applied sciences

MDPI

Article

Interpretable Single-dimension Outlier Detection (ISOD): An Unsupervised Outlier Detection Method Based on Quantiles and Skewness Coefficients

Yuehua Huang [1,2,*], Wenfen Liu [1,2], Song Li [1], Ying Guo [1] and Wen Chen [1]

[1] School of Computer Science and Information Security & School of Software Engineering, Guilin University of Electronic Technology, Guilin 541004, China
[2] Guangxi Key Laboratory of Cryptography and Information Security, Guilin 541004, China
* Correspondence: 1802102009@mails.guet.edu.cn

Abstract: A crucial area of study in data mining is outlier detection, particularly in the areas of network security, credit card fraud detection, industrial flaw detection, etc. Existing outlier detection algorithms, which can be divided into supervised methods, semi-supervised methods, and unsupervised methods, suffer from missing labeled data, the curse of dimensionality, low interpretability, etc. To address these issues, in this paper, we present an unsupervised outlier detection method based on quantiles and skewness coefficients called ISOD (Interpretable Single dimension Outlier Detection). ISOD first fulfils the empirical cumulative distribution function before computing the quantile and skewness coefficients of each dimension. Finally, it outputs the outlier score. This paper's contributions are as follows: (1) we propose an unsupervised outlier detection algorithm called ISOD, which has high interpretability and scalability; (2) massive experiments on benchmark datasets demonstrated the superior performance of the ISOD algorithm compared with state-of-the-art baselines in terms of ROC and AP.

Keywords: outlier detection; quantile; skewness coefficient; unsupervised

Citation: Huang, Y.; Liu, W.; Li, S.; Guo, Y.; Chen, W. Interpretable Single-dimension Outlier Detection (ISOD): An Unsupervised Outlier Detection Method Based on Quantiles and Skewness Coefficients. *Appl. Sci.* **2024**, *14*, 136. https://doi.org/10.3390/app14010136

Academic Editors: Katia Lida Kermanidis, Phivos Mylonas and Manolis Maragoudakis

Received: 24 November 2023
Revised: 20 December 2023
Accepted: 21 December 2023
Published: 22 December 2023

1. Introduction

Outlier detection, sometimes referred to as novelty detection, is the process of finding out what is different from normal data. According to Aggarwal, "outliers are also referred to as abnormalities, discordants, deviants, or anomalies in the data mining and statistics literature" [1].

Outlier detection has been an important field of research for industry and academia. By identifying outliers, researchers can obtain vital knowledge that assists in making better decisions or avoiding risks. So, outlier detection is widely used in many fields, such as network intrusion detection [2–5], intelligent transportation [6–9], video content analysis and detection [10–12], fraud detection [13–15], social media analysis [16–19], and data generation [20,21].

Over the past few decades, many outlier detection algorithms have been proposed [20,22–25]; depending on whether labeled data are utilized, they can be divided into three main categories: (1) supervised methods, (2) semi-supervised methods, and (3) unsupervised methods. We will provide more details on these methods in Section 2.

While these algorithms were shown to be effective in earlier applications, as the concept of big data has become more prevalent and data have become more multidimensional, they have increasingly become more problematic.

(1) Missing labeled data. Supervised algorithms require a large amount of labeled data that, in many cases, are difficult to implement or require incurring high costs. This can lead to unsatisfactory performance being demonstrated by these supervised algorithms.

(2) Curse of Dimensionality. In the era of big data, the dimensionality of data is increasing. The performance of supervised outlier detection algorithms, especially those based on proximity, will decrease rapidly with the increasing data dimensionality.

(3) Interpretability. In practical applications of anomaly detection, such as credit card fraud detection and medical imaging inspection for anomalies, we not only need to be able to detect anomalous data but also need to make a reasonable explanation as to why these data are anomalous. Due to the disparity in the distribution of outliers and normal instances, the peculiarities of various detection algorithms, and the complexity of data structures in particular applications, it might be challenging to explain abnormalities in outliers.

To avoid the above shortcomings, this paper proposes a new outlier detection algorithm based on quantiles and skewness coefficients: Interpretable Single dimension Outlier Detection (abbreviated as ISOD). In this method, the empirical cumulative distribution function of each sample's dimension is first determined using data from that sample; the skewness coefficient and quantile of the empirical cumulative distribution function are then computed. Finally, the skewness coefficient is used as a weight to summarize the anomaly score of that data point so that anomaly detection results can be obtained.

The rest of this paper is organized as follows: In Section 2, an overview of the current anomaly detection algorithms is provided. In Section 3, the focus is on the proposed algorithm (ISOD) and its analysis. In Section 4, the experiments we conduced and their results are analyzed, and Section 5 concludes the article.

2. Related Works

In data mining, training data are used to train models, and test data are used to measure performance. Based on the availability of labels, anomaly detection methods can be classified into supervised, semi-supervised, and unsupervised methods.

2.1. Supervised Methods

The availability of a training dataset with labeled cases for both normal and anomalous classes is assumed by supervised techniques. The training and testing datasets must then be chosen to perform cross-validation. The training dataset is modeled using a supervised learning technique. Creating a prediction model for normal vs. anomalous classes is a common strategy in these situations. The model is then assessed using the testing dataset.

The representative supervised method is a classification-based anomaly detection algorithm. A classifier is trained by using the labels in the training dataset so that it can distinguish between normal and abnormal data. Once this classifier is trained, it can accurately distinguish between normal data and abnormal data when facing new data.

Many algorithms exist for supervised anomaly detection in health monitoring [26], equipment failure detection, system state monitoring, and other industrial applications [27–30].

A detailed description of these methods can be found in [22]. The merits of supervised methods include the fact that they are (1) supervised, which means they are easy to use, and (2) robust to different data types. However, their shortcomings are obvious. Namely, (1) labeled data are difficult to obtain or require large costs to obtain, especially in industrial and commercial applications; also, (2) the abnormal result finally obtained by the supervised algorithm is binary, and the degree of abnormality cannot be further compared.

2.2. Semi-Supervised Methods

The main difference between semi-supervised anomaly detection algorithms and supervised anomaly detection algorithms is that not all data are labeled. In other words, a portion of the data has a label indicating whether it is normal data or abnormal data. But other data are unlabeled. The typical technique used in semi-supervised methods is to build a model for the class corresponding to normal behavior and use the model to identify anomalies in the test data.

A detailed description of these methods can be found in [22]. In semi-supervised anomaly detection, class support vector machines and support vector data description are widely used [31,32].

Such techniques may not always be widely used because, with their use, it is difficult to obtain a dataset that covers all possible anomalies. Even if there is such a dataset, the data will change over time, and abnormal data that have not appeared before may appear. Therefore, when there are no historical anomaly data, unsupervised measures can be used as a preliminary strategy for anomaly detection.

2.3. Unsupervised Methods

Unsupervised methods are the most extensively used methods since they do not need labeled training data. Modeling the training dataset is carried out using an unsupervised learning approach.

The underlying premise used by the strategies in this category is that regular cases in the test data are significantly more common than abnormalities. If this presumption is incorrect, the effectiveness of these methods will drop dramatically.

By employing a portion of the unlabeled dataset as training data, many semi-supervised approaches may be modified to function in an unsupervised manner. This kind of adaptation is predicated on the test data having relatively few abnormalities and the model to be trained being resilient to those anomalies.

A detailed description of these methods can be found in [22]. Isolation Forest is a representative for this type of method [33].

Some researchers have applied unsupervised anomaly detection for health testing [34] and time-series anomaly detection [35]. A novel method based on mutual information and reduced spectral clustering was developed in [36].

The advantage of this type of method is that it can perform anomaly detection without label data, which means it is more suitable in most situations; its main disadvantages are that the interpretability is relatively poor, and the decision process is less direct and more obscure than supervised anomaly detection, especially when using artificial intelligence technologies such as neural networks and deep learning.

2.4. Self-Supervised Methods

Self-supervised learning (SSL) is an AI-based method of training algorithmic models on raw, unlabeled data. Using various methods and learning techniques, self-supervised models create labels and annotations during the pre-training stage, aiming to iteratively achieve an accurate ground truth so a model can go into production.

Some self-supervised methods have been developed for outlier detection [37–39]. A detailed description of these methods can be found in [40].

Automating Feature Subspace Exploration, a preprocessing step in machine learning for improving outlier detection, was developed in [41].

3. Proposed Algorithm

3.1. Preliminaries

3.1.1. Quantiles

A quantile defines a particular part of a dataset; i.e., a quantile determines how many values in a distribution are above or below a certain limit. Special quantiles include the quartile (quarter), the decile (tenth), and percentiles (hundredth).

Although the term “quantile” lacks a uniform meaning, it is widely used to describe the proportion of values in data collection scores that are less than a particular number. A quantile shows how a given value compares to others. For example, if a value is in the kth percentile, it is greater than K percent of the total values.

$$Q_x = \frac{n_x}{n} \tag{1}$$

In Formula (1), n_x represents the number of values below x, n represents the total number of scores, and P_x represents the quantile of the data x.

3.1.2. Skewness Coefficient

The skewness coefficient is one way to measure the skewness of a distribution, a measure of a probability distribution's asymmetry. A distribution is said to be skewed if its curve is twisted either toward the left or the right. Karl Pearson's coefficient of skewness is the most significant measure of skewness. It is sometimes referred to as Pearson's skewness coefficient.

When a dataset's skewness is measured, it typically takes the form of a bell curve. The skewness of normal distributions is zero. As a result, the distribution becomes symmetrical concerning the mean. Still, there are situations in which skewness is not symmetric. It can be either positive or negative in these circumstances.

When a distribution's tail is more prominent on the right than the left, it is said to be positively skewed. The skewness coefficient is assumed to be positive since the distribution is positive. The majority of the values thus turn out to be to the left of the mean. This indicates that the values on the right side are the most extreme.

Negative skewness, on the other hand, occurs when the tail is more pronounced on the left rather than the right side. Contrary to positive skewness, most of the values are found on the right side of the mean in negative skewness. As such, the most extreme values are found to be further to the left.

Formula (2) describes how to calculate the skewness coefficient.

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^3}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2\right]^{\frac{3}{2}}} \tag{2}$$

In Formula (2), $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sometimes expressed as EX).

3.2. Definition of Outlier Detection

Outlier detection, without supervision, employs some criteria to find outlier candidates which deviate from major normal points. We have n data points $X_1, X_2, \ldots, X_n \in R^d$, which are sampled independently and identically distributed. We use the matrix $\boldsymbol{X} \in R^{n \times d}$ as the notation of the entire dataset, which is formed by stacking each data point's vectors as rows. After giving $\boldsymbol{X}$, an outlier detector obtains an outlier score $o_i \in R$ for each data point x_i, $1 \leq i \leq n$. Data points with higher outlier scores are more likely to be outliers.

3.3. The Proposed ISOD Algorithm

3.3.1. Construct the Empirical Cumulative Distribution Function

Anomaly detection is carried out to find data points in areas with less probability of occurrence in the data distribution. In the univariate normal distribution model, the degree of anomaly can be determined by the ratio of its distance to the mean and its variance. Starting from this idea, we can calculate the degree of anomaly in each dimension of the multivariate probability distribution and finally determine its anomaly score.

In each dimension, the data can be arranged from small to large to construct an empirical cumulative distribution function.

3.3.2. Compute the Quantiles

In the dataset $\boldsymbol{X} \in R^{n \times d}$, we use X_i $1 \leq i \leq n$ as a data sample, and X^j $1 \leq j \leq d$ is used as the j-th dimension of $\boldsymbol{X}$. Therefore, we use X_i^j as the j-th entry of data X_i.

According to Formula (1), we compute the quantile of X_i^j through Formula (3).

$$Q_{ij} = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\left\{X_k^j \leq X_i^j\right\} \quad \forall 1 \leq i \leq n, 1 \leq j \leq d \tag{3}$$

where $\mathbb{I}\{\cdot\}$ is an indicator function that is 1 when its argument is true and 0 when otherwise.

3.3.3. Compute the Skewness Coefficient

According to Formula (2), we compute the skewness coefficient of X^j $1 \leq j \leq d$ through Formula (4).

$$\gamma_j = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i^j - \overline{X^j}\right)^3}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i^j - \overline{X^j}\right)^2\right]^{\frac{3}{2}}} \tag{4}$$

where $\overline{X^j} = \frac{1}{n}\sum_{i=1}^{n} X_i^j$ is the mean of the j-th feature.

3.3.4. Obtain the Outlier Scores

Finally, we obtain an outlier score for each X_i through Formula (5).

$$\begin{gathered} O_i = \sum_{j=1}^{d} o_{ij} \\ o_{ij} = -\log_2 Q_{ij} \times \left(-\gamma_j\right) \text{ when } \gamma_j < 0 \\ o_{ij} = -\log_2\left(1 - Q_{ij}\right) \times \gamma_j \text{ when } \gamma_j > 0 \end{gathered} \tag{5}$$

We use γ_i as the weighting factor when calculating the anomaly score of each data point. We use o_{ij} to represent the abnormality degree of each dimension.

3.3.5. Pseudocode of ISOD

Based on the above steps, the pseudocode of the ISOD algorithm is given in Algorithm 1.

Algorithm 1: ISOD

Input: $\boldsymbol{X} = \left[x_{ij}\right]_{n \times d}$ with n samples and d features
Output: Outlier scores $\{O_1, O_2, \ldots, O_i, \ldots, O_n\}$
1. **for** each dimension $1 \leq j \leq d$:
2. calculate the quantile of each data in this dimension

$$Q_{ij} = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\left\{X_k^j \leq X_i^j\right\} \quad \forall 1 \leq i \leq n, 1 \leq j \leq d$$

3. calculate the skewness coefficient for each dimension:

$$\gamma_j = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i^j - \overline{X^j}\right)^3}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i^j - \overline{X^j}\right)^2\right]^{\frac{3}{2}}}$$

4.**end for**
5.**for** each data X_i $1 \leq i \leq n$:
6. calculate the anomaly score for each dimension

$$o_{ij} = -\log_2 Q_{ij} \times \left(-\gamma_j\right) \text{ when } \gamma_j < 0$$

$$o_{ij} = -\log_2\left(1 - Q_{ij}\right) \times \gamma_j \text{ when } \gamma_j > 0$$

7. calculate outlier score for X_i:

$$O_i = \sum_{j=1}^{d} o_{ij}$$

8.**end for**
9.**Return** $\{O_1, O_2, \ldots, O_i, \ldots, O_n\}$ **while** $\{O_1 \geq O_2 \geq \ldots \geq O_i \geq \ldots \geq O_n\}$.

3.4. Properties of ISOD

3.4.1. Time Complexity Analysis

According to Formulas (3) and (4), calculating the quantiles and skewness coefficients for all d dimensions using n samples leads to $\mathcal{O}(nd)$ time complexity. Similarly, according

to Formula (5), calculating the anomaly score for d dimensions using n samples also leads to $\mathcal{O}(nd)$ time complexity. Therefore, the overall time complexity of ISOD is $\mathcal{O}(nd)$.

3.4.2. Interpretability

Interpretability is an important aspect of the practical applications of anomaly detection. In network attack detection, for example, finding an anomaly is as important as identifying the cause of the anomaly. An algorithm with high interpretability has greater reliability, which not only means that it can provide a result but also the reason(s) behind such a result, which is good for improving the performance of the system and assisting in decision making. Therefore, interpretability is very important in the application of anomaly detection.

As can be seen from Formula (5), the ISOD algorithm aggregates the anomalies on each dimension to determine the final anomaly score. Where necessary, we can give the anomalies of the anomalous data in each dimension, which helps the expert to further identify the dimension in which the anomaly occurs. This involves taking anomaly detection from a "black box" to a "white box".

3.4.3. Sensitivity Analysis

As can be seen from the description of the algorithmic process in Section 3.3 above, the ISOD algorithm independently calculates the skewness coefficients for each dimension as weights to be combined with the quantiles in each dimension. Therefore, there are no special requirements on the distribution of data, slight data noise, or the percentage of outliers. Therefore, we can confidently say that the ISOD algorithm is a robust anomaly detection algorithm that is insensitive to data noise, and this property will have a positive impact on its practical application.

3.4.4. Hyperparameter-Free and Unsupervised

The ISOD algorithm is an easy-to-understand unsupervised anomaly detection algorithm with the following advantages: (1) The ISOD algorithm is a statistic-based algorithm that calculates the anomalies in each dimension and aggregates them to obtain final anomaly scores for the sample data. Therefore, the algorithm has no hyperparameters, and no parameter tuning is required. (2) The algorithm is an unsupervised algorithm that does not need to prepare a large amount of labeled data for training, which gives the algorithm high interpretability and, at the same time, lays a better foundation for the practical applications of the algorithm.

4. Experimental Results and Discussion

4.1. Performance Evaluation Metrics

4.1.1. ROC (Receiver Operating Characteristic)

The receiver operating characteristic (ROC) curve is frequently used for evaluating the performance of binary classification algorithms. It provides a graphical representation of a classifier's performance rather than a single value like most other metrics. The closer the ROC is to 1, the more effective that detection model is. This algorithm's ROC is equal to or lower than 0.5, which means that the inspection model has no value for use.

4.1.2. AP (Average Precision)

Another way to evaluate outlier detection models is to use the average precision (AP). The AP measures the average precision across all possible thresholds, with a higher value indicating a better model. The AP is more suitable for outlier detection problems with rare anomalies or imbalanced data, as it focuses more on the positive class (anomalies) than the negative class (normal instances). However, it may not reflect the overall accuracy or specificity of the model, as it does not account for the true negatives or false negatives. Evaluating outlier detection models can be challenging, especially when you do not have labeled data or ground truth data to compare with. One of the possible ways to evaluate

outlier detection models is to use external validation, which involves comparing the results with some other sources of information, such as domain experts, feedback, or historical data.

Overall, 30% of the data in experiments is reserved for testing, while the remaining 70% is used for training. The area under the receiver operating characteristic (ROC) and average precision (AP) are used to obtain the average score from ten separate trials to assess performance.

4.2. Experimental Settings

4.2.1. Experimental Environment and Baselines

In subsequent experiments, a Windows personal computer with AMD Ryzen 7 5800H CPU and 16G of memory will be used.

We compared the performance of the ISOD algorithm with eight state-of-the-art outlier detection algorithms. These eight outlier detection algorithms—k Nearest Neighbor (KNN) [42], Local Outlier Factor(LOF) [43], Isolation Forest (IForest) [33], Clustering-Based Local Outlier Factor (CBLOF) [44], Locally Selective Combination in Parallel Outlier Ensembles(LSCP) [45], One-Class Support Vector Machines (OCSVMs) [46], Deep Isolation Forest (DIF) [47], and GANomaly [48].

4.2.2. Dataset

To validate the effectiveness of the proposed method, we conducted a series of comparative experiments on ten real-world datasets with different types and sizes. They were collected from several domains and are available on the OODS website (https://odds.cs.stonybrook.edu/, accessed on 20 October 2023). These 10 datasets have been frequently used by researchers to evaluate the performance of anomaly detection methods.

Table 1 shows the 10 datasets from the OODS website with the highest dimensions which were selected for our study.

Table 1. Ten real-word benchmark datasets.

Dataset	Number of Samples	Number of Dimensions	Outliers (%)
Musk	3062	166	97 (3.2%)
Satimage-2	5803	36	71 (1.2%)
Letter Recognition	1600	32	100 (6.25%)
Speech	3686	400	61 (1.65%)
Satellite	6435	36	2036 (32%)
Arrhythmia	452	274	66 (15%)
Ionosphere	351	33	126 (36%)
Mnist	7603	100	700 (9.2%)
Optdigits	5216	64	150 (3%)
Heart	224	44	10 (4.4%)

4.3. Experimental Results

In this section, we give the experimental results of ISOD for the benchmark datasets in Tables 2 and 3. The highest ROC or AP score is marked in bold, which means that the algorithm achieves the best performance for this dataset.

4.3.1. Analysis of Experimental Results

The proposed ISOD algorithm achieved the best performance, with an average ROC of 0.813 and an average precision of 0.75. As shown in Table 2, the ISOD algorithm achieved the highest ROC in 6 of the 10 datasets. Additionally, as shown in Table 3, the ISOD algorithm achieved the highest AP (average precision) in 6 of the 10 datasets.

It is worth noting that, by analyzing the data in Tables 2 and 3, it can be found that the higher the data dimensionality, the better results the ISOD algorithm can achieve, as exemplified by the results for the Speech, Satellite, and Arrhythmia datasets. This confirms that the ISOD algorithm has low time complexity and good performance when working with data with high dimensionality, as noted in Section 3.4.1.

Table 2. ROC scores in terms of outlier detector performance (the highest ROC scores are marked in bold).

Dataset	KNN	LOF	IForst	CBLOF	LSCP	OCSVM	DIF	GANomaly	ISOD
Musk	0.73	0.86	0.709	0.755	0.472	0.111	0.874	0.248	**0.956**
Satimage-2	0.144	0.19	0.557	0.28	0.77	0.274	0.535	**0.929**	0.775
Letter Recognition	0.364	0.121	0.105	0.202	0.2	0.323	0.486	0.152	**0.797**
Speech	0.755	0.406	0.278	0.918	0.981	0.58	0.736	0.692	**0.833**
Satellite	0.625	0.393	0.475	0.664	0.837	0.593	0.22	0.58	**0.9**
Arrhythmia	0.769	0.433	0.168	0.798	0.297	0.171	**0.91**	0.324	0.538
Ionosphere	0.238	0.846	0.272	0.343	0.605	0.867	0.127	0.165	**0.912**
Mnist	0.476	0.454	0.115	0.369	0.418	**0.904**	0.787	0.603	0.808
Optdigits	0.794	0.478	0.798	0.518	0.264	0.896	**0.924**	0.67	0.747
Heart	0.705	0.331	0.488	0.115	0.311	0.56	0.355	0.429	**0.863**
Average ROC	0.56	0.451	0.397	0.496	0.516	0.528	0.545	0.479	**0.813**

Table 3. Average precision (AP) scores in terms of outlier detector performance (the highest AP scores are marked in bold).

Dataset	KNN	LOF	IForst	CBLOF	LSCP	OCSVM	DIF	GANomaly	ISOD
Musk	0.589	0.756	0.115	0.397	0.432	0.143	0.272	0.66	**0.814**
Satimage-2	0.188	0.321	0.786	0.146	0.154	0.315	0.494	0.807	**0.856**
Letter Recognition	0.257	0.474	0.948	0.52	**0.964**	0.885	0.802	0.345	0.754
Speech	0.536	0.214	0.195	0.138	0.32	0.164	0.349	0.822	**0.996**
Satellite	0.281	0.794	0.321	0.204	0.633	0.737	**0.962**	0.888	0.887
Arrhythmia	0.207	0.103	0.612	0.534	0.407	0.338	0.422	0.551	**0.663**
Ionosphere	0.929	**0.985**	0.979	0.138	0.827	0.88	0.715	0.741	0.566
Mnist	**0.744**	0.271	0.679	0.592	0.604	0.274	0.433	0.605	0.392
Optdigits	0.264	0.656	0.359	0.221	0.394	0.261	0.553	0.112	**0.776**
Heart	0.439	0.511	0.32	0.734	0.208	0.387	0.796	0.794	**0.8**
Average ROC	0.443	0.509	0.531	0.362	0.494	0.438	0.58	0.632	**0.75**

4.3.2. Additional Experimental Results and Analysis of Running Time

To further test the scalability of the ISOD algorithm, the running time of the algorithm on the 10 datasets mentioned above was tested, and the results are represented in the form of a scatter plot, as shown in Figure 1. In this figure, the horizontal axis represents the size of the dataset, the vertical axis represents the dimensionality of the data, and the dot size represents the running time of the ISOD algorithm on the dataset. The larger the dot, the longer the running time.

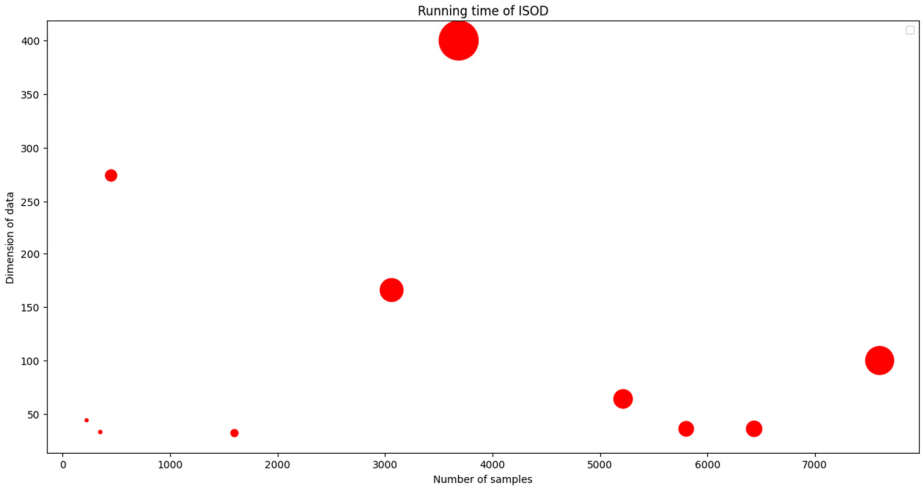

Figure 1. The running times for the ISOD algorithm on 10 benchmark datasets.(Larger dot mean longer running time).

Although Figure 1 does not provide a specific running time, by comparing the size of these scatter plots, we can see that a dataset with a large amount of data or data with a high dimensionality has a longer running time for the ISOD algorithm, which conforms the complexity analysis results mentioned earlier.

5. Conclusions

In this article, we proposed an effective unsupervised outlier detection method based on quantiles and skewness coefficients called ISOD. ISOD can be mainly divided into three stages: (1) the construction of the empirical cumulative distribution function; (2) the computation of the quantiles and skewness coefficients of each dimension; (3) summarizing the degree of anomalies in each dimension and ultimately obtaining the outlier score for each data point. After these stages, the method finally obtains the outlier scores.

The experimental results derived from applying the ISOD algorithm to 10 benchmark datasets show that the ISOD method has great competitive and promising performance in comparison to the state-of-the-art baseline anomaly detection algorithms. In addition to achieving better experimental results, the ISOD algorithm also has high interpretability and scalability, as explained in Section 4.

Based on Sections 3.4 and 4.3.1, it is clear that the ISOD algorithm does not require labeled data and that it is an unsupervised anomaly detection algorithm. At the same time, it has good scalability and can obtain good performance with ultra-high dimensional datasets. Finally, this algorithm is theoretically guaranteed to have high interpretability.

Author Contributions: Conceptualization, Y.H.; funding acquisition, W.L.; methodology, Y.H.; project administration, Y.H.; software, S.L., Y.G. and W.C.; supervision, W.L.; validation, S.L., Y.G. and W.C.; writing—original draft, Y.H.; writing—review and editing, Y.H. and W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (No. 61862011), the Guangxi Natural Science Foundation (No.2019GXNSFGA245004), and the Innovation Project of Guangxi Graduate Education (No.YCBZ2023128).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aggarwal, C.C.; Aggarwal, C.C. *An Introduction to Outlier Analysis*; Springer: Berlin/Heidelberg, Germany, 2017.
2. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv. (CSUR)* **2009**, *41*, 1–58. [CrossRef]
3. Ntroumpogiannis, A.; Giannoulis, M.; Myrtakis, N.; Christophides, V.; Simon, E.; Tsamardinos, I. A meta-level analysis of online anomaly detectors. *Vldb J.* **2023**, *32*, 845–886. [CrossRef]
4. Wang, Z.; Shao, L.; Cheng, K.; Liu, Y.; Jiang, J.; Nie, Y.; Li, X.; Kuang, X. ICDF: Intrusion collaborative detection framework based on confidence. *Int. J. Intell. Syst.* **2022**, *37*, 7180–7199. [CrossRef]
5. Heigl, M.; Weigelt, E.; Urmann, A.; Fiala, D.; Schramm, M. Exploiting the Outcome of Outlier Detection for Novel Attack Pattern Recognition on Streaming Data. *Electronics* **2021**, *10*, 2160. [CrossRef]
6. Zhang, H.; Zhao, S.; Liu, R.; Wang, W.; Hong, Y.; Hu, R. Automatic Traffic Anomaly Detection on the Road Network with Spatial-Temporal Graph Neural Network Representation Learning. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 4222827. [CrossRef]
7. Fournier, N.; Farid, Y.Z.; Patire, A. Erroneous High Occupancy Vehicle Lane Data: Detecting Misconfigured Traffic Sensors With Machine Learning. *Transp. Res. Rec.* **2022**, *2677*, 1593–1610. [CrossRef]
8. Dixit, P.; Bhattacharya, P.; Tanwar, S.; Gupta, R. Anomaly detection in autonomous electric vehicles using AI techniques: A comprehensive survey. *Expert Syst.* **2022**, *39*, e12754. [CrossRef]
9. Watts, J.; van Wyk, F.; Rezaei, S.; Wang, Y.; Masoud, N.; Khojandi, A. A Dynamic Deep Reinforcement Learning-Bayesian Framework for Anomaly Detection. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 22884–22894. [CrossRef]
10. Mansour, R.F.; Escorcia-Gutierrez, J.; Gamarra, M.; Villanueva, J.A.; Leal, N. Intelligent video anomaly detection and classification using faster RCNN with deep reinforcement learning model. *Image Vis. Comput.* **2021**, *112*, 104229. [CrossRef]
11. Zhao, Y.; Deng, B.; Shen, C.; Liu, Y.; Lu, H.; Hua, X.-S. Spatio-Temporal AutoEncoder for Video Anomaly Detection. In Proceedings of the 25th ACM International Conference on Multimedia (MM), Comp Hist Museum, Mountain View, CA, USA, 23–27 October 2017; pp. 1933–1941.
12. Dang, T.T.; Ngan, H.E.T.; Liu, W. Distance-Based k-Nearest Neighbors Outlier Detection Method in Large-Scale Traffic Data. In Proceedings of the IEEE International Conference on Digital Signal Processing (DSP), Singapore, 21–24 July 2015; pp. 507–510.
13. Wang, H.; Wang, W.; Liu, Y.; Alidaee, B. Integrating Machine Learning Algorithms With Quantum Annealing Solvers for Online Fraud Detection. *IEEE Access* **2022**, *10*, 75908–75917. [CrossRef]
14. Bhattacharjee, P.; Garg, A.; Mitra, P. KAGO: An approximate adaptive grid-based outlier detection approach using kernel density estimate. *Pattern Anal. Appl.* **2021**, *24*, 1825–1846. [CrossRef]
15. Zhang, Y.-L.; Zhou, J.; Zheng, W.; Feng, J.; Li, L.; Liu, Z.; Li, M.; Zhang, Z.; Chen, C.; Li, X.; et al. Distributed Deep Forest and its Application to Automatic Detection of Cash-Out Fraud. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–19. [CrossRef]
16. Chaudhry, H.N.; Javed, Y.; Kulsoom, F.; Mehmood, Z.; Khan, Z.I.; Shoaib, U.; Janjua, S.H. Sentiment Analysis of before and after Elections: Twitter Data of U.S. Election 2020. *Electronics* **2021**, *10*, 2082. [CrossRef]
17. Chalapathy, R.; Toth, E.; Chawla, S. Group Anomaly Detection Using Deep Generative Models. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Dublin, Ireland, 10–14 September 2020; pp. 173–189.
18. Chenaghlou, M.; Moshtaghi, M.; Leckie, C.; Salehi, M. Online Clustering for Evolving Data Streams with Online Anomaly Detection. In Proceedings of the 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Melbourne, Australia, 3–6 June 2018; pp. 506–519.
19. Sharma, V.; Kumar, R.; Cheng, W.-H.; Atiquzzaman, M.; Srinivasan, K.; Zomaya, A.Y. NHAD: Neuro-Fuzzy Based Horizontal Anomaly Detection in Online Social Networks. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 2171–2184. [CrossRef]
20. Souiden, I.; Omri, M.N.; Brahmi, Z. A survey of outlier detection in high dimensional data streams. *Comput. Sci. Rev.* **2022**, *44*, 100463. [CrossRef]
21. Pei, Y.; Zaïane, O. A Synthetic Data Generator for Clustering and Outlier Analysis. 2006. Available online: https://era.library.ualberta.ca/items/63beb6a7-cc50-4ffd-990b-64723b1e4bf9 (accessed on 20 October 2023).
22. Sikder, M.N.K.; Batarseh, F.A. Outlier detection using AI: A survey. In *AI Assurance*; Elsevier: Amsterdam, The Netherlands, 2023; pp. 231–291.
23. Chatterjee, A.; Ahmed, B.S. IoT anomaly detection methods and applications: A survey. *Internet Things* **2022**, *19*, 100568. [CrossRef]
24. Pang, G.; Shen, C.; Cao, L.; Hengel, A.V.D. Deep Learning for Anomaly Detection. *ACM Comput. Surv.* **2021**, *54*, 1–38. [CrossRef]
25. Boukerche, A.; Zheng, L.; Alfandi, O. Outlier Detection: Methods, Models, and Classification. *ACM Comput. Surv.* **2020**, *53*, 1–37. [CrossRef]
26. Samudra, S.; Barbosh, M.; Sadhu, A. Machine Learning-Assisted Improved Anomaly Detection for Structural Health Monitoring. *Sensors* **2023**, *23*, 3365. [CrossRef]
27. Qiu, J.; Shi, H.; Hu, Y.; Yu, Z. Enhancing Anomaly Detection Models for Industrial Applications through SVM-Based False Positive Classification. *Appl. Sci.* **2023**, *13*, 12655. [CrossRef]
28. Kerboua, A.; Kelaiaia, R. Fault Diagnosis in an Asynchronous Motor Using Three-Dimensional Convolutional Neural Network. *Arab. J. Sci. Eng.* **2023**, 1–19. [CrossRef]
29. Jiang, J.; Zhu, J.; Bilal, M.; Cui, Y.; Kumar, N.; Dou, R.; Su, F.; Xu, X. Masked swin transformer unet for industrial anomaly detection. *IEEE Trans. Ind. Inform.* **2022**, *19*, 2200–2209. [CrossRef]

30. Drost, B.; Ulrich, M.; Bergmann, P.; Hartinger, P.; Steger, C. Introducing mvtec itodd-a dataset for 3d object recognition in industry. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2200–2208.
31. Park, C.H. A Comparative Study for Outlier Detection Methods in High Dimensional Text Data. *J. Artif. Intell. Soft Comput. Res.* **2023**, *13*, 5–17. [CrossRef]
32. Sunny, J.S.; Patro, C.P.K.; Karnani, K.; Pingle, S.C.; Lin, F.; Anekoji, M.; Jones, L.D.; Kesari, S.; Ashili, S. Anomaly Detection Framework for Wearables Data: A Perspective Review on Data Concepts, Data Analysis Algorithms and Prospects. *Sensors* **2022**, *22*, 756. [CrossRef] [PubMed]
33. Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation Forest. In Proceedings of the 8th IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422.
34. Staffini, A.; Svensson, T.; Chung, U.-I.; Svensson, A.K. A Disentangled VAE-BiLSTM Model for Heart Rate Anomaly Detection. *Bioengineering* **2023**, *10*, 683. [CrossRef] [PubMed]
35. Sun, Z.; Peng, Q.; Mou, X.; Bashir, M.F. Generic and scalable periodicity adaptation framework for time-series anomaly detection. *Multimed. Tools Appl.* **2023**, *82*, 2731–2748. [CrossRef]
36. Huang, Y.; Liu, W.; Li, S.; Guo, Y.; Chen, W. A Novel Unsupervised Outlier Detection Algorithm Based on Mutual Information and Reduced Spectral Clustering. *Electronics* **2023**, *12*, 4864. [CrossRef]
37. Mozaffari, M.; Doshi, K.; Yilmaz, Y. Self-Supervised Learning for Online Anomaly Detection in High-Dimensional Data Streams. *Electronics* **2023**, *12*, 1971. [CrossRef]
38. Liu, Y.; Zhou, S.; Wan, Z.; Qiu, Z.; Zhao, L.; Pang, K.; Li, C.; Yin, Z. A Self-Supervised Anomaly Detector of Fruits Based on Hyperspectral Imaging. *Foods* **2023**, *12*, 2669. [CrossRef]
39. Zhang, X.; Mu, J.; Zhang, X.; Liu, H.; Zong, L.; Li, Y. Deep anomaly detection with self-supervised learning and adversarial training. *Pattern Recognit.* **2022**, *121*, 108234. [CrossRef]
40. Hojjati, H.; Ho, T.K.K.; Armanfard, N. Self-Supervised Anomaly Detection: A Survey and Outlook. *arXiv* **2022**, arXiv:2205.05173.
41. Liu, K.; Fu, Y.; Wang, P.; Wu, L.; Bo, R.; Li, X. Automating feature subspace exploration via multi-agent reinforcement learning. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 207–215.
42. Ramaswamy, S.; Rastogi, R.; Shim, K. Efficient algorithms for mining outliers from large data sets. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000; pp. 427–438.
43. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. *Sigmod Rec.* **2000**, *29*, 93–104. [CrossRef]
44. He, Z.; Xu, X.; Deng, S. Discovering cluster-based local outliers. *Pattern Recognit. Lett.* **2003**, *24*, 1641–1650. [CrossRef]
45. Zhao, Y.; Nasrullah, Z.; Hryniewicki, M.K.; Li, Z. LSCP: Locally selective combination in parallel outlier ensembles. In Proceedings of the 2019 SIAM International Conference on Data Mining, Calgary, AB, Canada, 2–4 May 2019; pp. 585–593.
46. Scholkopf, B.; Williamson, R.; Smola, A.; Shawe-Taylor, J.; Platt, J. Support vector method for novelty detection. *Adv. Neural Inf. Process. Syst.* **2000**, *12*, 582–588.
47. Xu, H.; Pang, G.; Wang, Y.; Wang, Y. Deep isolation forest for anomaly detection. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 12591–12604. [CrossRef]
48. Akcay, S.; Atapour-Abarghouei, A.; Breckon, T.P. GANomaly: Semi-supervised Anomaly Detection via Adversarial Training. In Proceedings of the 14th Asian Conference on Computer Vision (ACCV), Perth, Australia, 2–6 December 2018; pp. 622–637.