Outlier Detection on High-Dimensional Data

Wasan Abd_Majeed*,Murtadha M. Hamad

Department of Computer Science, College of Computer Science and Information Technology, University of Anbar, Ramadi, Iraq

ARTICLE INFO

Received: 04 / 02 /2024
Accepted: 05 / 03 / 2024
Available online: 18 / 06 / 2024

DOI:10.37652/juaps.2024.146609.1191

Keywords:
Outlier detection, High-dimensional data, Curse of dimensionality, Data preprocessing, Machine learning algorithms.
.

ABSTRACT

This paper provides an overview of anomaly detection in high-dimensional datasets, addressing challenges such as the curse of dimensionality. It highlights preprocessing techniques like handling missing data and normalization, and emphasizes the role of machine learning algorithms in anomaly detection. The paper discusses evaluation metrics and popular programming languages like Python for data analysis. It explores various anomaly detection methods including statistical-based, depth-based, deviation-based, distance-based, and density-based approaches, with a focus on deep learning techniques. Overall, the paper emphasizes the need for tailored approaches to extract meaningful insights from high-dimensional data and achieves an informational content rate ranging from 85% to approximately 96%.

Introduction

Outlier detection is a critical process in data analysis that involves identifying and analyzing data points that deviate significantly from the expected patterns or norms in a given dataset.

These outliers, which represent data points that are markedly different from the majority of other data points, can arise due to various factors, including measurement errors, data entry errors, or truly anomalous observations [1].

Numerous sources, such as weblogs, financial transactions, health records, surveillance logs, and the fields of commerce, telecommunication, and biosciences, continuously produce large volumes of data[1].

This phenomenon, which is commonly referred to as "large data," has attracted a great deal of scholarly interest because of the size and dispersion of these datasets. Large data is defined by Gartner [2] as high-volume, high-velocity, and high-variety datasets that require creative and economic data analytics in order to make well-informed decisions and extract useful insights. The main obstacles related to large data have been identified over time and are represented by the five Vs: value, veracity, variety, velocity, and volume [3], as seen in Fig. 1. Value is concerned with the outcomes of data analysis; veracity is the accuracy of the data; and diversity is the range of different kinds of data, including structured formats [4].

The three categories of datasets, which are semi-structured, unstructured, and structured, represent the various levels of organization. The volume of data, or the total amount of information collected, is one crucial

*Corresponding author at : Department of Computer Science, College of Computer Science and Information Technology, University of Anbar ,Ramadi, Iraq:
ORCID:https:./;Tel:+964 07822193376
Email: was21c1005@uoanbar.edu.iq

component; larger datasets are linked to higher dimensionality, or the quantity of characteristics or variables that may be examined. On the other hand, velocity refers to the speed at which data is generated, often involving multiple dimensions [5], as shown in Figure 1:

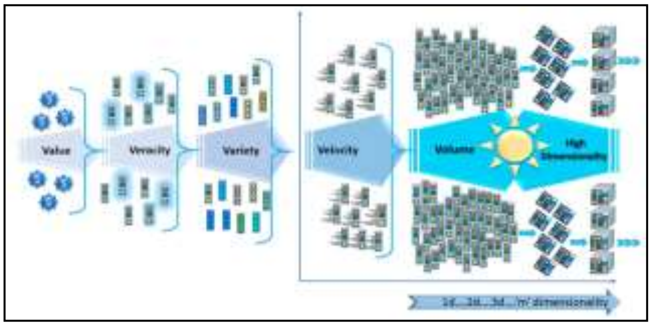

Fig 1. The High Dimensionality Problem In Large Data [4]

Although the commonly accepted definition of large data captures important issues with the five Vs: value, veracity, variety, velocity, and volume [5], but it frequently leaves out a vital element: "dimensionality." Dimensionality, or the total number of components or qualities in the data, is crucial for real-world data analysis [6]. A significant issue posed by the growth in dimensionality is anomaly identification in large-scale databases.

The primary goal of anomaly detection is to locate unusual patterns that deviate from the conventional distribution of data; these patterns are also known as anomalies or outliers. When dealing with large dimensionality, anomaly detection is challenged since the more qualities or features there are, the more data is required for accurate generalization. Data sparsity results from the increased isolation and dispersion of data points as a result. Real anomalies are hidden by the sparsity, which is brought on by extraneous variables or high noise of multiple insignificant qualities. These challenges often lead to a lower efficacy when compared to more traditional. techniques, such as density-, distance-, and clustering-based procedures [7].

This thesis proposes a novel privacy-preserving approach to distance-based outlier detection. In this approach, the dataset is distributed among multiple parties, with the designated data stewards being authorized to utilize the data without disclosing it to others. The main objective is to identify outliers without revealing any information beyond the knowledge that certain items are outliers. The proposed approach involves a computation process where each party contributes their portion of the answer, followed by a secure sum to compute the total distance. To ensure privacy, the total distance is randomly split between the participating sites, ensuring that no party possesses knowledge of the actual distance. Furthermore, a secure protocol is employed to determine if the actual distance between any two points exceeds the predefined threshold.

In summary, the proposed approach aims to strike a delicate balance between the benefits of outlier detection and the imperative to protect privacy. By design, the approach ensures that data is not disclosed beyond what is inherently revealed in the identification of outlier items. This approach holds potential for applications in various domains where outlier detection is essential, while also respecting and addressing privacy concerns.

1. Multi-Dimensional Data

Multi-dimensional data refers to data that has more than one dimension or attribute. In other words, it involves datasets where each data point is described by multiple variables or features. Each dimension represents a specific aspect or characteristic of the data. For example, consider a dataset of houses with features such as price, square footage, number of bedrooms, and number of bathrooms. In this case, the data is multi-dimensional because each house is represented by multiple attributes or dimensions [8].

Multi-dimensional data is commonly encountered in various fields, including but not limited to Data Science and Machine Learning: Many real-world datasets used for tasks like classification, regression, clustering, or recommendation systems are multi-dimensional. These datasets often contain a combination of numerical and categorical features. Geospatial Data: Spatial data, such as maps, satellite imagery, or geographical coordinates, are multi-dimensional, typically represented by latitude, longitude, and potentially additional attributes like elevation or time [9].

Physics and Engineering: Experimental or simulation data in fields like physics, engineering, or biology often involve multiple dimensions. For instance, measurements in particle physics experiments may include attributes like energy, momentum, or particle identification. Analyzing and visualizing multi-dimensional data can be challenging due to the increased complexity. Techniques such as data reduction, feature selection, and dimensionality reduction are commonly used to simplify the data and extract meaningful patterns. Visualization methods like scatter plots, heatmaps, parallel coordinates, or dimensionality reduction techniques (e.g., t-SNE or PCA) can help explore and interpret multi-dimensional data [10].

Overall, multi-dimensional data allows for a richer representation of complex phenomena and provides opportunities for detailed analysis and insight generation across various domains [11].

2. Common Techniques For Outlier Detection

Outlier detection is a challenging task that requires careful consideration of various factors, such as data quality, data distribution, and the specific problem domain. In addition, outlier detection methods need to be able to handle high-dimensional data, streaming data, and other types of data that can be difficult to analyze [12].

One challenge in outlier detection is the definition of outliers. Outliers can be defined in various ways, and the choice of definition can have a significant impact on the performance of outlier detection methods. In addition, outlier detection methods need to be able to handle different types of outliers, such as global outliers, contextual outliers, and collective outliers [13].

Another challenge in outlier detection is the selection of appropriate methods. There are many outlier detection methods available, and each method has its strengths and weaknesses [14].

Therefore, it is important to carefully choose appropriate methods for the specific problem domain and to validate their performance on different datasets [15].

A third challenge in outlier detection is the handling of imbalanced datasets. Outliers are often rare events, and datasets may be imbalanced, with a small proportion of outliers compared to normal data points. This can lead to biased performance evaluations and require the use of specialized methods to handle imbalanced datasets [16].

In addition, outlier detection methods need to be scalable, efficient, and able to handle large-scale datasets. Outlier detection methods also need to be able to handle real-time data streams and be able to detect outliers in a timely manner. Future directions in outlier detection include the development of new methods that can handle high-dimensional data, streaming data, and other types of data that can be difficult to analyze. In addition, there is a need for methods that can handle multiple types of outliers and for methods that can handle imbalanced datasets [17].

Another area of future research is the integration of outlier detection methods with other machine learning techniques, such as clustering, classification, and regression. This can lead to more accurate and robust outlier detection methods that can be used in a wide range of applications [18].

Finally, there is a need for benchmark datasets and evaluation metrics that can be used to compare different outlier detection methods and to validate their performance on different problem domains. This can help advance the field of outlier detection and lead to more effective and reliable methods for outlier detection [19].

3.1 Pre – Processing Stage

In the pre-processing stage of data analysis, one important step is data reduction. Data reduction refers to the process of reducing the size or dimensionality of a dataset while preserving its essential characteristics and minimizing information loss. This is done to improve the efficiency and effectiveness of subsequent data analysis tasks [2].

3.2 Data Reduction Techniques

There are several techniques commonly used for data reduction [3]:

1. Feature Selection: Feature selection aims to select a subset of the most relevant features from the original dataset. This helps reduce the dimensionality of the data by eliminating less informative or redundant features. By selecting the most informative features, it is possible to simplify the analysis process and

improve the performance of machine learning algorithms.

2. Feature Extraction: Feature extraction involves transforming the original set of features into a lower-dimensional representation while preserving the most important information. Techniques like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are commonly used for feature extraction. These methods project the data onto a new set of orthogonal variables that capture the maximum amount of variance or discriminative information.
3. Sampling Techniques: Sampling techniques involve selecting a representative subset of the original dataset. This can be done through methods such as random sampling, stratified sampling, or cluster sampling. By selecting a smaller sample from the original dataset, computational resources can be conserved while still capturing the main characteristics of the data.

3.2. Data Reduction Techniques Benefits

Data reduction techniques have several benefits, including [4]:

1. Improved computational efficiency: By reducing the size or dimensionality of the dataset, data analysis tasks can be performed more efficiently, especially when dealing with large datasets.
2. Removal of noise and irrelevant information: Data reduction techniques can help eliminate noisy or irrelevant features, which can improve the accuracy and interpretability of the analysis results.
3. Mitigation of the curse of dimensionality: High-dimensional datasets often suffer from the curse of dimensionality, where the data becomes sparse, and the performance of learning algorithms deteriorates. Data reduction techniques can help mitigate this issue by reducing the dimensionality and improving algorithmic performance.

Overall, data reduction is an essential step in the pre-processing stage of data analysis. It helps streamline subsequent analysis tasks, improve computational efficiency, and enhance the quality of the analysis results by focusing on the most relevant information in the dataset [5].

3.3 Data Cleaning

Data cleaning plays a pivotal role in the initial stages of data preprocessing, with the overarching goal of ensuring the quality and reliability of the dataset. This multifaceted process encompasses several crucial tasks, including addressing missing values, eliminating duplicates, handling outliers, rectifying inconsistent data, standardizing and formatting information, correcting errors, resolving data integrity issues, and diligently documenting changes [6].

Specifically, managing outliers is imperative to prevent their undue influence on subsequent data analysis and modeling. Various approaches, such as visual inspection or statistical methods, are employed to identify outliers, followed by a thorough assessment of their underlying causes. Strategies for handling outliers encompass the removal or transformation of values, utilization of fissurization or capping techniques, and, when necessary, treating outliers separately. The selection of a particular outlier handling method is contingent upon the specific data context, analysis objectives, and domain knowledge, emphasizing the nuanced nature of this critical data preprocessing step[7].

3.4 Data Integration

Data integration is a comprehensive process that amalgamates data from diverse sources to establish a unified and cohesive perspective. The key facets and techniques involved in this integration endeavor encompass various stages. Beginning with the identification of relevant data sources, including databases, files, and APIs, the process proceeds to schema mapping, aligning data attributes across different source schemas with the target unified schema. Data transformation plays a crucial role, involving the conversion of data formats, standardization of units, and harmonization of data representations. Ensuring data quality is pivotal, achieved through activities such as data cleaning, deduplication, and handling missing values [8].

The culmination involves data consolidation, where transformed and cleaned data are merged into a singular unified dataset. Robust error handling and recovery mechanisms are implemented, and due consideration is given to data governance and metadata management for documentation and organization. Ongoing data integration strategies, whether periodic or real-time, are employed to facilitate synchronization. The holistic impact of data integration is profound, empowering data-driven decision-making, improving data analysis, and enhancing reporting capabilities, ultimately providing a comprehensive understanding of business operations [9].

3.5 Data Transformation

Data transformation, a pivotal stage in both data preprocessing and integration, involves employing various techniques to enhance the quality and utility of the data. These techniques encompass several common practices such as scaling and normalization, which adjust data ranges for uniformity, and logarithmic or exponential transformations, particularly useful for addressing skewed data. Binning is applied to simplify complex patterns, while encoding categorical variables into numerical representations facilitates analysis [10].

Feature engineering is utilized to extract valuable insights, and aggregation aids in summarizing data based on attributes or dimensions. Time-series transformation is implemented for temporal data, and dimensionality reduction serves to decrease the number of variables. The selection of these techniques is contingent upon the characteristics of the data, the objectives of the analysis, and the requirements of downstream tasks. Ultimately, data transformation ensures the production of meaningful and interpretable data, laying the foundation for accurate analyses and model building [3].

4. Outlier Detection Methods, Models, And Classification

Outlier detection is a crucial task in data analysis and machine learning, aimed at identifying observations that deviate significantly from the majority of the data points. Various methods and techniques have been developed to address this challenge, each with its strengths and limitations [14]. This flowchart serves as a guide to understanding and navigating the landscape of outlier detection methodologies. From traditional statistical approaches to cutting-edge deep learning techniques, this flowchart provides an overview of the different paths available for detecting and categorizing outliers [15]. By categorizing outliers into distinct groups and exploring multiple detection strategies, analysts and data scientists can effectively identify anomalies and tailor their mitigation strategies accordingly. The flowchart outlines the key steps and decision points involved in outlier detection, offering a structured approach to handling anomalous data points and ensuring the integrity and reliability of analytical insights derived from the data [16].

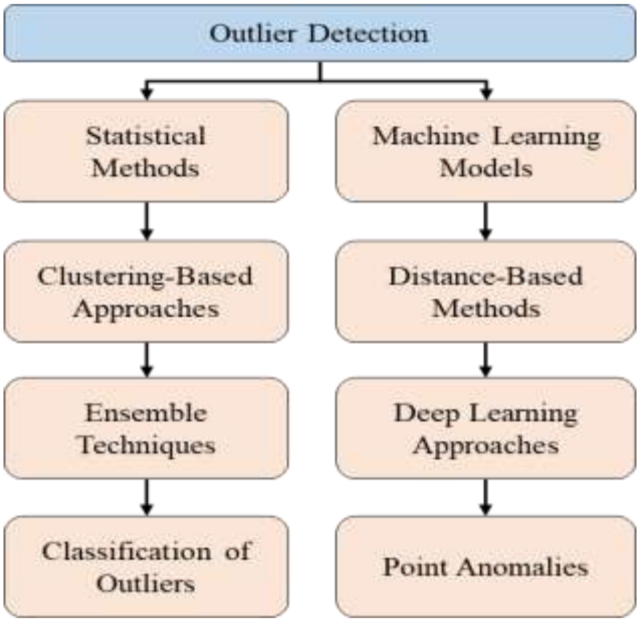

Fig 2. Outlier Detection Methods, Models, And Classification [4]

5. Outlier Detection Algorithms

A widely used outlier detection algorithm is the k-nearest neighbors (k-NN) method, which assesses the outlier status of an observation by evaluating its proximity to its k nearest neighbors within the dataset. The algorithm follows a step-by-step process: first, a value for k, representing the number of neighbors to consider, is selected. Subsequently, the distance between each observation and its k nearest neighbors is calculated and sorted in ascending order. An outlier score is then assigned to each observation based on its distance to the kth nearest neighbor, with higher distances indicating a greater likelihood of being an outlier. To classify observations as outliers or non-outliers, a threshold value or a statistical approach is employed based on their outlier scores. The k-NN algorithm operates under the assumption that outliers are situated far from their neighbors, effectively detecting local outliers with distinct characteristics from neighboring observations [27].

An analogous technique to k-NN is the Local Outlier Factor (LOF) algorithm, which extends the k-NN approach by computing the local density of an observation in comparison to its neighbors. Observations with significantly lower densities than their neighbors are identified as outliers [28].

Beyond k-NN and LOF, several other commonly used outlier detection techniques include Isolation Forest, One-Class Support Vector Machines (SVM), DBSCAN (Density-B Clustering Applications with Noise), Robust CovariGaussian Mixture Models (GMM). For instance, Isolation Forest constructs random forests to isolate outliers, leveraging the intuition that outliers require fewer partitions to stand out. One-Class SVM trains a model on normal observations, classifying new observations based on learned boundaries. DBSCAN identifies outliers as points that do not belong to any cluster or have low-density neighborhoods. Robust Covariance considers the presence of outliers when estimating the covariance matrix, identifying outliers with large Mahalanobis distances. GMMs model data distribution as a mixture of Gaussian components, designating observations with low probabilities under the GMM as outliers. These techniques illustrate the diversity of approaches, each with its own assumptions and characteristics. The selection of a specific algorithm hinges on factors such as dataset specifics, outlier characteristics, and the desired balance between false positives and false negatives in the detection process [29].

The document presents a comprehensive overview of various outlier detection techniques. It begins by discussing clustering-based approaches, which group similar data points into clusters and identify outliers based on deviations from cluster patterns. Examples include DBSCAN, OPTICS, and LOF. Deep learning methods are then explored, leveraging neural networks to detect outliers by learning patterns in high-dimensional data. Statistical-based approaches, such as Z-Score and Grubbs' Test, analyze deviations from expected distributions. Depth-based methods measure the centrality of data points within a dataset, while deviation-based approaches focus on quantifying deviations from central locations. Distance-based methods assess similarity between data points using metrics like Euclidean distance, while density-based techniques identify outliers based on data density. Additionally, under-sampling techniques mitigate class imbalance within datasets. Each method offers unique advantages and considerations, requiring careful parameter tuning and evaluation for effective outlier detection.[45].

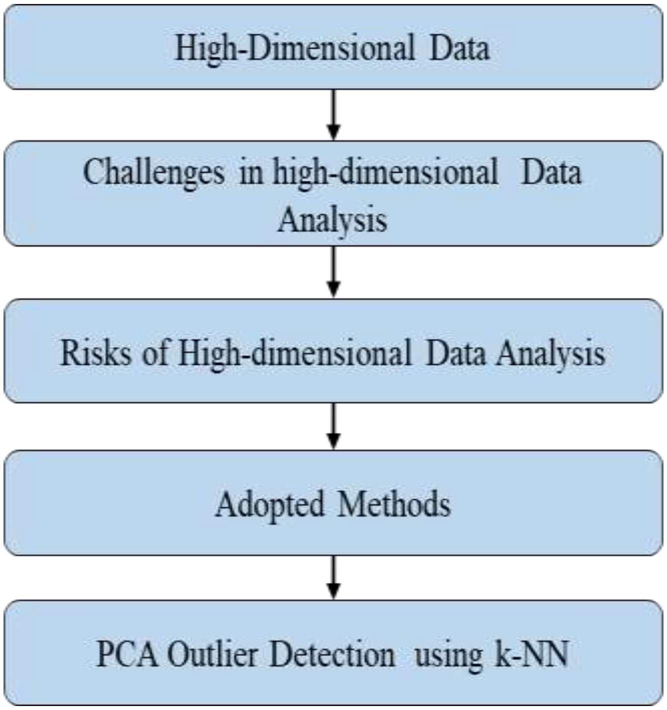
High-Dimensional Data
Challenges in high-dimensional Data Analysis
Risks of High-dimensional Data Analysis
Adopted Methods
PCA Outlier Detection using k-NN

Fig 3. Model Methodology

6. Results

In this section, we apply different outlier detection methods using the calculated distances to the k nearest neighbors. We consider methods such as the average distance method, percentage-based method, and z-score method. Additionally, we provide insights on incorporating domain-specific criteria for identifying outliers.

In this section of the code, we apply different outlier detection methods based on the calculated distances to the k nearest neighbors. We implement the average distance method, percentage-based method, and z-score method. We also provide insights into incorporating domain-specific criteria for identifying outliers. Finally, we visualize the results of each outlier detection method using histograms, helping us understand the distribution of outliers detected by each method, as shown in Figure6:

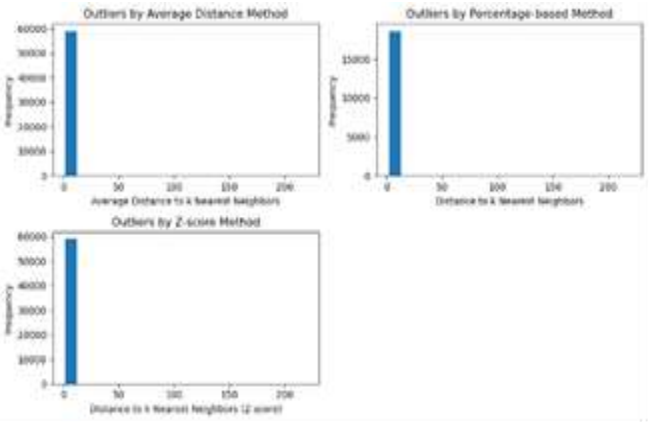

Fig 4. Outlier Detection Methods and Visualization

The PUBG dataset was obtained from Kaggle, containing 1.8 million observations and 25 features. After data cleaning and initial analysis, Principal Component Analysis (PCA) was applied to reduce dimensions while retaining the most variance.

Subsequently, the k-Nearest Neighbors (k-NN) algorithm was applied to the transformed data to calculate the distance to the nearest k neighbors for each data point. Outliers were identified based on the farthest distances.

7. Analyzing Chess Game Data Using Artificial Intelligence Techniques

This study aims to analyze chess game data using artificial intelligence techniques to detect outliers. A dataset containing 20,058 chess matches with 17 features was collected. Exploratory data analysis was conducted to understand the distributions of the numerical variables. Principal Component Analysis (PCA) was applied to reduce dimensions. To detect outliers, the nearest neighbor's algorithm (k-NN) was used to calculate the distance to the nearest k neighbors. The results demonstrated the ability of k-NN with PCA to identify outliers based on distance. This study provides a systematic framework for analyzing chess game data using artificial intelligence techniques.

Table (1) provides a comprehensive overview of the dataset, including the number of instances, number of features, data types of each feature, and summary statistics for key numerical features.

Table 1. Comprehensive Overview Of The Dataset

	game_ id	turns	white_ rating	black_ rating	opening_ moves
count	**20,058**	**20,058**	**20,058**	**20,058**	**20,058**
mean	**10,029.50**	**60.466**	**1,596.63**	**1,588.83**	**4.817**
std	**5,790.39**	**33.571**	**291.25**	**291.04**	**2.797**
min	**1**	**1**	**784**	**789**	**1**
25%	**5,015.25**	**37**	**1,398**	**1,391**	**3**
50%	**10,029.50**	**55**	**1,567**	**1,562**	**4**
75%	**15,043.75**	**79**	**1,793**	**1,784**	**6**
max	**20,058**	**349**	**2,700**	**2,723**	**28**

The dataset comprises 20,058 chess matches, each represented as a case with 17 descriptive features. These features encompass various aspects of the matches, such as turn counts, victory status, player ratings, and opening moves. The data types for these features vary, including integers, Booleans, and objects. Descriptive statistics offer insights into the central **tendencies and variability of key numerical features.** Notably, the dataset shows no identified potential outliers. Furthermore, Principal Component Analysis (PCA) was applied to reduce the dimensionality from 17 features to 5 principal components, providing a more concise representation of the data. Additionally, the k-Nearest Neighbors (K-NN) algorithm was employed for outlier detection based on the distances between data points and their nearest neighbors.

8. Conclusions

This coordinated conversation gives a succinct outline of the advancement in exception identification for non-IID high-layered information, featuring the difficulties and exploration valuable open doors. While both present day and conventional techniques can include information investigation, dimensionality decrease utilizing PCA, and k-NN for anomaly discovery, current strategies are supposed to offer more prominent versatility, refinement, and possibly further developed execution. The decision between these methodologies ought to consider the particular attributes of the dataset and the ideal degree of interpretability.

In rundown, tending to difficulties in exception identification for high-layered, non-free, and evenly circulated information requires a complete comprehension of both conventional and current techniques, with an emphasis on flexibility, vigorous assessment, and taking care of dynamic conditions. The reconciliation of kNN and PCA further improves the general viability of anomaly discovery in complex datasets.

References

[1] S. Aliesawi, C. Tsimenidis, B. S. Sharif, and M. Johnston, Performance comparison of IDMA receivers for underwater acoustic channels. 2010. doi: 10.1109/ISWCS.2010.5624402.

[2] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," Glob. Transitions Proc., vol. 3, no. 1, pp. 91–99, 2022, doi: https://doi.org/10.1016/j.gltp.2022.04.020.

[3] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," Glob. Transitions Proc., vol. 3, no. 1, pp. 91–99, 2022, doi: https://doi.org/10.1016/j.gltp.2022.04.020.

[4] H. Palo, S. Sahoo, and A. Subudhi, "Dimensionality Reduction Techniques: Principles, Benefits, and Limitations," 2021, pp. 77–107. doi: 10.1002/9781119785620.ch4.

[5] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," J. Bus. Res., vol. 70, pp. 263–286, 2017, doi: https://doi.org/10.1016/j.jbusres.2016.08.001.

[6] T. Dasu and T. Johnson, Exploratory Data Mining and Data Cleaning, vol. 101. 2003. doi: 10.1002/0471448354.

[7] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," SIGMOD Rec. (ACM Spec. Interes. Gr. Manag. Data), vol. 30, no. 2, pp. 37–46, 2001, doi: 10.1145/376284.375668.

[8] Y. K. Dwivedi et al., "Opinion Paper: 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," Int. J. Inf. Manage., vol. 71, p. 102642, 2023, doi: https://doi.org/10.1016/j.ijinfomgt.2023.102642.

[9] D. Kaur and D. Singh, "Critical Data Consolidation in MDM to Develop the Unified Version of Truth," Int. J. Adv. Comput. Sci. Appl., vol. 12, Jan. 2021, doi: 10.14569/IJACSA.2021.0121242.

[10] S. Ali et al., "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," Inf. Fusion, vol. 99, p. 101805, 2023, doi: https://doi.org/10.1016/j.inffus.2023.101805.

[11] D. Cousineau and S. Chartier, "Outliers detection and treatment: A review," Int. J. Psychol. Res., vol. 3, Jun. 2010, doi: 10.21500/20112084.844.

[12] C. H. Park, "A Comparative Study for Outlier Detection Methods in High Dimensional Text Data," J. Artif. Intell. Soft Comput. Res., vol. 13, no. 1, pp. 5–17, 2023, doi: 10.2478/jaiscr-2023-0001.

[13] B. Angelin and A. Geetha, Outlier Detection using Clustering Techniques – K-means and K-median. 2020. doi: 10.1109/ICICCS48265.2020.9120990.

[14] G. Orair, C. Teixeira, Y. Wang, W. Meira Jr, and S. Parthasarathy, "Distance-Based Outlier Detection: Consolidation and Renewed Bearing," PVLDB, vol. 3, pp. 1469–1480, Oct. 2010.

[15] M. Re and G. Valentini, "Ensemble methods: A review," in Advances in Machine Learning and Data Mining for Astronomy, 2012, pp. 563–594.

[16] X. Zhang, J. Mu, X. Zhang, H. Liu, L. Zong, and Y. Li, "Deep anomaly detection with self-supervised learning and adversarial training," Pattern Recognit., vol. 121, p. 108234, 2022, doi: https://doi.org/10.1016/j.patcog.2021.108234.

[17] T. Doctoral, D. Series, and C. Science, OUTLIER DETECTION ANALYSIS Data Mining Approaches for Outlier Detection Analysis Shahrooz Abghari. 2020.

[18] T. Falahi, G. Nasserddine, and J. Younis, "Detecting Data Outliers with Machine Learning," Al-Salam J. Eng. Technol., vol. 2, May 2023, doi: 10.55145/ajest.2023.02.02.018.

[19] E. Calikus, S. Nowaczyk, M. R. Bouguelia, and O. Dikmen, Wisdom of the contexts: active ensemble learning for contextual anomaly detection, vol. 36, no. 6. Springer US, 2022. doi: 10.1007/s10618-022-00868-7.

[20] E. Calikus, Together We Learn More : Algorithms and Applications for User-Centric Anomaly

Detection Together We Learn More : Algorithms and Applications for User-Centric Anomaly Detection, no. 89. 2022.

[21]A. Boukerche, L. Zheng, and O. Alfandi, “Outlier Detection: Methods, Models, and Classification,” ACM Comput. Surv., vol. 53, pp. 1–37, Jun. 2020, doi: 10.1145/3381028.

[22]Z. Zhou, G. Si, H. Sun, K. Qu, and W. Hou, “A robust clustering algorithm based on the identification of core points and KNN kernel density estimation,” Expert Syst. Appl., vol. 195, p. 116573, 2022, doi: https://doi.org/10.1016/j.eswa.2022.116573.

[23]S. Naeem, A. Ali, S. Anam, and M. Ahmed, “An Unsupervised Machine Learning Algorithms: Comprehensive Review,” IJCDS J., vol. 13, pp. 911–921, Apr. 2023, doi: 10.12785/ijcds/130172.

[24]Z. Zhao, “Ensemble Methods for Anomaly Detection,” no. December, 2017, [Online]. Available: https://surface.syr.edu/etd/817

[25]L. H. Chiang, R. J. Pell, and M. B. Seasholtz, “Exploring process data with the use of robust outlier detection algorithms,” J. Process Control, vol. 13, no. 5, pp. 437–449, 2003, doi: https://doi.org/10.1016/S0959-1524(02)00068-9.

[26]R. Purohit, J. P. Verma, R. Jain, and M. Bhavsar, “WePaMaDM-Outlier Detection: Weighted Outlier Detection using Pattern Approaches for Mass Data Mining,” 2023 Int. Conf. Adv. Comput. Comput. Technol. InCACCT 2023, pp. 243–248, 2023, doi: 10.1109/InCACCT57535.2023.10141778.

[27]A. Boukerche, L. Zheng, and O. Alfandi, “Outlier Detection: Methods, Models, and Classification,” ACM Comput. Surv., vol. 53, pp. 1–37, Jun. 2020, doi: 10.1145/3381028.

[28]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, pp. 1–24, 2021, doi: 10.3390/bdcc5010001.

[29]H. Ali Mohammadi and S. Chen, “Performance Evaluation of Outlier Detection Techniques in Production Timeseries: A Systematic Review and Meta-Analysis,” Expert Syst. Appl., vol. 191, p. 116371, Dec. 2021, doi: 10.1016/j.eswa.2021.116371.

[30]M. B. Al- Zoubi, “An effective clustering-based approach for outlier detection,” Eur. J. Sci. Res., vol. 28, pp. 310–316, Jan. 2009.

[31]M. Hahsler, M. Piekenbrock, and D. Doran, “Dbscan: Fast density-based clustering with R,” J. Stat. Softw., vol. 91, no. 1990, 2019, doi: 10.18637/jss.v091.i01.

[32]Z. He, X. Xu, and S. Deng, “Discovering cluster-based local outliers,” Pattern Recognit. Lett., vol. 24, no. 9, pp. 1641–1650, 2003, doi: https://doi.org/10.1016/S0167-8655(03)00003-5.

[33]M. Landauer, S. Onder, F. Skopik, and M. Wurzenberger, “Deep learning for anomaly detection in log data: A survey,” Mach. Learn. with Appl., vol. 12, p. 100470, 2023, doi: https://doi.org/10.1016/j.mlwa.2023.100470.

[34]J. Saeedi and A. Giusti, “Semi-supervised visual anomaly detection based on convolutional autoencoder and transfer learning,” Mach. Learn. with Appl., vol. 11, p. 100451, 2023, doi: https://doi.org/10.1016/j.mlwa.2023.100451.

[35] Ahmed Subhi Abdalkafor, Salah A. Aliesawi., Data aggregation techniques in wireless sensors networks (WSNs): Taxonomy and an accurate literature survey. AIP Conference Proceedings 2400, 020012 (2022); 2022.

[36]M. Landauer, S. Onder, F. Skopik, and M. Wurzenberger, “Deep learning for anomaly detection in log data: A survey,” Mach. Learn. with Appl., vol. 12, p. 100470, 2023, doi: https://doi.org/10.1016/j.mlwa.2023.100470.

[37]B. Dastjerdy, A. Saeidi, and S. Heidarzadeh, “Review of Applicable Outlier Detection Methods to Treat Geomechanical Data,” Geotechnics, vol. 3, no. 2, pp. 375–396, 2023, doi: 10.3390/geotechnics3020022.

[38]C. H. Sim, F. F. Gan, and T. C. Chang, “Outlier Labeling With Boxplot Procedures,” J. Am. Stat. Assoc., vol. 100, pp. 642–652, Feb. 2005, doi: 10.1198/016214504000001466.

[39]X. Dang and R. Serfling, “Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties,” J. Stat. Plan. Inference, vol. 140, no. 1, pp. 198–213, 2010, doi: https://doi.org/10.1016/j.jspi.2009.07.004.

[40]J. Huang, D. Cheng, and S. Zhang, “A novel outlier

detecting algorithm based on the outlier turning points," Expert Syst. Appl., vol. 231, p. 120799, 2023, doi: https://doi.org/10.1016/j.eswa.2023.120799.

[41]J. Huang, D. Cheng, and S. Zhang, "A novel outlier detecting algorithm based on the outlier turning points," Expert Syst. Appl., vol. 231, p. 120799, 2023, doi: https://doi.org/10.1016/j.eswa.2023.120799.

[42]F. Yoseph, M. Heikkila, and D. Howard, "Outliers identification model in point-of-sales data using enhanced normal distribution method," Proc. - Int. Conf. Mach. Learn. Data Eng. iCMLDE 2019, pp. 72–78, 2019, doi: 10.1109/iCMLDE49015.2019.00024.

[43]W. P. Bensken, F. M. Pieracci, and V. P. Ho, "Basic Introduction to Statistics in Medicine, Part 1: Describing Data," Surg. Infect. (Larchmt)., vol. 22, no. 6, pp. 590–596, 2021, doi: 10.1089/sur.2020.429.

[44]R. Dyckerhoff and P. Mozharovskyi, "Exact computation of the halfspace depth," Comput. Stat. Data Anal., vol. 98, pp. 19–30, 2016, doi: 10.1016/j.csda.2015.12.011.

[45]E. Blanco-Mallo, L. Morán-Fernández, B. Remeseiro, and V. Bolón-Canedo, "Do all roads lead to Rome? Studying distance measures in the context of machine learning," Pattern Recognit., vol. 141, p. 109646, 2023, doi: https://doi.org/10.1016/j.patcog.2023.109646.

[46]M. Pachgade and S. Dhande, "Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach," IJARCSSE, vol. 2, pp. 1–5, Jun. 2012.

[47]A. Lotfi, P. Moradi, and H. Beigy, "Density peaks clustering based on density backbone and fuzzy neighborhood," Pattern Recognit., vol. 107, p. 107449, 2020, doi: https://doi.org/10.1016/j.patcog.2020.107449.

[48]R. Bhuyan and S. Borah, "A survey of some density based clustering techniques," arXiv Prepr. arXiv2306.09256, vol. 1, 2023, doi: 10.13140/2.1.4554.6887.

[49]Y. Zhu, K. M. Ting, Y. Jin, and M. Angelova, "Hierarchical clustering that takes advantage of both density-peak and density-connectivity," Inf. Syst., vol. 103, p. 101871, 2022, doi: https://doi.org/10.1016/j.is.2021.101871.

[50]A. D. Outlier and I. Using, "Datasets for Validation of Stroke Clinical Outcomes," pp. 1–23, 2020, doi: 10.1016/j.ijmedinf.2019.103988.Applying.

[51]A. Guzmán-Ponce, J. S. Sánchez, R. M. Valdovinos, and J. R. Marcial-Romero, "DBIG-US: A two-stage under-sampling algorithm to face the class imbalance problem," Expert Syst. Appl., vol. 168, p. 114301, 2021, doi: https://doi.org/10.1016/j.eswa.2020.114301.

[52]A. Guzmán-Ponce, J. S. Sánchez, R. M. Valdovinos, and J. R. Marcial-Romero, "DBIG-US: A two-stage under-sampling algorithm to face the class imbalance problem," Expert Syst. Appl., vol. 168, p. 114301, 2021, doi: https://doi.org/10.1016/j.eswa.2020.114301.

[53]E. Debie and K. Shafi, "Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses," Pattern Anal. Appl., vol. 22, no. 2, pp. 519–536, 2019, doi: 10.1007/s10044-017-0649-0.

[54]P. Hajibabaee, F. Pourkamali-Anaraki, and M. A. Hariri-Ardebili, "Dimensionality reduction techniques in structural and earthquake engineering," Eng. Struct., vol. 278, p. 115485, 2023, doi: https://doi.org/10.1016/j.engstruct.2022.115485.

[55] Salah A. Aliesawi, Dena S. Alani, Abdullah M. Awad, " Secure image transmission over wireless network", International Journal of Engineering and Technology (UAE), Science Publishing Corporation Inc., Qatar, 7 (4) 2758-2764, 2018.

الكشف الخارجي عن البيانات عالية الأبعاد

وسن عبدالمجيد ، مرتضى محمد حمد

قسم علوم الحاسوب ، كلية علوم الحاسبات وتكنلوجيا المعلومات ،جامعة الانبار، الرمادي، العراق.
was21c1005@uoanbar.edu.iq

الخلاصة:

في مجال فحص المعلومات، تمثل مجموعات البيانات عالية الطبقات صعوبات استثنائية تتطلب منهجيات محددة للتعرف على الشذوذ. تقدم هذه الورقة مخططًا موجزًا لاكتشاف الاستثناءات في المعلومات ذات الطبقات العالية، ومعالجة الصعوبات ذات الصلة وتقديم نطاق من الإجراءات للتعامل معها حقًا. إن المعلومات ذات الطبقات العالية، التي تصورها ثروة من العناصر المتناقضة مع التصورات، تمثل "آفة الأبعاد". ويؤدي هذا الأمر إلى زيادة التعقيد الحسابي، وتناثر المعلومات، والتحديات في التمثيل والترجمة. لمحاربة هذه المشكلات، تعد أساليب المعالجة المسبقة المحددة أمرًا أساسيًا، بما في ذلك الاهتمام بالمعلومات المفقودة، والتوحيد القياسي، والتطبيع. تلعب خوارزميات التعلم الآلي دورًا أساسيًا في تحديد الاستثناءات. تقدم هذه الورقة تجارب حول الأسس الافتراضية لحسابات الذكاء الاصطناعي ذات الصلة بهذه المهمة. كما أنه يبحث في قياسات التقييم لتقييم عرض تقنيات اكتشاف الاستثناءات ويتميز بلهجات البرمجة المعروفة المستخدمة في هذا. ويتم عرض الأساليب القائمة على التعلم العميق لتحديد الشذوذ، وذلك باستخدام قوة منظمات الدماغ لفصل الأمثلة المعقدة عن المعلومات ذات الطبقات العالية. تتناول هذه الورقة الاستراتيجيات القابلة للقياس، والعمق، والانحراف، والمسافة، والسمك، مما يكشف عن نظرة ثاقبة لتطبيقاتها وفوائدها. علاوة على ذلك، ظهرت لغة بايثون كلغة القرار في التحقيق في المعلومات بسبب دعمها القوي للمنطقة المحلية، والمكتبات الغنية، والتوثيق الشامل، والقدرة على التكيف عبر مساحات مختلفة. في الملخص، تقدم هذه الورقة تحقيقًا شاملاً لموقع الاستثناء في المعلومات ذات الطبقات العالية، ومعالجة الصعوبات المرتبطة بآفة الأبعاد. مع إجراءات المعالجة المسبقة الخاصة والاعتماد على الذكاء الاصطناعي وحسابات التعلم العميقة، تستكشف الورقة استراتيجيات مختلفة، وتعرض تطبيقاتها. ظهرت لغة بايثون باعتبارها اللغة المفضلة لبيئتها القوية. يسلط هذا المساعد الموحد الضوء على الحاجة إلى طرق مخصصة للتعامل مع استنتاج أجزاء كبيرة من المعرفة، ويقدم مخططًا شاملاً مع معدل مادة تعليمية يتراوح من 85% إلى 96% تقريبًا.