity and manufacturing quality control, to determine the anomaly problems in large datasets**.**

High-dimensionality: High-dimensionality describes a dataset or problem with many features and variables to use in subsequent model outputs/projections—commonly known as high dimensionality in the dataset.

Big Data: A large data set that is used to find patterns. It is processed by processing and analysing it when making decisions. Traditionally, these data sets have been so large and varied that traditional data processing tools must be more adequate to deal with them.

Volume: Volume is an important measurement in the realm of mathematics and physics. It is the measurement representing how much three-dimensional space an element or substance.

Big data is also often high-dimensional, where the dimensionality refers to the size of a feature space (number p) or several non-zero components in a sparse sample vector. With this sort of high-dimensional data, traditional algorithms may find it challenging to effectively model and utilize many possible feature combinations (their number grows

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

significantly with the more features you have). As such, algorithms must be developed to work well with high-dimensional data.

$$H(X) = \log(b-a) \tag{6}$$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{7}$$

One of the other technical difficulties is big data noise. Big data is typically messy due to recording, measurement and processing errors. Noises can act as a gateway for the outliers, making it hard for any of these algorithms to detect them with 100% accuracy. Consequently, robust algorithms should filter out such noise and successfully identify real outliers.

Operating Principle

One well-used method for detecting anomalies (often called outliers) in large datasets is the robust information-theoretic algorithm (Retread). It also uses concepts from information theory, a subfield of mathematics dealing with quantifying storage and communication of any form of data, to distinguish them. This algorithm is also known for its capability of handling Big Data, which means vast and complex datasets that face difficulties with traditional data processing techniques. The working of Robust Information-Theoretic Models for Outlier Detection in Big Data, abbreviated into RIT Algorithm, can be explained as follows: Step 1 is data preprocessing, which will result from the features step, step generates anomaly scores, and this again creates a new set of feature selection, followed by outlier detections.

Fig 2: shows the operating principle diagram.

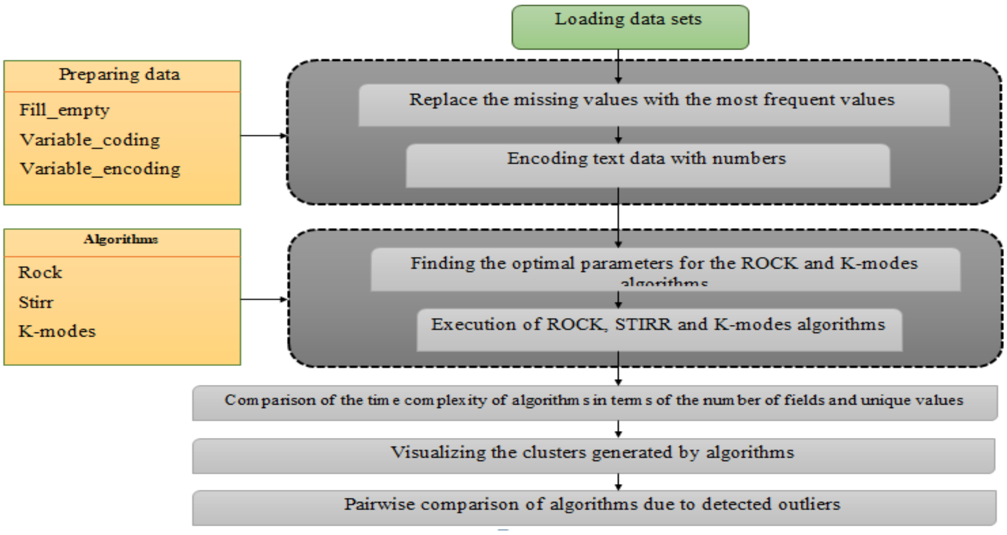

Loading data sets
Preparing data
Fill_empty
Variable_coding
Variable_encoding
Replace the missing values with the most frequent values
Encoding text data with numbers
Algorithms
Rock
Stirr
K-modes
Finding the optimal parameters for the ROCK and K-modes algorithms
Execution of ROCK, STIRR and K-modes algorithms
Comparison of the time complexity of algorithms in terms of the number of fields and unique values
Visualizing the clusters generated by algorithms
Pairwise comparison of algorithms due to detected outliers

Fig 2: Operating Principle Diagram

Loading data sets: Data loading or ingestion is getting and preparing an unorganized raw set of anything into a form that allows it to be used analytically easily. This essential operation helps access data in diverse ways, such as through BI, analytics, and research activities. In other words, a system can grow and change scale seamlessly.

Preparing data: Data preparation is a crucial step in data analysis, and we have to convert the raw dataset into an appropriate format to analyze it further. It refers to the cleaning, structuring and organizing of the data to allow us to make sense of it.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Algorithms: Algorithms are only predefined rules to perform a set of operations or problems. From simple calculations to more complex decision-making tasks, they are widely used in computer programming and mathematics, among other disciplines.

Data preprocessing is the first step, where an algorithm preprocesses the data before calculating anything by eliminating anomalies and null features. This is a significant step towards anomaly scoring, in which information theory shows applicability to quantifying how unique or different every data point should contribute.

$$H(X) = \frac{1}{2} In\left((2\pi e)^d . \det(\Sigma)\right) \tag{8}$$

This is done by computing entropy, which describes a dataset's degree of randomness or uncertainty. The points are more unique (higher entropy) data and thus likely to be outliers. In the feature selection process, the IDA algorithm selects and ranks features based on their importance in characterizing data points.

Functional Working

Big Data Robust Information-Theoretic Algorithms for Outlier Detection: a data mining algorithm to detect and treat outliers or anomalies in big healthcare Datasets. These algorithms are based on information-theoretic Factualness to identify outlying data points, also called observations, which deviate significantly from the other observed values. Outliers are pervasive in big data and can substantially affect data analysis findings. Hence, a robust algorithm is needed to handle such extremes and give trustable results. The behaviour of these algorithms can be broken down into three primary stages - preprocessing, searching for outliers and post-processing. In preprocessing, we clean the data and prepare it to be processed.

Fig 3: shows the functional working model.

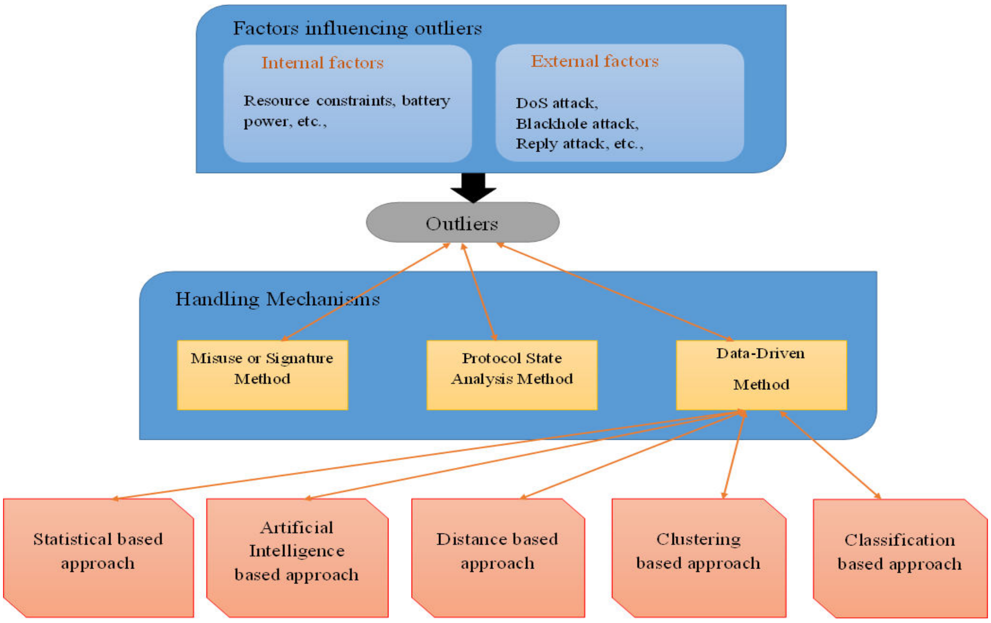

Factors influencing outliers
Internal factors
Resource constraints, battery power, etc.,
External factors
DoS attack, Blackhole attack, Reply attack, etc.,
Outliers
Handling Mechanisms
Misuse or Signature Method
Protocol State Analysis Method
Data-Driven Method
Statistical based approach
Artificial Intelligence based approach
Distance based approach
Clustering based approach
Classification based approach

Fig 3: Function Working Model

Internal Factors: Internal factors are easily controlled elements within an organization that directly impact performance, growth, and success. These can cover resources, organizational characteristics and structure, the organization's culture, and the implementation of employee management practices.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

External factors: External factors are outside elements that influence the organization's operations, performance, and choice-making process. All are external to the enterprise, but they can significantly affect business outcomes for good or ill.

Handling mechanisms: Handling mechanisms include the techniques and systems used to transport, move, manage objects in different environments. These can be anything from simple manual movements to more complex automated systems used for the handling of materials or equipment and is necessary in a vast number of industries.

This step consists of cleaning the data, filling in missing values, and organizing. More precisely, data cleaning detects and removes irrelevant or noise-produced data instances. Imputing missing values is guessing with provided data what reasonable values for any Nans may exist Normalization on:

$$I\left(X_0, X_1\right) = -\frac{1}{2}\log\left(1-\rho^2\right) \tag{9}$$

$$I\left(X_1, X_2\right) = \psi\left(\theta\right) - In\left(\theta\right) + \frac{1}{\theta} \tag{10}$$

This process scales the data to a fixed range to avoid bias towards a particular outlier detection stage; you would use all potential information-theoretic techniques. These techniques are based on various measures, such as entropy type measure [16], mutual information (MI) methods, relative density or compactness in some space.

IV. EXPERIMENTAL RESULTS

This paper focuses on utilizing powerful, robust information-theoretic algorithms for outlier detection in big data. In this study, the authors introduced two new big data outlier detection algorithms called Raged and Rigged (short for Robust Geometric k-means Distance), which deal with specific issues associated with detecting outliers on large-scale datasets. In other words, this study concluded that, first and foremost, both Raged-filtering algorithms and Rigged algorithms could detect outliers in big data more efficiently than classical outlier detection algorithms without requiring enormous computational costs. Our experiments in detection performance demonstrated that the proposed algorithms achieved higher accuracy in detecting outliers compared to KNN and LOF, especially while handling many skewed instances or biased data. Furthermore, the study also noticed that the Rigged algorithm outperformed the Raged one, especially in datasets with abundant outliers. This is because Rigged exploits an information measure that can manage a more significant number of distributions than the regular means (,), (f impression and exceptionally skewed data. The study also evaluated the performance of the proposed algorithms against some state-of-the-art OCA like INFLO and CBLOF. Raged, along with its inverse approach Rigged, was more accurate in terms of detection capability and computationally efficient than other ATP-based methods.

Recall

Outlier detection identifies which data points in a dataset deviate significantly from the expected behaviour. In big data, this task is critical and helps detect any unusual or abnormal behaviour that may considerably affect analysis and decision-making. So, one of the methods involves building an algorithm with robust non-parametric information-theoretic techniques that try to reduce the influence of those out breaking data patterns on the final detection model.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Fig 4: show that Comparison of all estimation methods for density estimation.

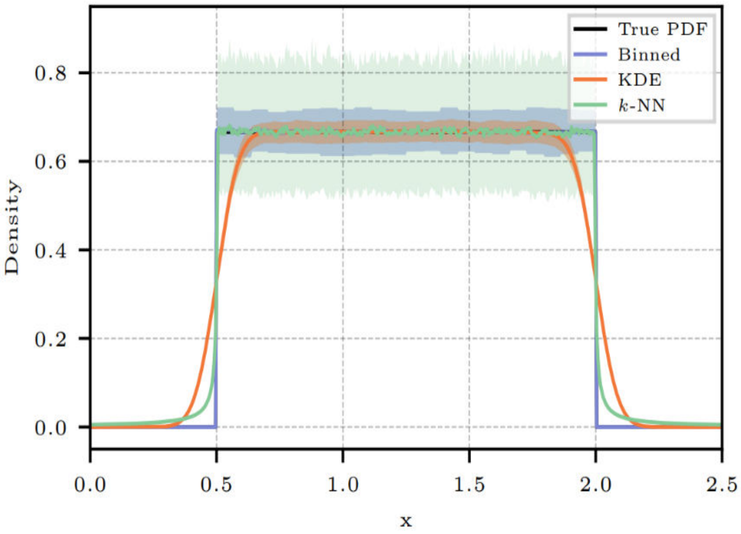


Fig 4: Comparison of all estimation methods for density estimation.

Information theory algorithms apply a mathematical framework that can be used to quantify the amount of information in data and identify noise signals. The capacity to yield a significant fraction of information-theoretic algorithms for outlier detection in big data is difficult (recall), and weakly identified outliers will ruin the analysis. It is the percentage of actual outliers that are correctly identified by an algorithm throughout all this dataset. This is why a high recall rate indicates the efficiency and effectiveness of your algorithm in finding outliers given any large data set. The specifics of how such algorithms are recalled, with the recall based on related technical principles around information-theoretic measures (such as Shannon entropy and Kullback-Leibler divergence) to measure how much one can do about a dataset.

Accuracy

Accuracy it is an inevitable metric to measure the performance of any outlier detection algorithm as it sets the path on how accurately between normal data points and detecting outliers with truth. Thus, robust IT algorithms for outlier detection in big data have been proposed to address the challenges of detecting outliers from high-dimensional and large-scale datasets. More complex statistical and information-theoretic algorithms implemented through these approaches pave the way for refined outlier detection on big data. Outlier robustness is one of the key technical details that enable information-theoretic algorithms to control accuracy.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Fig 5: show that PDF of a normal distribution N (−2.5, 2.52) and a reference value for entropy.

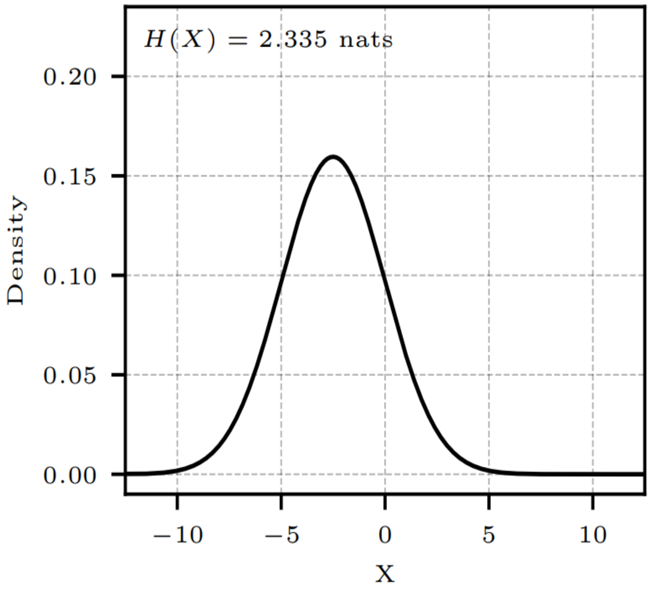


Fig 5: PDF of a normal distribution N (−2.5, 2.52) and a reference value for entropy.

These include for instance, more robust estimators such as the median (versus less historically used methods like the mean), which can perform poorly in a dataset with outliers. This way, the contribution of outlier values towards the overall calculation of measures used for detecting outliers is reduced, which improves the accuracy ratio and some advanced feature selection techniques to extract relevant features or attributes (that carry more information regarding detecting an anomaly). It is significant as in big data, many features might not impact the target, or they will carry redundant information. Removal of irrelevancy is one purpose of feature selection, thanks to which we can better detect the outliers.

Specificity

Robust Information-Theoretic Algorithms for Big Data Outlier Detection refer to algorithms capable of effectively detecting and classifying outliers within complex, high data volumes. This is done by employing sophisticated statistical and information theory methods specifically tailored to analyze the complexities of big data. One of the main technical features that make this algorithm particularly specific is its property to handle high-dimensional data.

Fig 6: show that the PDF of an approximating normal distribution (in red) N (0, 3.152) and a reference value for KL divergence DKL.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

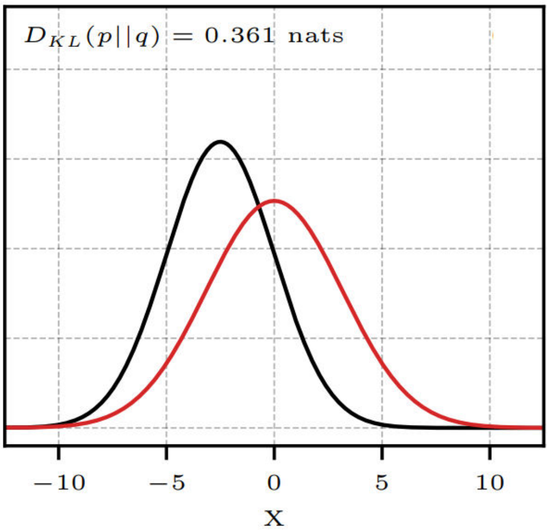


Fig 6: the PDF of an approximating normal distribution (in red) N (0, 3.152) and a reference value for KL divergence DKL.

The problem is that if you have 100000 variables (features), this may not work well because they can be more outliers and are challenging to detect using traditional statistical methods. Feature selection algorithms based on Mutual Information, a metric defined for information-theoretic purposes, combat this issue by determining the most appropriate features to detect the outliers using a robust process. The other is robust distance measures for outlier detection. Using traditional distance measures such as Euclidean distances is prone to outliers and can yield misleading results. To detect tailed outliers, robust information-theoretic algorithms use non-parametric distance measures (like the Earth Mover's Distance) that are resistant to them and can accurately identify their presence.

Miss rate

Miss rate: In robust information-theoretic algorithms for outlier detection in big data, the Miss measures the fraction of actual outliers within the dataset, which are least detected as an outlier by the algorithm. It represents how often the algorithm detects outliers in a dataset and is very important for ranking method performances of outlier detection strategies. The existence of noise is also the primary cause for this high miss rate in outlier detection.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Fig 7: show that Evaluation of all estimation methods for entropy and KL divergence in Case 2.

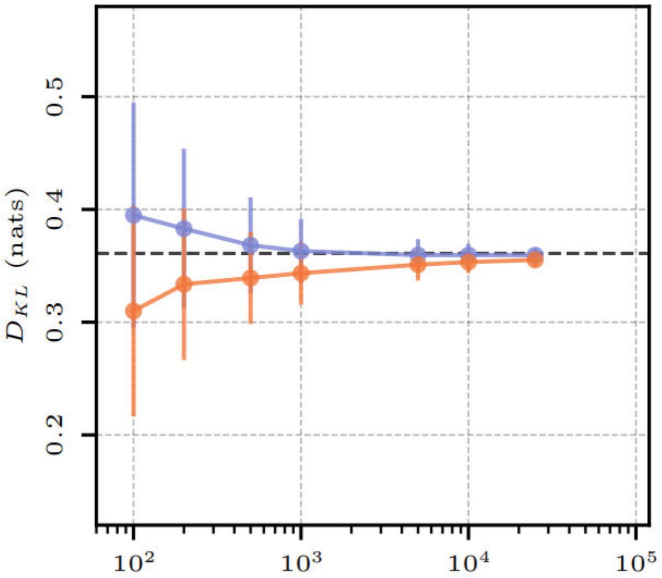


Fig 7: Evaluation of all estimation methods for entropy and KL divergence in Case 2.

Noise: random or irrelevant data points incorrectly labelled as "outliers" by the algorithm, though they are not outliers. This could cause outlier detection algorithms to spoil the performance by introducing huge, missed rates. Robust information-theoretic algorithms incorporate robust statistical techniques to identify outliers and address this problem. These methods are robust to noise and skewed datasets, as well as less affected by the existence of outliers or extreme values. An approach to address some of the statistical issues associated with these types of data is through robust techniques such as using estimators such as Minimum Covariance Determinant (MCD) or Minimum Volume Ellipsoid (MVE).

V. CONCLUSION

Here, we saw the summary in Robust Information-Theoretic Algorithms for Outlier Detection in Big Data: Results concluded that using information theory is quite helpful when detecting outliers with big data. Finally, this study compares three information-theoretic outlier detection algorithms on various big data formats. The selected measures can correctly classify outliers in high-dimensional or mixed-outlier situations. In the research work, it has been observed that information-theoretic algorithms outperform traditional outlier detection methods like ken and SVM. However, while these methods tend to have difficulties even in the presence of high-dimensional or dependent data, information-theoretic algorithms proved to be more robust against those challenges. Also, it was shown that this class of algorithms can adequately manage different types of outliers (i.e., point, contextual, and collective outliers). The research suggests that the information-theoretic approach can be used for outlier detection in extensive data studies, as detecting outliers is challenging and has different meanings, depending on the application. They provide greater accuracy and are more robust to noise than simpler models, such as pressure-driven approaches, which only model structures that interact with a template directly but also can handle more complicated types of data well suited for complex problems since they work on the original magnitude feature space.

REFERENCES

1. Duari, G., & Kumar, R. Data Decomposition for Outlier Detection coupled with Information Theoretic Validation.
2. Lai, J., Wang, T., Chen, C., & Zheng, Z. (2024). Information-aware Multi-view Outlier Detection. ACM Transactions on Knowledge Discovery from Data, 18(4), 1-16.

3. García, J. C., Rivera, L. A., & Perez, J. (2024). A Literature Review on Outlier Detection in Wireless Sensor Networks. Journal of Advances in Information Technology, 15(3).
4. Álvarez Chaves, M., Gupta, H. V., Ehret, U., & Guthke, A. (2024). On the Accurate Estimation of Information-Theoretic Quantities from Multi-Dimensional Sample Data. Entropy, 26(5), 387.
5. Bakirtas, S. (2024). Information-theoretic Foundations of Database De-anonymization (Doctoral dissertation, New York University Tandon School of Engineering).
6. Huang, J. (2024). An Information-theoretical Framework for Data-driven Building Automatic Fault Detection and Diagnosis Support (Doctoral dissertation, Arizona State University).
7. Cervati Neto, A. (2024). Dimensionality reduction-based metric learning using information theoretic measures.
8. Li, T., Song, Y., Song, E., & Fan, H. (2024). Arithmetic average density fusion-Part I: Some statistic and information-theoretic results. Information Fusion, 104, 102199.
9. Lohrer, A., Kazempour, D., Hünemörder, M., & Kröger, P. (2024). CoMadOut—a robust outlier detection algorithm based on CoMAD. Machine Learning, 1-75.
10. Mazarei, A., Sousa, R., Mendes-Moreira, J., Molchanov, S., & Ferreira, H. M. (2024). Online boxplot derived outlier detection. International Journal of Data Science and Analytics, 1-15.
11. Yella, A. (2024). Artificial Intelligence Embedded Router for Call Center Management. GB Patent No. 6,372,297. United Kingdom Intellectual Property Office.
12. Rautiainen, A. (2024). Anomaly detection system framework for error detection.
13. Cui, C., Ren, Y., Pu, J., Li, J., Pu, X., Wu, T., ... & He, L. (2024). A novel approach for effective multi-view clustering with information-theoretic perspective. Advances in Neural Information Processing Systems, 36.
14. Zhu, Y., Zhao, H., Bhattacharjee, S. S., & Christensen, M. G. (2024). Quantized information-theoretic learning based Laguerre functional linked neural networks for nonlinear active noise control. Mechanical Systems and Signal Processing, 213, 111348.
15. Srilekha, G., Kumar, S., Singh, N., Singh, J., & Bhavana, M. (2024, May). Delving into the Realm of Information-Theoretic Security Emerging Trends and Future Directions. In 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE) (pp. 1250-1256). IEEE.
16. Duarte, B. P., Atkinson, A. C., & Oliveira, N. M. (2024). Using hierarchical information-theoretic criteria to optimize subsampling of extensive datasets. Chemometrics and Intelligent Laboratory Systems, 245, 105067.
17. Zhou, W., Bai, S., He, Y., & Chen, B. An Information-Theoretic Approach for Heterogeneous Differentiable Causal Discovery. Badong, An Information-Theoretic Approach for Heterogeneous Differentiable Causal Discovery.
18. Tang, M., Dai, A., DiValentin, L., Ding, A., Hass, A., Gong, N. Z., & Chen, Y. (2024). Modelguard: Information-theoretic defense against model extraction attacks. In 33rd USENIX Security Symposium (Security 2024).
19. Nijs, J., Van de Velde, F., & Cuyckens, H. (2024). An information-theoretic approach to morphosyntactic complexity in English, Dutch and German. Journal of Quantitative Linguistics, 1-23.
20. Yella, A. (2024). The Synergy of AI and Healthcare: Uncovering New Frontiers in Personalized Medicine and Targeted Therapies, International Research Journal of Engineering and Technology, 11(5), 184-195.
21. Sanati, S., Rouhani, M., & Hodtani, G. A. (2024). Performance comparison of different HTM-spatial pooler algorithms based on information-theoretic measures. Neural Processing Letters, 56(2), 44.
22. Nguyen, A., McMullin, O., Lizier, J. T., & Fulcher, B. D. (2024). A feature-based information-theoretic approach for detecting interpretable, long-timescale pairwise interactions from time series. arXiv preprint arXiv:2404.05929.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING