RECENT STRATEGIES FOR DETECTING OUTLIERS :A REVIEW

Abstract

Among the study fields, outlier detection is one of the most significant and rapidly evolving. For millennia, outlier defection has been used to identify anomalous occurrences in the data and, if necessary, remove them. One of the most significant and quickly evolving fields of research is outlier detection. Investigations into the critical problem of outlier detection have been conducted in a variety of academic and practical sectors. forth order to provide solutions for properly handling outliers, researchers are putting forth endless effort to develop dependable approaches. In this survey, the researcher offers a thorough analysis and methodical review of how different approaches to outlier detection have evolved over the past 20 years. In this paper, the researcher covers the basic ideas behind outlier detection as well as applications of outlier detection techniques and different approaches to outlier detection. This article offers a deeper understanding of outlier detection approaches and recent advancements in outlier detection techniques, along with an overview of the structure of existing outlier identification tactics. The researcher also categorises the various outliers detecting techniques into different techniques such as distance, clustering, density ensemble, and learning-based methods.

Authors

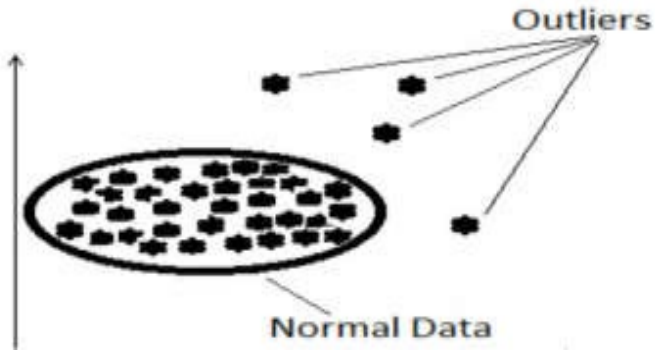
C. Jayaramulu
Research Scholar
Dayananda Sagar University
Bangalore, India.
jayaramcomp@gmail.com

Bondu Venkateswarlu
Associate Professor
Department of CSE Dayananda Sagar University
Bangalore, India.

SK Althaf Hussain Basha
Professor
Department of CSE
Krishna Chaitanya Institute of Technology & Sciences
Markapur, India.

I. THE FIRST REMARKS

The amount of data in the globe has been increasing at an exponential rate overall. The goal of the computer science discipline known as "datamining" is to extract meaningful information from vast amounts of unstructured data sources. It involves going through databases looking for recurrent patterns. Finding unique data items that deviate from the bulk of data sets and data objects is the process of identifying outliers. If a data item deviates significantly from the bulk of other data items in the data set, raising doubts about its inclusion in the data collection, it is said to be an outlier[1]. A data object that deviates from the majority of other data objects in the dataset is called an outlier, as shown in figure 1. The process of identifying data objects within a dataset that are inconsistent with the majority of other data objects in the dataset is referred to as "detecting outliers," and it is defined as the discovery of these data items. In the last several years, the detection of outliers has become a significant field of research since it may provide valuable information in a number of disciplines. Techniques for identifying outliers may be used in a number of situations, such as identifying fraudulent credit card transactions in the banking sector, seeing anomalies in the study of medical data, and spotting unauthorised access to computer network resources. Numerous more uses included finding differences in astronomical data, finding discrepancies in census data, identifying anomalies in robot behaviour, locating faults in image processing, identifying problems in industrial objects, and identifying abnormalities in web applications. There are many different types of outliers, such as point outliers, collective outliers, and contextual outliers.



II. APPLICATION AREA(S) FOR OUTLIER DETECTION

1. **Data and Process Logs:** Organisations may get access to hidden information on linked websites with the use of outlier identification tools, which save time and financial costs. More work and money would be needed if outlier identification methods weren't offered. The enormous amount of data in the logs requires the use of a variety of automated data mining techniques in order to search for unexpected patterns [3]. These logs are a valuable source of information that can be used for outlier detection monitoring.

2. **Fraud Detection as well as Intrusion Detection:** The usual buyer's behaviour will change in the process of identifying fraudulent activity if a card is lost or stolen. It is possible for you to notice an odd buying pattern. Similarly, illegal access inside computer networks might result in a distinct pattern [4]. It is crucial to identify these anomalous patterns, sometimes referred to as outliers, in order to preserve security.

3. **Monitoring and Safety Apparatus:** It needed to be protected and monitored in relation to the cyber security subject. In computer networks, the processes of producing secured logging and log management are essential because they increase authenticity and security intelligence. The detection of outliers in surveillance footage is an interesting and useful research question in the present context.

4. **Misinformation, Fake News, and Social Media Websites:** Social media has given people a platform to constantly spread misleading information in recent years. Sometimes it could be difficult to distinguish between news that is true and that is fake. On the other hand, incorrect news items may be mistakenly identified as outliers because of how much they stand out when they originate from a reliable source [6]. It is essential to acknowledge the destructive impact that the spread of misinformation has on both people and society at large.

5. **Assessment of Health Care and medical Diagnosis:** These devices often produce strange patterns or readings that, when considered in the context of medical applications and the health care system, usually imply the diagnosis of a disease condition [25–26]. Medical practitioners may then take the appropriate preventive action after accurately diagnosing the ailment and its underlying causes thanks to the detection and understanding of the abnormal patterns.

6. **Data Sources of Transactions:** Data bases' operations are documented in the audit logs of financial transactions. The audit logs are useful for assessing the accuracy, legality, and riskiness of the information being reported. It is crucial to keep a close eye on the audit logs to spot and report any unusual conduct [7].

7. **Sensor Networks and Databases:** Effective network routing and reliable sensor outputs have been made possible in sensor settings such wireless networks[8,9], target tracking environments[10], and body sensor networks [11] by the identification of outliers. It is useful for tracking computer network performance in several aspects, including identifying network bottlenecks.

III. METHODS BASED IN STATISTICS

Statistical techniques may be used to identify outliers in supervised, semi-supervised, unsupervised, and unsupervised learning environments. The phrases "outliers" and "inliers" have varied meanings based on the data distribution model. The two primary categories of statistically based methods are the parametric approach and the nonparametric approach. The main difference between the two methods is that the former uses the available data to estimate the parameters of the distribution model, based on an assumption about the underlying distribution model in the supplied data. In this study, we divide the existing body of work on using statistical techniques to find outliers into three groups: parametric strategies, non-parametric strategies, and other types of statistical strategies.

1. **Methods of Measurement:** Two popular techniques for identifying outliers are the regression model and the Gaussian mixture model. The distribution model behind the data being studied is assumed in each of these approaches.

- **Gussian Mixture Model methods:** The Gaussian model is one of the statistical techniques most frequently employed to find outliers. The mean and variance estimates of the Gaussian distribution in the model are computed using the maximum likelihood estimate (MLE) approach [12]. The mean and variance estimates are obtained using this method. Statistical discordance tests, such as the box plot and the mean variance test, are employed in the testing phase.

- **Regression methods and Techniques:** Regression models are a simple tool for identifying outliers and overcoming outlier identification difficulties. Depending on the kind of problem that has to be handled, the user may choose between a linear and a nonlinear model. When using this approach, the first stage, referred to as the training stage, often entails building a regression model that is consistent with the data. A data point that has a significant discrepancy between its actual value and the value predicted by a regression model is termed an outlier. This is done in the following phase, known as the test step, when the regression model is tested by comparing each data instance to the model. A significant departure from both the actual value and the predicted value characterises such a point. Regression techniques are often used to identify outliers. Typical methods include the Mhalanobis distance threshold robust least squares with bi square weights, mixed models, and an additional vibration model. The Bayesian technique for doing regression is one more potential strategy [13]. Although Satman [14] suggested a different method for locating outliers in linear regression, all methods rely on regression models to find outliers. The least trimmed square estimation method is based on concentration stages and a non-interactive covariance matrix. The technique has the advantage of quickly identifying a large number of outliers, which lowers the amount of processing work needed. However, since egression models can occasionally exhibit minute preference, more research may be done to minimise the bias and variance of the intercept estimator in order to improve the model's performance.

2. **NON-Parametrical Methods:** Kernel density estimation [15]: The KDE approach is a popular nonparametric technique for finding outliers. Using kernel functions, Latecki described an unsupervised approach to outlier detection in [16]. A crucial stage in the outlier identification process is comparing each point's local density to that of its neighbouring points. When the proposed techniques are compared to other prominent density-based approaches, the experimental results show that the recommended strategies outperform them. However, the method is not suitable for extremely big and high-dimensional real-world datasets. This might be an extension of the current research. Afterwards, Goa et al. [17] devised a more effective way to address some of the issues that had existed before. Using kernel-based strategies, the methods show improved performance and scalability for big data sets while reducing computation time when compared to the approaches presented by LOF and Latecki et al. [18]. They also resolved the problem of incorrect outlier identification in complex and large-scale data sets by employing variable kernel density estimates. Finally, consider the degree to which the LOF depends on the parameter k, which denotes the relative significance of the neighbourhood under consideration. They came up with a plan that included an estimate of the weighted neighbourhood density in order to fix the problem. Generally speaking, the technique reduces processing time while simultaneously increasing speed and

scalability for large data sets. In order to identify rogue nodes in the network, Kumar and Verma[19] utilise KDE to estimate the distribution of sensor data.

3. **Other Statistical:** approaches Numerous statistical procedures have been suggested; nevertheless, some of the most straightforward statistical methods for identifying outliers are the histogram[20] and other statistical tests[21], such as the box plot, trimmed mean, extreme student zed deviation, and Dixon type test. The trimmed mean, for instance, is less susceptible to outliers than the other methods; however, the extreme student zied deviation test is the most effective method for identifying individual outliers. The Dixon type test's main advantage is that it can function well even with a very small sample size since it does not need the premise that the data are normal.

IV. DISTANCE-BASED APPROACHES

Algorithms based on distance are used to compute the distances between sites in order to find outliers. The definition of distance-based outlier detection that is most frequently used is based on the idea of the local neighbourhood, k closest neighbour (KNN)[22], and the conventional distance threshold. A data point is considered an outlier if it differs significantly from the one that is closest to it.

K Nearest Neighbour Approaches: These methods are among the most effective for identifying outliers, according to a number of researchers who have used them to calculate outliers. However, this classification has nothing to do with the k-nearest neighbour classification. These techniques are most often used in an attempt to find global outliers. They spend a lot of time analysing the details of an object's neighbourhood to ascertain whether or not it is situated unusually close to its neighbours or has a low density. The first step in the process is to search for every record's k closest neighbours. These neighbours are then used to calculate the outlier score. Using low-density data or close neighbours is one of the most important guidelines to keep in mind while searching for outliers.

V. METHODS PREDICATED ON CLUSTERING

Utilising clustering to characterise the behaviour of the data is standard procedure in clustering-based approaches. Outliers are smaller overall and have a much lower number of data points than other clusters. It is crucial to remember that there are differences between the methods used for clustering and the process of finding outliers. Cluster sites are found using clustering methods, while outlier locations are found using outlier detection techniques. The ability of clustering-based approaches to capture the cluster structure of normal instances is influenced by a number of factors, including the quality of the clustering algorithm [22], Deep learning-based [23,24], Clustering methods [23], Ensembling techniques [24], Partial Least Square approach [27], and Data Science approach [28]. Five final remarks: In today's fast-paced economy, the identification of outliers is an essential component in a wide number of application disciplines. In this study, we address the significance of detecting outliers in data mining as well as the features of numerous outlier detection techniques that can be found in the literature. We examine the relative performances of these algorithms in particular. In this paper, the researcher presented a thorough analysis that was methodically arranged and included the most current and cutting-edge techniques for using categorization to find outliers. The many types of outlier detection algorithms are thoroughly analysed in this work

and are categorised as density-based, statistical-based, distance-based, ensemble-based, and learning-based techniques, in that order. We examined some of their biggest benefits, drawbacks, and challenges in the comment section. Furthermore, the investigator tried to enumerate and provide the latest unresolved research challenges and obstacles. The applications for which outlier detection techniques are employed were also examined. Prior to selecting a course of action for an outlier detection problem, the researchers were able to obtain a thorough understanding of the essential prerequisites for these approaches.

VI. CONCLUSION

The ability to identify outliers is crucial in many application disciplines in the fast-paced economy of today.The significance of recognising and comprehending anomalies inData mining is covered in this book along with other related subjects.characteristics of several outliers found in anomaly detectionalgorithms that are available in academic publications. The Inquiriersupplied a thorough questionnaire in a systematic way.within the parameters of this investigation, which looks at innovative practisesa technique for classifying the data in order to discover outliers. This paper presents a thorough examination of theThere are several varieties of techniques for identifying outliers.Classifying them into distance-based clustering, ensemble-based ensembles, density-based statistics, andas well as learning-focused techniques.We examined a few of their most notable benefits. Drawbacks of our conversation as well as challengespassage of writing. Furthermore, the researcher made an effort to look intoand list a few of the unanswered questions that are currently being investigated.issues to be resolved. Additionally, there were outlier detection techniques. Assessed in light of their intended use.The researchers were successful in recruiting subjects for their investigations.Armed with a thorough understanding of the prerequisitesmany methods before selecting one to use for a finding outliers is a challenging process.

REFERENCES

[1] E. L. Paula, M. Ladeira, R. N. Carvalho, and T. Marzagáo, ‘‘Deep learning anomaly detection as support fraud investigation in Brazilian exports and anti-money
[2] laundering’’ in Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Anaheim, CA, USA, Oct. 2016, pp.954–960.
[3] U. Porwal and S. Mukund, ‘‘Credit card fraud detection in e-commerce: An outlier detection approach,’ 2018, arXiv:1811.02196. [Online]. Available: https://arxiv.org/abs/1811.02196
[4] K. Alrawashdeh and C. Purdy, ‘‘Toward an online anomaly intrusion detection system based on deep learning,’’ in Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Anaheim, CA, USA, Dec. 2016, pp. 195–200.
[5] G. Gebremeskel, C. Yi, Z. He, and D. Haile, ‘‘Combined data mining techniques based patient data outlier detection for healthcare safety,’’ Int. J. Intell. Comput. Cybern., vol. 9, no. 1, pp. 42–68, 2016.
[6] V. J. Hodge and J. Austin, ‘‘A survey of outlierdetection methodologies,’’ Artif. Intell. Rev., vol. 22, no. 2, pp. 85– 126, 2004.
[7] C. C. Aggarwal and P. S. Yu, ‘‘An effective andefficient algorithm for high-dimensional outlierdetection,’’ Int. J. Very Large Data Bases,vol. 14, no.2, pp. 211–221, 2005.
[8] F. Angiulli, S. Basta, and C. Pizzuti, ‘‘Distancebaseddetection and prediction of outliers,’’ IEEE Trans.Knowl. Data Eng., vol. 18, no. 2, pp. 145–160, Feb.
[9] 2006.
[10] M. Breunig, H. Kriegel, R. T. Ng, and J. Sander,‘‘LOF: Identifying density-based local outliers,’’ ACMSIGMOD Rec., vol. 29, no. 2, pp. 93–104, 2000.
[11] F. Cao, M. Ester, W. Qian, and A. Zhou, ‘‘Densitybasedclustering over an evolving data stream withnoise,’’ in Proc. SIAM Conf. Data Mining, Apr. 2006,pp. 328–339.
[12] Z. Zheng, H. Y. Jeong, T. Huang, and J. Shu,‘‘KDE based outlier detection on distributed data streams in multimedia network,’’ Multimedia ToolsAppl., vol. 76, no. 17, pp. 18027–18045, Sep. 2017.

[13] L. L. Sheng, ‘‘Fractal-based outlier detection algorithm over RFID data streams,’’ Int. J. Online Eng., vol 12, no. 1, pp. 35–41, Feb. 2016. D. van Hieu and P. Meesad, ‘‘A fast outlier
[14] detection algorithm for big datasets,’’ in Recent Advances in Information and Communication Technology (Advances in Intelligent Systems and Computing), vol. 463, P. Meesad, S. Boonkrong, and H. Unger, Eds. Cham, Switzerland: Springer, 2016.
[15] X. T. Wang, D. R. Shen, M. Bai, T. Z. Nie, Y. Kou, and G. Yu, ‘‘An efficient algorithm for distributed outlier detection in large multi-dimensional datasets,’’ J. Comput. Sci. Technol., vol. 30, no. 6, pp. 1233–1248, Nov. 2015.
A. Ayadi, O. Ghorbel, A. M. Obeid, and M. Abid, ‘‘Outlier detection approaches for wireless sensor networks: A survey,’’ Comput. Netw., vol. 129, pp. 319–333, Dec. 2017.
[16] J. Mao, W. Tao, C. Jin, and A. Zhou, ‘‘Feature grouping-based outlier detection upon streaming trajectories,’’ IEEE Trans. Knowl. Data Eng., vol. 29, no. 12, pp. 2696–2709, Dec. 2017.
[17] C. D’Urso, ‘‘EXPERIENCE: Glitches in databases, how to ensure data quality by outlier detection techniques,’’ J. Data Inf. Qual., vol. 7, no. 3, 2016, Art. no. 14.
[18] H. P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, ‘‘Outlier detection in axis-parallel subspaces of high dimensional data,’’ in Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining, Berlin, Germany: Springer, 2009,pp. 831–838.
[19] Z. Cai, Z. He, X. Guan, and Y. Li, ‘‘Collective data-sanitization for preventing sensitive information inference attacks in social networks,’’ IEEE Trans. Depend. Sec. Comput., vol. 15, no. 4, pp. 577–590, Aug. 2018.
[20] Y. Yu, L. Cao, E. A. Rundensteiner, and Q. Wang, ‘‘Outlier Detection over Massive-Scale Trajectory Streams,’’ ACM Trans. Database Syst., vol. 42, no. 2, pp. 10:1–10:33, 2017.
[21] Y. Djenouri, A. Belhadi, J. C.-W. Lin, D. Djenouri, and A. Cano, ‘‘A survey on urban traffic anomalies detection algorithms,’’ IEEE Access, vol. 7, pp. 12192– 12205, 2019.
[22] Y. Djenouri, A. Zimek, and M. Chiarandini, ‘‘Outlier detection in urban traffic flow distributions,’’ in Proc. IEEE Int. Conf. Data Mining (ICDM), Nov. 2018, pp. 935–940.
[23] V. Chandola, A. Banerjee, and V. Kumar, ‘‘Anomaly detection: A survey,’’ ACM Comput. Surv., vol. 41, no. 3, pp. 1–58, Jul. 2009.
[24] BonduVenkateswarlu, SV Prasad Raju, Performance Analysis And Validation Of Clustering Algorithms Using Soft Computing Technique, International Journal of Computer
[25] Science and Information Technology & Security (IJCSITS),Volume 6,Issue 2,pp. 55 61,2016.
[26] Ekta Maini, BonduVenkateswarlu, Improving the performance of heart disease prediction system usingensembling techniques, AIP Conference Proceedings,Volume 2316,Issue 1, AIP Publishing LLC,2016.
[27] P, Samson AnoshBabu., Annavarapu, C.S.R. Deep learning-based improved snapshot ensemble technique for COVID-19 chest X-ray classification. ApplIntell 51, 3104–3120 (2021). https://doi.org/10.1007/s10489-021- 02199-4
[28] P, Samson AnoshBabu, Annavarapu, C.S.R, Suresh Dara, ‘‘Clustering-based hybrid feature selection approach for high dimensional microarray data’’ Chemometrics and
[29] Intelligent Laboratory Systems,, vol. 213, no. 2021, pp. 104305, Jun. 2021.
[30] SK AlthafHussainBasha,Naga Raju Devarakonda,ShaikSubhani,“OutliersDetectionin
[31] Regression Analysis using Partial Least Square Approach”, ICT and Critical Infrastructure:
[32] proceedings of the 48th Annual Convention of Computer Society of India- Springer, Vol II
[33] Advances in Intelligent Systems andComputing,Volume 249, pp. 125-135,
[34] Visakhapatnam, December 2013,ISBN: 978-3- 319-03095-1.
[35] Sk. Althaf Hussain Basha, Y. Sri Lalitha, Y. Gayatri, and M. V Aditya Nag "Student Performance Prediction – A Data Science Approach",2nd International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA 2020), Goa, India.(Springer Conference August 29-30,2020,)

