Pegem Journal of Education and Instruction, Vol. 13, No. 4 , 2023 (pp. 462-471)

RESEARCH ARTICLE WWW.PEGEGOG.NET

PAAFDA: Inclusive Data Fudging Detection Algorithm

[1]Mr. P. Prashanth Kumar, [2]V. Supraja ,[3]K. Pranathi ,[4]E. Naveena

[1]Associate Professor, Computer Science and Engineering, Sridevi Women's Engineering College, Hyderabad, India
Email: swecprashanth@gmail.com
[2,3,4,]B.Tech Student, Department of Computer Science and Engineering, Sridevi Women's Engineering College, Hyderabad, India

ABSTRACT

Modern technology has elevated data and its analysis from the status of scattered spreadsheet values and characteristics to that of a tool to revolutionize any major industry. It is critical to create a reliable system that can detect and properly highlight all instances of corrupted data in the dataset, as data fudging may originate from many different unethical and unlawful sources. A difficult challenge is the detection of damaged data and the recovery of data from a corrupted dataset. Unless this is handled early on, it could cause issues when processing data using machine or deep learning techniques later on. Next, we provide PAAFDA, an acronym for "Proximity based Adamic Adar Fudging Detection Algorithm," and then we compile the findings, highlighting how it identifies damaged data instead of outliers. Isolation forest, DBSCAN (which stands for "Density-Based Spatial Clustering of Applications with Noise"), and other state-of-the-art models depend on parameter tuning to achieve high recall and accuracy, but they introduce a large degree of uncertainty when it comes to corrupted data. For linear and clustered damaged datasets, the authors of this paper investigate the most specific performance concerns of several unsupervised learning techniques. Furthermore, a new PAAFDA technique is suggested; it surpasses conventional unsupervised learning benchmarks on fifteen widely used baselines, such as LOF (Local Outlier Factor), K-means clustering, and isolation forests, achieving an accuracy of 96.35% for clustered data and 99.04% for linear data. Along with presenting the aforementioned viewpoints, this article thoroughly examines the relevant literature. We conclude from this study what has to be done in this area going forward by identifying all the problems with the current methods.

INTRODUCTION:

Nearly every aspect of human experience has seen tremendous advancement in technology from the birth of our species. The vast quantities of data are largely responsible for the continuous progress in technology; without them, much of this sector would grind to a halt. The firms with more data seem to have a stranglehold on the market since data has

Corresponding Author e-mail: swecprashanth@gmail.com
How to cite this article: 1Mr. P. Prashanth Kumar, 2V. Supraja ,3K. Pranathi ,4E. Naveena. PAAFDA: Inclusive Data Fudging Detection Algorithm.Pegem Journal of Education and Instruction, Vol.
13, No. 4, 2023, 462-471
Source of support: Nil **Conflicts of Interest:** None. **DOI:** 10.48047/pegegog.13.04.55

Received: 12.10.2023

Accepted: 22.11.2023 **Published:** 24.12.2023

 fying data fudging, we investigated additional approaches that have been developed for detecting data fudging, with a focus on outliers. Through extensive research, we learned about many algorithms with different degrees of accuracy when evaluated on a dataset with damaged data instead of outliers. To solve problems with categories and their categorization, Kmeans clustering employs clusters and their cancroids as an unsupervised method. As it detects core samples of high density and builds upon them, DBSCAN - another clustering-based technique-tends to perform better with data containing clusters of comparable density. It was clear that both approaches were effective when used to spot outliers in the given dataset. Methods including Isolation Forest, Elliptic Envelope, and Histogram-Based Outlier Detection were investigated as the study progressed. We may find the outliers that go over the specified range by using the method that isolation forest gives us to split the dataset features. 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. An further successful unsupervised technique for anomaly identification is the Histogram based algorithm for outlier detection (HBOD) method, which likewise involves displaying and analyzing histograms.

When it comes to spotting outliers in the dataset, these algorithms are also rather accurate. To further quantify the degree of accuracy in forecasting the outliers for the artificially produced dataset, algorithms including

become so crucial. An algorithm or piece of technology can't go beyond its first stage without data. Nowadays, data is vital

to every company, making it even more important to safeguard this important resource from harmful manipulation. Even little changes to a dataset may have farreaching consequences due to the snowball effect. Even though there are various immoral methods to corrupt data, continuous study has been undertaken throughout time to develop effective strategies to learn about data fudging, to mention a few outstanding efforts in detecting data fudging including. Prior to devising a novel strategy for identify VOLUME 10, 2022 1 The IEEE Access journal has approved this piece for publication. The material may be revised before final publication, since this is the author's version that has not been completely edited. The publication details are as follows: DOI 10.1109/ACCESS.2023.3253022. The

"Principal Component Analysis" (PCA), "DeepSVDD," and "Rotation based Outlier Detection" (ROD) were explored. To mention a few PCA, ROD, Local Outlier Factor, DeepSVDD and more were utilized. 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 fudging detection is an area where Adamic Adar's potential is growing, since it is an algorithm that shows promise for data correlation in graph networks. Based on the findings of the aforementioned investigation, it is possible to circumvent the inefficiencies of the present work. The goal of this work is to improve the current fudging detection algorithm's accuracy by using the Adamic Adar algorithm's strong data correlation capabilities and to integrate the existing research in this area. Adamic Adar is the central figure in the research's suggested innovative approach, which is based on a graph-based algorithm. The Adamic Adar index, made available to us by Adamic Adar, helps in link prediction, especially in domains like social networks. Consideration of the number of shared connections between two nodes yields the Adamic Adar index.

The study proposes PAAFDA, an abbreviation for "proximity based Adamic Adar fudging detection algorithm," as an alternative to the aforementioned algorithms for detecting data fudging. It outperforms them all when applied to realworld scenarios. After a thorough analysis of both the current techniques for fudging identification and the new approach introduced in this study, the focus shifted to finding practical ways to restore the original data for the corrupted ones. But that's not going to be covered in this research. For data regeneration, the linear regression method works well with datasets that only have two characteristics, however most datasets deal with massive volumes of data that have several features. One such method for restoring damaged data utilizing the generator 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. The parts that follow are the meat and potatoes of this article. Section 2 provides an overview of relevant literature about the methodologies used for this research. Section 3 details the data and techniques utilized, and Section 4 presents the recommended methodology to address the problem. The data that show how effective our tactics are are presented in Section 5, and

the conclusions and future directions of our study are presented in Section 6.

RELATED WORK:

A survey on anomaly detection

Many different fields of study and fields of application have investigated anomaly detection since it is a significant challenge. While some anomaly detection methods are more general, others have been tailored to particular use cases. An organized and all-encompassing synopsis of anomaly detection studies is the goal of this study. Based on the basic principle used by each method, we have classified the current procedures into several groups. The methods distinguish between typical and out-of-the-ordinary actions based on the assumptions we've defined for each category. These assumptions may be used as recommendations to measure the efficiency of a given approach in a specific area when applying it to that domain. We provide a foundational anomaly detection method for each category and demonstrate how the many current methods within that category are variations on that method. The methods that fall under each heading are more clearly and concisely laid forth in this template. In addition, we list the pros and cons of the strategies that fall into each category. Since the computational complexity of the methods is a significant concern in practical application fields, we also address it. We anticipate that our study will shed light on the many tracks taken by researchers interested in this problem, as well as the ways in which methods established in one field may be transplanted to other, unanticipated contexts.

Applying the BACON technique to fuzzy multivariate outliers

It is neither linguistically relevant nor enlightening to rely on a precise cut-off number to detect outliers for dependable decision-making. Instead of a hard cut-off threshold, this study suggests two fuzzy treatment approaches for the Blocked Adaptive Computationally-efficient Outlier Nominator (BACON) algorithm. The experimental findings demonstrate that compared to the crisp version of BACON, the suggested fuzzy treatments capture the uncertainty at the data's inlier and outlier boundaries and offer more meaningful interpretations to the final results.

Finding clusters in noisy big geographical datasets using a densitybased approach

Class identification in geographical datasets is a desirable job for clustering algorithms. The following criteria for clustering algorithms are raised, however, when used to big geographic datasets: efficient identification of clusters with various shapes, minimum domain knowledge requirements for determining the

input parameters, and high performance on huge databases. When these needs are combined, the popular clustering algorithms fail to provide a solution. This study introduces DBSCAN, a novel clustering technique that uses a density-based concept of clusters to find clusters with any form. There is only one input parameter that DBSCAN needs, and it helps the user find the right value. We tested DBSCAN experimentally with both simulated and actual data from the SEQUOIA 2000 benchmark to determine its efficacy and efficiency. Our findings show that compared to the famous technique CLAR-ANS, DBSCAN is more efficient in finding clusters of any shape, and that the efficiency gap between the two algorithms is around 100 times wider.

Hidden fake data injection attacks in smart grid control systems may be detected using an elliptic envelope.

Power transmission systems rely on state estimate. State estimation systems are vulnerable to stealthy false data injection attacks (SF-DIA), which may lead to power theft, small disruptions, or even outages. To avoid or lessen the impact of these assaults, accurate and precise detection is crucial. In this study, 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. A dataset from a simulated IEEE 14-bus system is used to train and evaluate all five models. Out of the five unsupervised approaches tested, the elliptic envelope based strategy had the lowest false alarm rate and the highest detection rate.

Methods for detecting outliers in large data streams using the Local Outlier Factor

The goal of the statistical process known as "outlier detection" is to identify data points or occurrences that deviate significantly from the typical distribution. The data mining and ML communities have taken a keen interest in it. Many applications rely on outlier detection, such as those that identify network intrusions and credit card fraud. Outlier detection may be either global or local. 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 Local Outlier Factor (LOF). Data streams are a significant kind of large data, but many LOF techniques designed for static data environments are incompatible with them. It is clear that current methods for detecting local outliers in data streams are inadequate, and that new algorithms are required to adequately analyse the very fast data streams in order to do this task. 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 examines what makes each one unique. 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:

The DJANGO WEB framework, which includes the following modules, was used to construct this project.

1) User Login: The system may be accessed by using the username "admin" and the password "admin."

2) Next, the user will need to log in before they can load and execute the dataset.

3) Execute LOF: To do this, first, train the loaded dataset using the 'Local Outlier Factor' method; second, compare the discovered corrupted values to the real values in order to determine accuracy.

4) run the "Isolation Forest" algorithm on the supplied dataset. This will help identify corrupted values; after that, you can compare the found corrupted values to the genuine values to see how accurate it was. 5) Execute One Class SVM: Prepare the dataset by training it using the 'OCS' method; compare the resulting distorted values to the genuine values to determine accuracy.

6) Execute PAACDA: To do this, first, train the dataset using the 'PAACDA' method; second, compare the identified corrupted values to the real values to determine the accuracy.

7) Execute Extension Hybrid PAAFDA: the loaded dataset is trained using the 'PAAFDA and Random Forest' method to identify corrupted values. Then, the reported corrupted values are compared to the actual values in order to determine the accuracy.

CONCLUSION:

To perform good research, one must have access to dependable and correct data. This is because inaccurate or misleading data leads to misleading conclusions. The healthcare and military industries are particularly vulnerable to the catastrophic consequences that might result from the accidental input of inaccurate data into computers. Data fudging may occur when it is being written, modified, or moved to another disc. Furthermore, files might be corrupted by viruses. In most cases, the goal is to damage important system files. Discovering hidden outliers in a dataset is only half the battle; corrupted data severely reduces the accuracy of models and the results of data analytics. For this reason, precision is key when checking the sources for this kind of information. 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. We begin by presenting the fundamental ideas of outlier identification and then go on to show how these models and methods may be used to identify tainted data. Next, we use three high-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 encapsulate the quality improvement methodologies for data fudging detection. When compared to the other algorithms, PAAFDA achieved the highest accuracy rates (96.35 percent for clustered data and 99.04 percent for linear data). Lastly, 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 synthesised the results of different statistical and probabilistic models and detailed their use of the novel PAAFDA algorithm to achieve their data goals. Among the other top performers for the clustering dataset, HBOS and MAD had accuracy scores of 95.05% and 94.46%, respectively. With accuracy scores of

92.4391.95%, 87.01%, 72.17%, 86.06%, 82.71, and 82.37%, respectively, COPOD, GMM, LUNAR, Elliptic Envelop, ECOD, and Isolation Forest are among the middle performers. Accuracy rates of 76.82%,

72.25%, 72.53%, 62.71%, 59.47%, and 39.60% were recorded by the one class

Class SVM, DeepSVDD, PCA, ROD, LOF, and DBSCAN, in that order. Among the prior top-performing models, HBOS, MAD, COPOD, and GMM all achieved superior results on the linear dataset, with respective accuracies of 95.00%, 94.77%, 92.27%, and 92.15%. The following algorithms achieved varying degrees of accuracy: K-Means clustering (86.70%), LUNAR (86.87%), Isolation forest (82.22%), ECOD (82.83%), and

DeepSVDD (76.25%). The outcomes were better for models that were designed to handle linear data. Accuracy rates of

73.01%, 72.28%, 62.83%, 58.79%, and 43.20% were recorded using PCA, One

Class SVM, ROD, LOF, and DBSCAN Clustering, in that order. Accuracy was same regardless of the amount of the dataset. The problem was the same as before: when fudging rose, performance declined.

REFERENCES:

[1] E. Burgdorf, Predicting the impact of data fudging on the operation of cyberphysical systems. 2017. [2] V. Chandola, 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. Touny, A. S. Moussa, and A. S. Hadi, "Fuzzy multivariate outliers with application on BACON algorithm," in 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2020.

[5] S. Thudumu, P. Branch, J. Jin, 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/s40537-02000320-x.

[6] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, "Application of k Means Clustering algorithm for prediction of Students Academic Performance," arXiv [cs.LG], 2010. [Online]. Available: http://arxiv.org/abs/1002.2425

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

[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," In kdd, 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-

[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: 10.1088/17426596/1237/5/052023.

[12] M. Ashrafuzzaman, S. Das, A. A. Jillepalli, Y. Chakhchoukh, 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. Koukoura, D. Infield, and C. Soraghan, "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 Dengel. "Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm." KI-2012: poster and demo track 9 (2012).

[15] N. Paulauskas and A. Baskys, "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. Cadima, "Principal component analysis: a review and recent developments," Philos. Trans. A Math. Phys. Eng. Sci., vol. 374, no. 2065, p. 20150202, 2016, doi: 10.1098/rsta.2015.0202.

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

[18] A. Karimian, Z. Yang, and R. Tron, "Rotational outlier identification in pose graphs using dual decomposition," in Computer Vision – ECCV 2020, Cham: Springer International Publishing, 2020, pp. 391–407.

[19] Y. Almardeny, N. Boujnah, and F. Cleary, "A novel outlier detection method for multivariate data," IEEE Trans. Knowl. Data Eng., vol. 34, no. 9, pp. 4052–4062, 2022, doi: 10.1109/tkde.2020.3036524.

[20] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, “A review of Local Outlier Factor algorithms for outlier detection in
big data streams,” Big Data Cogn. Comput., vol. 5, no. 1, p. 1, 2020, doi: 10.3390/bdcc5010001.

[21] M. M. Breunig, R. T. Kriegel, and J. Ng, “LOF: identifying density-based local outliers,” in Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000, pp. 93–10