Vol.15 No.3(2024),144-153
DOI: https://doi.org/10.61841/turcomat.v15i3.14785

Comprehensive Data Corruption Identification Using Machine Learning Algorithms (PAACDA)

Dr. M. Vanitha [1], K. Maneesha [2], K. Uma Renu Sri [3], K. Nancy[4]

[1] Professor, Department of CSE, Malla Reddy Engineering College for Women, Autonomous, Hyderabad,

[2],[3],[4]Student, Department of CSE, Malla Reddy Engineering College for Women, Autonomous, Hyderabad

ABSTRACT:

Data and analysis have evolved from being scattered numbers and qualities in spreadsheets to being seen as a means to revolutionize any substantial industry, thanks to the rise of technology. There are many unethical and unlawful ways that data may get corrupted, thus it's important to find a way to effectively detect and highlight all the corrupted data in the dataset. It is not an easy task to detect damaged data or to restore information from a corrupted dataset. This is crucial and could cause issues with data processing using machines or deep learning methods later on if not handled early enough. Rather than focusing on outlier identification, this study introduces its PAACDA: Presence driven Adamic Adar Corruption identification Algorithm and then consolidates the findings. Even though they rely on parameter tuning to achieve high accuracy and remember, current state-of-the-art models like Isolation forest and DBSCAN (which stands for "Density-Based the spatial the process of clustering of the applications with Noise") have a lot of uncertainty when they factor in corrupted data. This study investigates the specific performance problems with several unsupervised learning methods on corrupted linear and clustered datasets. In addition, we provide a new PAACDA technique that achieves a higher precision of 96.35% for cluster data and 99.04% for linear data compared to previous unsupervised training benchmarks on 15 prominent baselines, including as K-means clustering, Isolation forest, and LOF (Local Outlier Factor). From the aforementioned angles, this essay delves deeply into the relevant literature as well. In this study, we identify all the problems with current methods and suggest ways forward for research in this area.

INTRODUCTION

Vol.15 No.3(2024),144-153
DOI: https://doi.org/10.61841/turcomat.v15i3.14785

Nearly every aspect of human experience has seen tremendous advancement in technology from the birth of our species. Huge volumes of data are largely responsible for the continuous improvement in technology; without them, much of this sector would grind to a halt. Businesses with more data appear to have a stranglehold on the market since data has become so crucial. In many cases, data is the building block that allows the algorithm or technology to flourish and eventually reach its full potential. Data is essential to any firm, and protecting it from harmful manipulation has become more important in the current world. The effects of changing only a tiny subset of a dataset may be enormous. In spite of the fact that there are a lot of unethical methods to corrupt data, people have worked tirelessly throughout history to find effective techniques to learn about data abuse, such as... Prior to devising a novel strategy for determining VOLUME 10, 2022 1 the editors of IEEE Access have decided to publish this piece. Before final publication, material may vary from this author's version, which was not completely edited. This work is referenced as DOI 10.1109/ACCESS.2023.3253022. The Creative Commons Attribution-Non-commercial-No Derivatives 4.0 License governs this piece of art. Looking for additional information about flying data corruption? Check out https://creativecommons.org/licenses/by-nc-nd/4.0/. We also thoroughly explored various approaches that are already available for detecting data corruption, especially when it comes to outliers. After extensive research, we learned that different algorithms perform differently when evaluated on a dataset with damaged data instead of outliers, and this information shed light on the topic. An unsupervised method for dealing with classification and grouping problems, K-means clustering makes use of clusters and their centroids. As it detects primary samples of high density and builds upon them, DBSCAN—another clustering-based technique—tends to perform better with data containing clusters of comparable density. Both of these approaches produced adequate results when used to detect outliers in the given dataset. We began to delve into methods like histogram-based outlier identification, elliptic envelope outlier detection, and isolation forests as our study progressed. Isolation forest gives us a method to divide the characteristics of the dataset into subsets and find the ones that go above the specified range of values. Any points beyond the elliptic envelope model's bounds indicate outliers in the dataset, and the model tends to construct an ellipse around the dataset's scatter plot. Histogram oriented algorithm for outlier identification (HBOD) is another successful unsupervised

Vol.15 No.3(2024),144-153
DOI: https://doi.org/10.61841/turcomat.v15i3.14785

technique for detecting abnormalities; it too involves plotting and analysing histograms. When it comes to finding outliers in the dataset, these methods are also quite accurate. In order to compile a table showing the accuracy level of outlier prediction for the synthetic dataset, algorithms including "Principal Component Analysis" (PCA), "Deep SVDD," and "Rotation based Outlier Detection" (ROD) were also examined. The following methods were utilized: principal component analysis (PCA), regression on data (ROD), deep support vector machine (SVDD), and many more. Despite the different approaches offered by different models, the novel approaches suggested in this study performed very well on the following metrics: F1 score, Accuracy, Recall, Precision, and Sensitivity. Data corruption detection is an area where Adamic Adar's potential is growing, since it is an algorithm that shows promise regarding information correlation in graph networks. It is possible to prevent the inefficiencies of the existing job, according to the research cited above. By capitalizing on the Adamic Adar algorithm's popularity in data correlation, this work aims to integrate the existing work in this area of research and improve the accuracy of the current corrupt detection method. The study's innovative approach is based on Adamic Adar, a graph-based algorithm. Thanks to Adamic Adar, we have accessibility to the Adamic Adar's index, which is useful for link prediction, especially in contexts like social networks. Considering the number of shared connections between two nodes allows one to calculate the Adamic Adar index. The study proposes PAACDA, an abbreviation for "proximity driven Adamic Adar corrupt detection algorithm," as an alternative to the aforementioned algorithms for detecting data corruption. It outperforms them all when applied to real-world scenarios. After a thorough analysis of both the current techniques for corruption identification and the new approach introduced in this study, the focus shifted to finding practical ways to restore original information for the corrupted ones. Nevertheless, this research does not cover this topic. While datasets with just two characteristics are ideal for the linear regression method of data regeneration, the vast majority of datasets deal with massive volumes of data that include numerous features. One such method for restoring damaged data utilizing the generation and discriminator paradigm is GANs, which stand for Generative Adversarial Networks. Still uncharted territory is the use of different GAN evolutions, most notably tabular GANs, to remediate polluted regeneration efforts. What follows is a breakdown of the rest of the piece. After reviewing relevant literature on the topics under

Vol.15 No.3(2024),144-153
DOI: https://doi.org/10.61841/turcomat.v15i3.14785

investigation in Chapter 2, we go on to Section 3 to provide examples of the data and methodologies used, and finally to Section 4 to lay out the steps needed to resolve the problem. Section 6 presents the findings and potential directions for further study, whereas Section 5 presents the outcomes that show how effective our tactics were.

RELATED WORK

Applying the BACON technique to fuzzy multivariate outliers

For trustworthy decision-making, relying on a sharp cut-off value to detect outliers lacks linguistic significance and understanding. Instead of a hard cut-off threshold, this study suggests two fuzzy treatment approaches for the Blocking Adapt Computationally-efficient Outliers Nominator (BACON) algorithm. Findings from experiments demonstrate that the suggested fuzzy treatment for BACON capture uncertainty at the data's inlier and outlier boundaries and give more insightful interpretations of the final findings compared to the crisp version of the algorithm.

Finding clusters in noisy big geographical datasets using a density-based approach

When it comes to class identification in geographical datasets, clustering methods shine. But certain clustering method needs come up when applied to big geographical databases: identifying clusters of any shape with little to no subject expertise needed, and performing well on massive datasets. When these needs are combined, the popular clustering algorithms fail to provide a solution. For the purpose of finding clusters of any form, we introduce DBSCAN, a novel clustering technique that uses a density-based concept of clustersDBSCAN helps the user choose the right value for the one input parameter it needs. We tested DBSCAN experimentally with both simulated and actual data from the SEQUOIA 2000 standard to determine its efficacy and efficiency. Our experimental findings show that compared to the famous method CLAR-ANS, DBSCAN is more efficient in finding clusters of any shape, and the efficiency gap between the two is more than 100.

Protecting smart grid management systems against covert bogus data injection attacks using an elliptic envelope

In power transmission networks, state estimation is a crucial step. State estimation systems are vulnerable to stealth attacks that use false data injection (SF-DIA), which may lead to power theft,

Vol.15 No.3(2024),144-153
DOI: https://doi.org/10.61841/turcomat.v15i3.14785

small disruptions, or even outages. In order to avoid or lessen the impact of these assaults, it is crucial to identify them accurately and precisely. Here, we provide a method for detecting SFDIA in state estimation that relies on unsupervised learning. The plan employs an elliptic envelope to identify these assaults as outliers and a random forest classifier to reduce the scheme's complexity. We evaluate the elliptic envelope technique with four more unsupervised approaches. Datasets taken from a replicated IEEE 14-bus system are used to train and evaluate all five models. Out of the five unsupervised approaches tested, the elliptical envelope based strategy had the lowest false alarm rate and the highest detection rate.

A survey of LOC algorithms for huge data stream outlier identification

The goal of the statistical process known as "outlier detection" is to identify data points or occurrences that deviate significantly from the typical distribution. Many people in the data mining and ML communities are interested in it. Credit card fraud detection and network surveillance are only two of many significant applications that rely on outlier detection. Global and local outlier identification are the two main categories. When data points are considered out of the ordinary for the whole dataset, they are called global outliers. On the other hand, when data points are considered out of the ordinary for their immediate vicinity, they are called local outliers. The identification of local outliers is the focus of this research. One of the most well-known density-based methods for detecting local outliers is the Location Outlier Factor (LOF). Although there are several LOF methods designed for static data environments, they are not immediately applicable to data streams—a crucial kind of big data. There is a clear need for improved algorithms that can handle the high velocity stream of information streams in order to identify local outliers, since current methods for this purpose are inadequate. With a focus on LOF algorithms, this study surveys the research on local outlier identification in both static and stream settings. It gathers and sorts all the local outlier identification algorithms that are already out there, then reviews and compares their features. In addition, the article delves into the pros and cons of such algorithms and suggests other encouraging avenues for enhancing current techniques of local outlier identification in data streams.

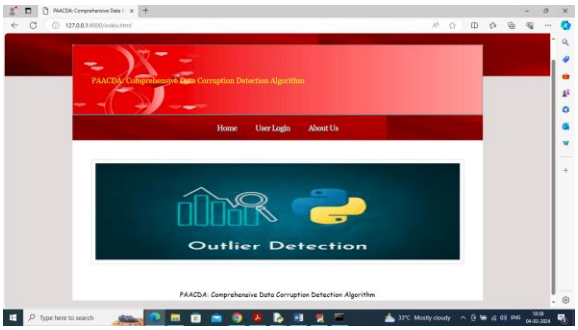
METHODOLOGY

Vol.15 No.3(2024),144-153
DOI: https://doi.org/10.61841/turcomat.v15i3.14785

1) Login: The username and password for the system may be "admin and admin."
2) Dataset loading: Datasets will be uploaded and processed when the user logs in.
3) Execute LOF: Train the loaded dataset using the 'Local Outlier Factor' algorithm to identify corrupted values. After that, compare the discovered corrupted values with the real values to determine accuracy.
4) Explore the Isolation Forest: We will train the 'Isolation Forest' algorithm on the supplied dataset to identify corrupted values. After that, we will compare the discovered corrupted values with the genuine values to determine the accuracy.
5) Carry out a One Class SVM: In order to determine the correctness, the loaded dataset is trained using the 'OCS' algorithm, which can identify corrupted values. Then, the found corrupted values are compared with the genuine values.
6) Deploy PAACDA: The loaded dataset is trained using the 'PACCDA' algorithm to identify corrupted values. The accuracy of the algorithm is then evaluated by comparing the discovered corrupted values with the real values.
7) Combination PAACDA Run Extension:

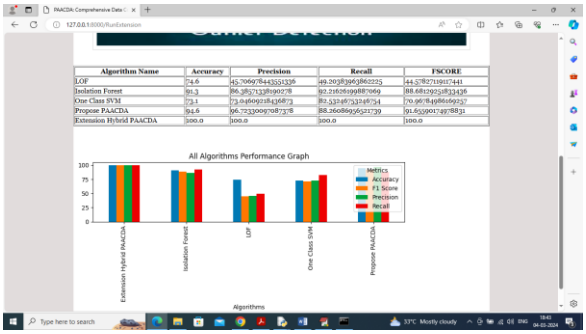
The loaded dataset is trained using the 'PACCDA and Random Forest' technique to identify corrupted values. The accuracy of the algorithm is then evaluated by comparing the discovered corrupted values with the real values.

RESULT AND DISCUSSION



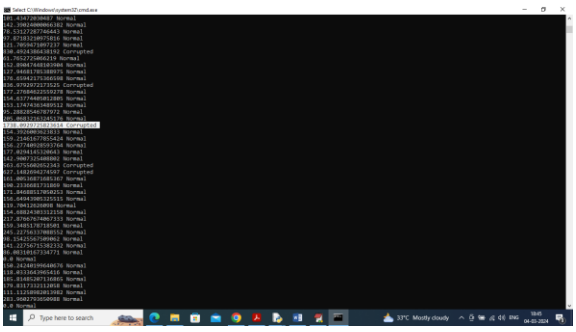
Select "User Login" from the results up top.

Vol.15 No.3(2024),144-153
DOI: https://doi.org/10.61841/turcomat.v15i3.14785



Each algorithm's accuracy along with additional metrics are shown in the above result. Out of all the methods, extension achieved 100% accuracy. The graph's x-axis shows the names of the algorithms, while the y-axis shows the other metrics, including accuracy, represented by different colored bars.

The results below show the detected values under normal and corrupted conditions.



The dataset in the above result is typically between 0 and 300, however the algorithm has identified several values as damaged since they are 850 or over 1300.

CONCLUSION

In order to perform good research, it is vital to have access to dependable and correct data. This is due to the fact that inaccurate and unreliable data leads to misleading outcomes. In industries like healthcare and defense, the consequences of accidentally inputting wrong data onto a computer might be catastrophic. The corruption of data may occur when it is being written, updated, or moved to another destination disk. Furthermore, files might be corrupted by viruses. In most cases, the goal is to damage important system files. Data with high corruption rates may severely hinder the precision of models and the results of data analytics; finding hidden outliers is only half the problem. To validate the data gathered from the sources, precision is essential in this situation.

Vol.15 No.3(2024),144-153
DOI: https://doi.org/10.61841/turcomat.v15i3.14785

Verifying the veracity of data collected is an important part of every research project. Because of this, verifying the reliability of any survey is essential. To begin, this article provides an overview of outlier identification, including its key principles, models, and strategies for identifying tainted data. Next, we use three different highly structured synthetic datasets—small, medium, and medium-large—to separate the data into two groups according to their distribution: linearly distributed and clustered. This allows us to better contain the quality improvement methods for data corruption detection. With an overall accuracy rating 96.35% for cluster results and 99.04% with linear data, PAACDA proved to be the most effective method. Finally, we present an experimental comparison of numerous state-of-the-art quality improvement methods using a wide range of quality evaluation metrics; the authors have synthesized the results of different statistical and probabilistic models, and they detail how they applied the novel PAACDA algorithm to the data in order to achieve their goals. Other greatest performers in the grouping dataset received accuracy rates of 94.46% and 95.05%, respectively, alongside HBOS and MAD. A few examples of methods with decent accuracy in the middle range include COPOD (92.43%), GMM (87.01%), LUNAR (82.37%), Elliptic Envelop (72.17%), K-Means clustering (86.06%), ECOD (82.37%), and Isolation Forest (82.37%). Every class—SVM, Deep SVDD, PCA, ROD, LOF, and DBSCAN—recorded an accuracy index of 76.72%, 72.25%, 72.53%, 62.71%, 59.47%, and 39.50%, respectively, indicating the worst performance. Many previous top-performing models, including HBOS, MAD, COPOD, and GMM, performed better on the linear dataset, with approximate precision levels of 95.00%, 94.77%, 92.27%, and 92.15%, respectively. The following algorithms were found to have high accuracy rates: K-Means clustering (86.70%), LUNAR (86.87%), Isolation forest (82.22%), ECOD (82.83%), and Deep SVDD (76.25%). More linearly oriented models outperformed their non-linear counterparts. Accuracy rates of 73.01%, 72.28%, 62.83%, 58.79%, & 43.20% were recorded using PCA, Single Class SVM, ROD, LOF, & DBSCAN Clustering, in that order. Accuracy was same regardless of the amount of the dataset. The problem was the same as before: when corruption rose, performance declined.

REFERENCES

Vol.15 No.3(2024),144-153
DOI: https://doi.org/10.61841/turcomat.v15i3.14785

[1] E. Bergdorf, Predicting the impact of data corruption on the operation of cyber-physical systems. 2017. [2] V. Chandelle, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM computing surveys (CSUR), vol. 41, no. 3, pp. 1–58, 2009.

[3] M. Pang-Ning and V. Steinbach, Introduction to data mining. Pearson Education India, 2016.

[4] H. M. Tony, A. S. Moussa, and A. S. Hadid, "Fuzzy multivariate outliers with application on BACON algorithm," in 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2020.

[5] S. Thamud, P. Branch, J. Jinn, and J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," J. Big Data, vol. 7, no. 1, 2020, DOI: 10.1186/s40537-020-00320-x.

[6] O. J. Oyemade, O. O. Oladipo, and I. C. Barbuda, "Application of k Means Clustering algorithm for prediction of Students Academic Performance," arrive [clog], 2010. [Online]. Available: http://arxiv.org/abs/1002.2425

[7] H. L. Sari, D. Sorani Mrs, and L. N. Zulia, "Implementation of k-means clustering method for electronic learning model," J. Phys. Conf. Ser., vol. 930, p. 012021, 2017, doe: 10.1088/1742-6596/930/1/012021.

[8] M. Ester, H.-P. Krieger, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," In kid, vol. 96, no. 34, pp. 226–231, 1996.

[9] D. Deng, "DBSCAN clustering algorithm based on density," in 2020 7th International Forum on Electrical Engineering and Automation (IFEEA), 2020.

[10] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in 2008 Eighth IEEE International Conference on Data Mining, 2008. 24 VOLUME 10, 2022 this article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3253022 this work is licensed under a Creative Commons Attribution-Non-commercial-No Derivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

[11] R. Gao, T. Zhang, S. Sun, and Z. Liu, "Research and improvement of Isolation Forest in detection of local anomaly points," J. Phys. Conf. Ser., vol. 1237, no. 5, p. 052023, 2019, DOI: 10.1088/1742-6596/1237/5/052023.

Vol.15 No.3(2024),144-153
DOI: https://doi.org/10.61841/turcomat.v15i3.14785

[12] M. Ashrafi zaman, S. Das, A. A. Illegally, Y. Chak-chak, and F. T. Sheldon, "Elliptic envelope based detection of stealthy false data injection attacks in smart grid control systems," in 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 2020.

[13] C. McKinnon, J. Carroll, A. McDonald, S. Kaikoura, D. Infield, and C. Strachan, "Comparison of new anomaly detection technique for wind turbine condition monitoring using gearbox SCADA data," Energies, vol. 13, no. 19, p. 5152, 2020, DOI: 10.3390/en13195152.

[14] Goldstein, Markus, and Andreas Denel. "Histogram-based outlier score (hobs): A fast unsupervised anomaly detection algorithm." KI-2012: poster and demo track 9 (2012).

[15] N. Kazlauskas and A. Basks, "Application of histogram-based outlier scores to detect computer network anomalies," Electronics (Basel), vol. 8, no. 11, p. 1251, 2019, DOI: 10.3390/electronics8111251.

[16] I. T. Jolliffe and J. Kadima, "Principal component analysis: a review and recent developments," Philos. Trans. A Math. Phys. Eng. Sci., vol. 374, no. 2065, p. 20150202, 2016, doe: 10.1098/rsta.2015.0202.

[17] S. Mishra et al., "Principal Component Analysis," Int. J. Lives. Res., p. 1, 2017, DOI: 10.5455/ijlr.20170415115235.