Journal of Computer Technology and Applied Mathematics
Vol. 1, No. 1, 2024

Research on the Effectiveness of Different Outlier Detection Methods in Common Data Distribution Types

SONG, Qingqing [1*] XIA, Shaoliang [1]

[1] Belarusian State University, Belarus

**** SONG, Qingqing is the corresponding author, E-mail: fpm.sunC@bsu.by***

Abstract: Outlier detection are widely applied in areas such as network performance optimization and pre-processing of machine learning data. In the field of machine learning, the objective is to enhance data quality, thereby improving the performance of subsequent statistical analyses or machine learning models. Currently, there are numerous effective and reliable outlier analysis methods, and their effectiveness varies significantly when dealing with different types of data distributions. Therefore, it is essential to select an appropriate outlier analysis method. In this study, we conducted outlier detection on sample data from five continuous probability distributions (including Normal, Chi-square, Exponential, Gamma, and T distributions) and four discrete probability distributions (including Binomial, Poisson, Geometric, and Hypergeometric distributions). This paper employs five outlier detection methods, namely Z-Score, IQR, DBScan, Isolation Forest, and Random Forest, and evaluates the detection effectiveness of these methods. Through comparison and analysis, this paper summarizes the characteristics of various outlier detection methods when dealing with sample data from different types of distributions. These findings will assist us in making more rational method selections when facing different outlier detection scenarios.

Keywords: Outlier Detection, Outlier Analysis, Machine Learning, Data Distribution, Performance Evaluation, Data Preprocessing.

DOI: https://doi.org/10.5281/zenodo.10888672

1 Introduction

Outlier detection [1] is a statistical method used to identify and handle outliers in a dataset. Outliers, also known as anomalies, refer to one or several values in the data that significantly differ from other values. These outliers may be due to variability in measurements or could represent experimental errors. Currently, outlier detection has a wide range of applications in various fields such as network performance optimization, pre-processing of machine learning data, healthcare, and more.

Especially in recent years, the importance of outlier analysis in the field of machine learning has been widely recognized by scholars, and it has been proven to have a positive effect on the research of machine learning. In the field of machine learning, the purpose of outlier analysis is to improve the quality of data, which in turn affects further statistical models and machine learning models. For example, outliers may distort the model, leading to extended training time, reduced accuracy, and degraded performance.

Currently, there are a variety of proven effective and reliable methods for outlier analysis, such as Z-score, IQR, DBScan clustering, Isolation Forest, Random Forest, and other different methods. The effectiveness of each different outlier detection method varies significantly in datasets with different distribution types, so it is necessary to choose an outlier analysis method that meets the needs according to different data types.

This paper will evaluate the effectiveness of five outlier analysis methods, namely Z-score, IQR, DBScan clustering, Isolation Forest, and Random Forest, in datasets of different continuous probability distributions and different discrete probability distributions. Based on the results, this paper will analyze and summarize the characteristics of different outlier detection methods, and generalize the applicable scenarios of each outlier analysis method.

2 Outlier Detection Method

In the field of statistics, widely used and proven effective outlier analysis methods include histogram analysis, box plot analysis, Z-score, IQR, DBScan clustering, Isolation Forest, and Random Forest. Since the principle of the box plot analysis method is the same as the IQR method, this paper does not directly discuss the box plot analysis method. Furthermore, as histogram analysis relies on human visual interpretation, this method is also not within the scope of this paper.

2.1 Z-Score (Standard Score)

The Z-Score is a statistical concept [2], which describes the distance of a raw score from the mean in units of standard deviation. The calculation formula for Z-Score is $z = \frac{x-\mu}{\sigma}$. Here, x is the raw score, μ is the population mean, and σ is the population standard deviation. The value of the Z-Score tells you how many standard deviations away from the mean. If the Z-Score is equal to 0, then it is on the mean. A positive Z-Score indicates that the raw score is above the mean, while a negative Z-Score indicates that the raw score is below the mean.

2.2 IQR (Interquartile Range)

The Interquartile Range (IQR) is a concept that measures statistical dispersion and data variability [3], which is achieved by dividing the dataset into quartiles. The IQR is the difference between the third quartile and the first quartile (IQR = Q3 - Q1). In this case, outliers are defined as observations that are below (Q1 - 1.5x IQR) or above (Q3 + 1.5x IQR).

2.3 DBScan Clustering

DBScan is a clustering algorithm that groups data [4]. It can also serve as a density-based anomaly detection method, applicable to both unidimensional and multidimensional data. DBScan has three important concepts: core points, boundary points, and noise points. Core points and boundary points are in the same cluster, while noise points are data points that do not belong to any cluster. They could be anomalous or non-anomalous, requiring further investigation.

2.4 Isolation Forest

Isolation Forest is an unsupervised learning algorithm [5], belonging to the ensemble decision tree family. This method is different from all previous methods. All previous methods tried to find the normal region of the data, and then identify any outliers or anomalies outside this defined region. Isolation Forest explicitly isolates outliers, rather than analyzing and constructing normal points and regions by assigning scores to each data point. It takes advantage of a fact that anomalies are minority data points, and their attribute values are significantly different from the attribute values of normal instances.

2.5 Random Forest

This method involves training a Random Forest model and using this model to predict the original data, then identifying the data point corresponding to the maximum value in the prediction results, thereby achieving outlier detection. This is a common unsupervised learning method used for detecting anomalies or outliers in data. The advantage of this method is that it can handle high-dimensional data, and it can process both real-time streaming data and offline data.

3 Common Data Distribution Types

Data distribution types are generally divided into two categories: continuous probability distribution and discrete probability distribution. The subject of this paper is outlier analysis. Theoretically, the Z-Score and IQR outlier analysis methods have better analysis effects only on samples of continuous probability distribution. However, DBScan clustering, Isolation Forest, and Random Forest have better analysis effects on both samples of discrete probability distribution and continuous probability distribution. Therefore, this paper will categorize and study these two types of samples.

3.1 Continuous Probability Distribution

This study will select five types of continuous probability distributions: the normal distribution, chi-square distribution, exponential distribution, gamma distribution, and T-distribution, to evaluate the effectiveness of outlier detection. The histogram samples of data from these five continuous probability distributions are shown in Figure 1.

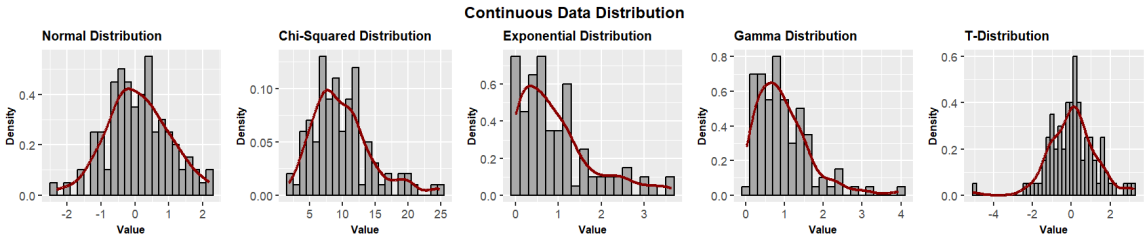

Figure 1 Five continuous distribution data histograms

3.1.1 normal distribution

The normal distribution, also known as the Gaussian distribution [7], is a probability distribution that is extremely important in fields such as mathematics, physics, and engineering. In its graphical representation, the middle part is high, the two ends are low, and it presents a bell-shaped curve. As shown in Figure 2, these are the histogram samples of data from the normal distribution under different parameters.

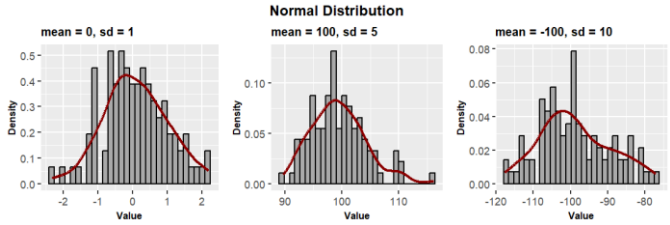

Figure 2 Histogram of normally distributed data under different parameters

The normal distribution can be described by two parameters: the mean (μ) and the standard deviation (σ). Its probability density function is shown in formula (1).

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

Where x is the value we are interested in.

3.1.2 Chi-Square Distribution

If k independent random variables all follow a standard normal distribution, then the sum of the squares of these k random variables forms a new variable. This new variable follows a chi-square distribution [8]. As shown in Figure 3, these are the histogram samples of data from the chi-square distribution under different parameters.

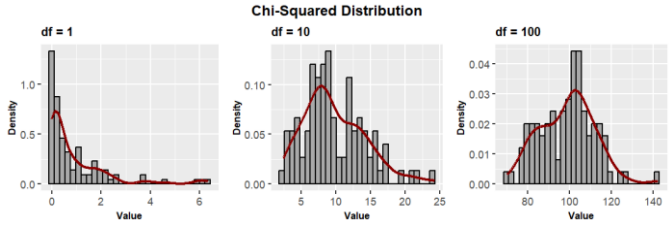

Figure 3 Histogram of chi-square distributed data under different parameters

The probability density function of the chi-square distribution is shown in formula (2).

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \quad (2)$$

Where: x is the value we are interested in, k is the degrees of freedom of the distribution, and Γ(k/2) is the gamma function.

3.1.3 Exponential Distribution

In probability theory and statistics, the exponential distribution is a continuous probability distribution used to represent the time interval between independent random events [9, 10]. If a random variable X follows an exponential distribution with parameter λ, we usually write it as X~Exp(λ). As shown in Figure 4, these are the histogram samples of data from the exponential distribution under different parameters.

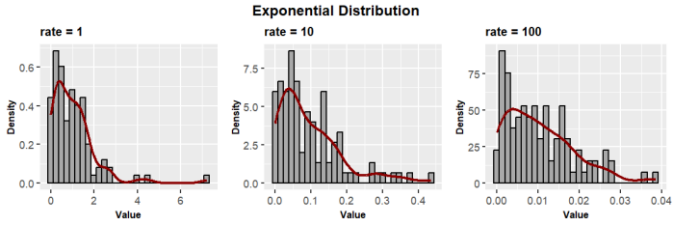

Figure 4 Histogram of exponentially distributed data under different parameters

The probability density function of the exponential distribution is shown in formula (3).

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases} \quad (3)$$

Where: x is the value we are interested in, λ is the parameter of the distribution, also known as the rate parameter, which is the reciprocal of the mean of the distribution.

3.1.4 Gamma Distribution

The gamma distribution is a two-parameter continuous probability distribution [11, 12], widely used in engineering, physics, statistics, and other fields. If a random variable X follows a gamma distribution with parameters (k, θ), we usually write it as X~Gamma (k, θ). As shown in Figure 5, these are the histogram samples of data from the gamma distribution under different parameters.

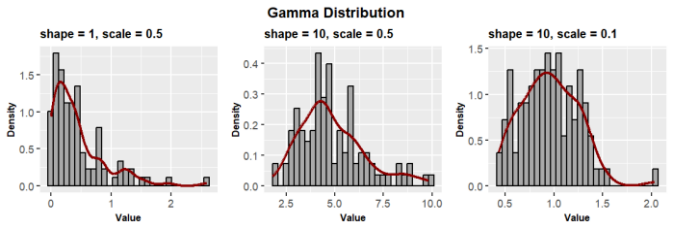

Figure 5 Histogram of gamma distributed data under different parameters

The probability density function of the gamma distribution is shown in formula (4).

$$f(x;k,\theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)} \quad (4)$$

Where: x is the value we are interested in, k is the shape parameter, θ is the scale parameter, and $\Gamma(k)$ is the gamma function.

3.1.5 T-Distribution

In statistics, the t-distribution is a continuous probability distribution commonly used in experiments with small sample sizes where the population standard deviation is unknown [13-15].

If a random variable T follows a t-distribution with degrees of freedom v, we usually write it as T~t(v). As shown in Figure 6, these are the histogram samples of data

from the t-distribution under different parameters.

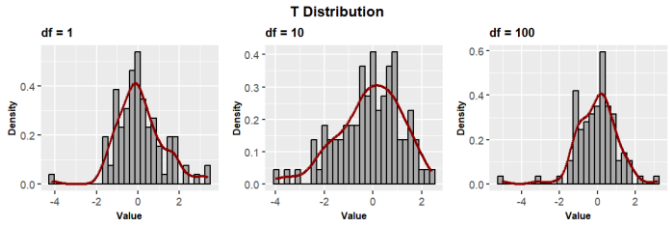

Figure 6 Histogram of T-distributed data under different parameters

The probability density function of the t-distribution is shown in formula (5).

$$f(t;v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}} \quad (5)$$

Where: t is the value we are interested in, v is the degrees of freedom of the distribution, and Γ is the gamma function.

3.2 Discrete Probability Distributions

This study will select four types of discrete probability distributions: the binomial distribution, Poisson distribution, geometric distribution, and hypergeometric distribution, to evaluate the effectiveness of outlier detection. The histogram samples of data from these four discrete probability distributions are shown in Figure 7.

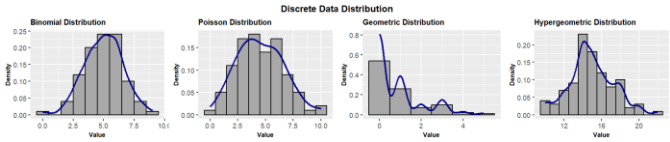

Figure 7 Data distribution histograms of four discrete probability distributions

3.2.1 Binomial Distribution

The binomial distribution is a discrete probability distribution that describes the number of successes in a series of independent yes/no experiments, where the probability of success for each experiment is fixed [16].

If a random variable X follows a binomial distribution with parameters (n, p), we usually write it as X~B(n,p). As shown in Figure 8, these are the histogram samples of data from the binomial distribution under different parameters.

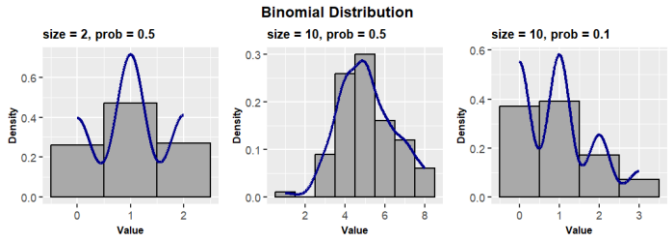

Figure 8 Histogram of binomially distributed data under different parameters

The probability mass function of the binomial distribution is shown in formula (6).

$$P(X=k) = C_n^k p^k (1-p)^{n-k} \quad (6)$$

Where: k is the number of successes we are interested in, n is the number of experiments, p is the probability of success in each experiment, and C_n^k is the combination number, representing the number of ways to select k elements from n different elements.

3.2.2 Poisson Distribution

The Poisson distribution is a discrete probability distribution that describes the average rate of events occurring in a fixed time or space [17]. If a random variable X follows a Poisson distribution with parameter λ, we usually write it as X~Poisson(λ). As shown in Figure 9, these are the histogram samples of data from the Poisson distribution under different parameters.

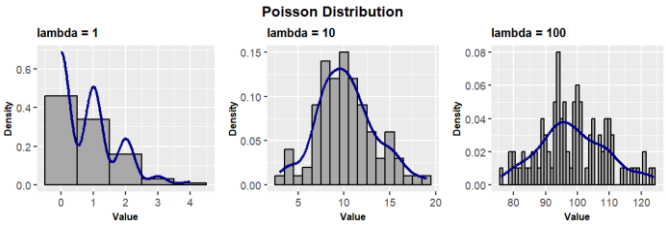

Figure 9 Histogram of poisson distributed data under different parameters

The probability mass function of the Poisson distribution is shown in formula (7).

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (7)$$

Where: k is the number of events we are interested in, λ is the average rate of events per unit time (or unit space), and k! is the factorial of k.

3.2.3 Geometric Distribution

The geometric distribution is a discrete probability distribution that describes the number of independent Bernoulli trials required before the first success [18]. If a random variable X follows a geometric distribution with parameter p, we usually write it as X~Geo(p). As shown in Figure 10, these are the histogram samples of data from the geometric distribution under different parameters.

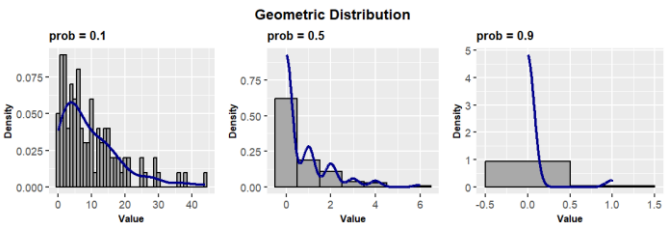

Figure 10 Histogram of geometrically distributed data under different parameters

The probability mass function of the geometric

distribution is shown in formula (8).

$$P(X = k) = (1 - p)^{k-1}p \qquad (8)$$

Where: k is the number of trials we are interested in, p is the probability of success in each trial.

3.2.4 Hypergeometric Distribution

The hypergeometric distribution is a discrete probability distribution that describes the number of successes when drawing samples from a finite population [19-21]. If a random variable X follows a hypergeometric distribution with parameters (N, K, n), we usually write it as X~Hypergeometric (N, K, n). As shown in Figure 10, these are the histogram samples of data from the hypergeometric distribution under different parameters.

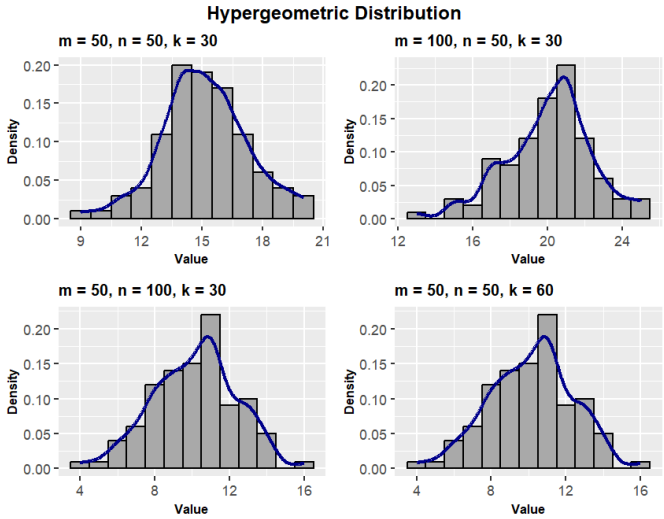

Figure 11 Histogram of hypergeometrically distributed data under different parameters

The probability mass function of the hypergeometric distribution is shown in formula (9).

$$P(X = k) = \frac{C_K^k C_{N-K}^{n-k}}{C_N^n} \qquad (9)$$

Where: k is the number of successes we are interested in, N is the size of the population, K is the number of successful states in the population, n is the number of sampling times, and C_n^k is the combination number, representing the number of ways to select k elements from n different elements.

4 Outlier Detection Effectiveness Evaluation

In this section, we will use five different outlier detection methods to perform outlier detection on samples that include both continuous probability distribution and discrete probability distribution. For the parameters in these five outlier detection methods, we choose values that are often selected in actual use. In the Z-Score method, we set the Z-Score threshold to 2; in the IQR method, we set the first quartile to 0.25 and the third quartile to 0.75; in the DBScan method, we set the distance threshold (eps) to 0.5 and the minimum number of neighbors (minPts) to 5; in the Isolation Forest and Random Forest methods, we set the number of individual decision trees in the forest (ntree) to 100.

4.1 Effectiveness of Outlier Detection Methods in Samples from Continuous Probability Distributions

This section will evaluate the performance of various outlier detection methods on samples from five continuous probability distributions. In order to more accurately obtain the characteristics of each outlier detection method, we will generate two sets of data with different parameters for different continuous data distributions. The specific distribution parameters are shown in Table 1.

Table 1 Distribution Parameters of Samples from Different Continuous Probability Distributions

Distribution Type	Distribution Parameter A	Distribution Parameter B
Normal Distribution	Standard Deviation = 1	Standard Deviation = 5
Chi-square Distribution	Degrees of Freedom = 10	Degrees of Freedom = 50
Exponential Distribution	Rate = 1	Rate = 0.1
Gamma Distribution	Shape = 2, Scale = 0.5	Shape = 20, Scale = 0.5
T-Distribution	Degrees of Freedom = 10	Degrees of Freedom = 50

4.1.1 Effectiveness of different outlier detection methods on normally distributed sample

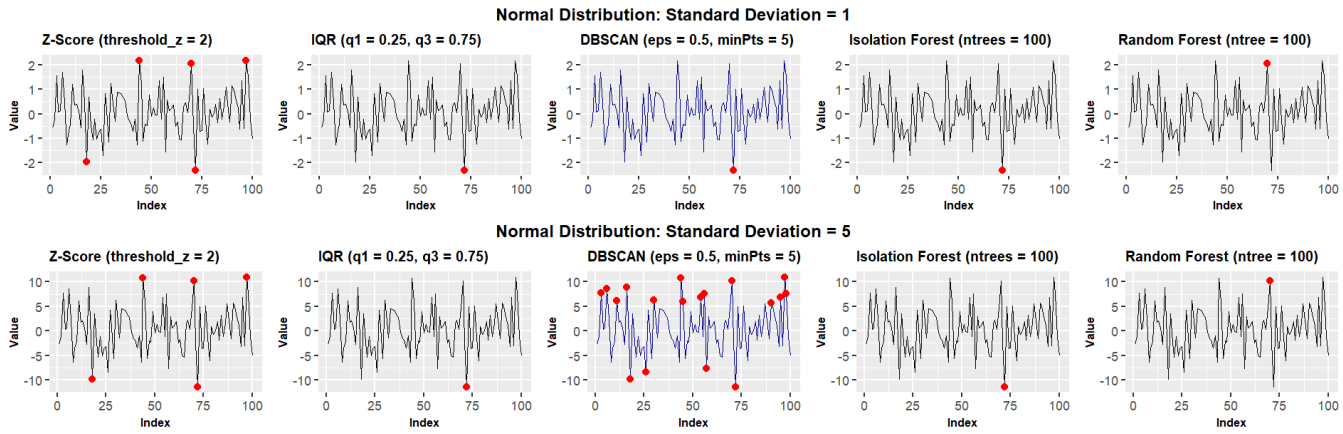

Figure 12 Effectiveness of different outlier detection methods on normally distributed sample under different parameters

In the data samples of the normal distribution under two different parameters, it can be observed that, except for DBScan, the five outlier detection methods are not sensitive to the parameters of the normal distribution of the data samples. DBScan, due to the constant distance threshold (eps) parameter of 0.5 in this experiment, will have more points determined as outliers when the span between data is larger. In addition, the Z-Score method with a threshold parameter set to 2 has a higher sensitivity for outlier detection.

4.1.2 Effectiveness of different outlier detection methods on Chi-square distributed sample

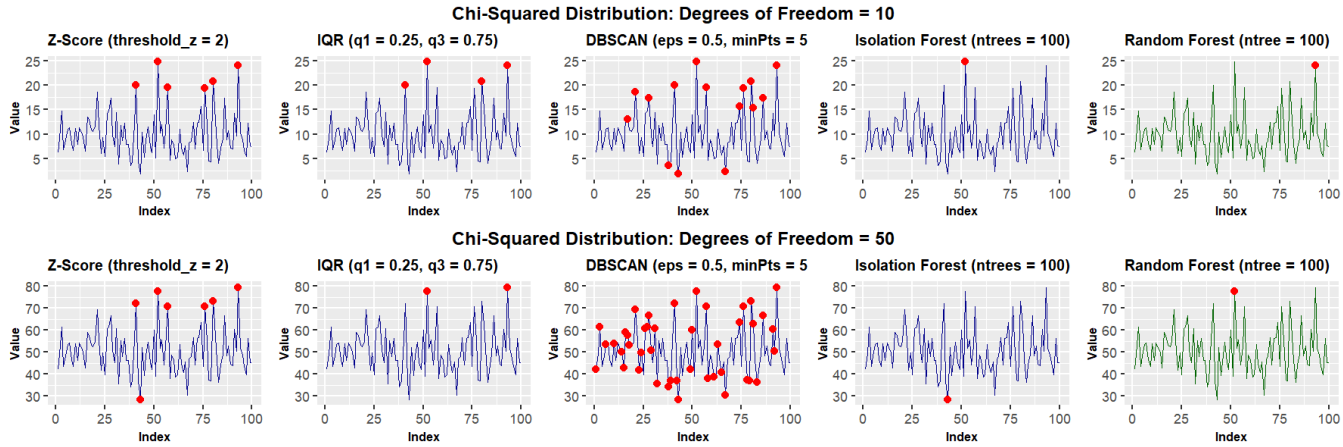

Figure 13 Effectiveness of different outlier detection methods on Chi-square distributed sample under different parameters

Under the two chi-square distribution samples with degrees of freedom of 10 and 50, we can observe that the detection effect of the Z-Score method is relatively stable. The IQR method has a higher sensitivity for data with a smaller numerical span between data. However, DBScan has an overly high sensitivity under the condition of eps=0.5.

It is worth noting that when the random forest outlier detection method is applied to the chi-square distribution sample with a degree of freedom of 10, a phenomenon occurs where the second largest value is determined as an outlier, but the maximum value is not determined as an outlier.

4.1.3 Effectiveness of different outlier detection methods on exponentially distributed sample

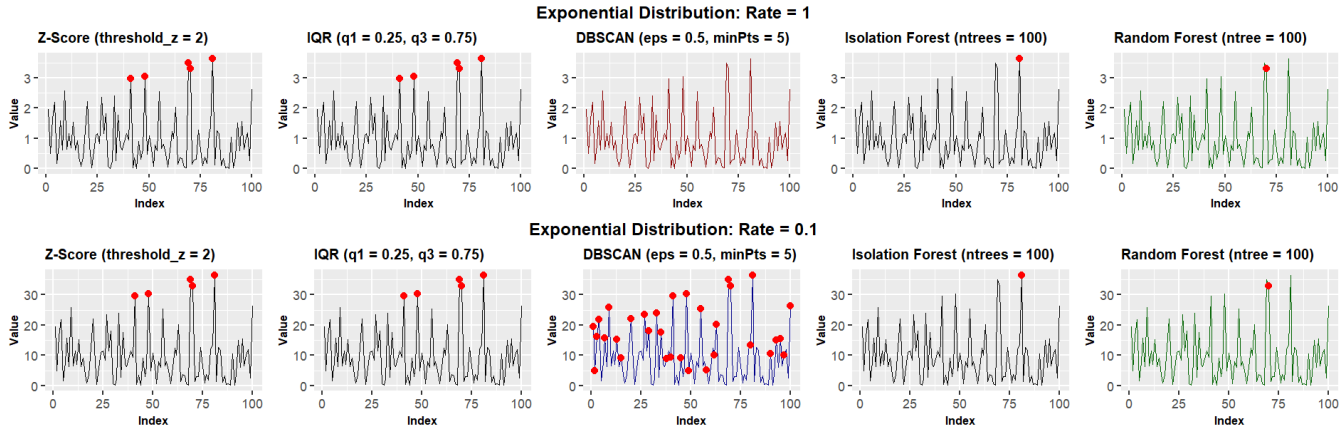

Figure 14 Effectiveness of different outlier detection methods on exponentially distributed sample under different parameters

Under the two exponential distribution samples with rate parameters of 1 and 0.1, we can observe that the detection effects of the Z-Score, Isolation Forest, and IQR methods are relatively stable. In the sample with a rate parameter of 1, DBScan did not detect any outliers under the condition of eps=0.5.

Similar to the phenomenon in section 4.1.2, the random forest outlier detection method in both samples showed the phenomenon where the second largest value was determined as an outlier, but the maximum value was not determined as an outlier.

4.1.4 Effectiveness of different outlier detection methods on gamma distributed samples

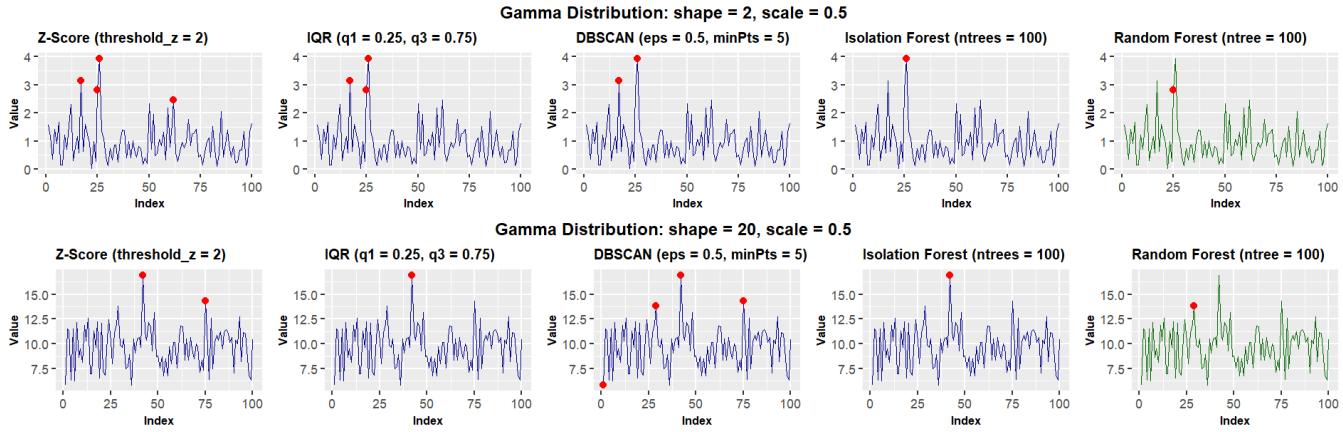

Figure 15 Effectiveness of different outlier detection methods on gamma distributed samples under different parameters

The sample characteristics of the gamma distribution under different parameters are inconsistent, so all outlier detection methods have certain differences in the detection effects for the two samples. In the gamma distribution samples, the DBScan outlier detection method is relatively stable at eps=0.5 compared to samples of other types of distributions.

Similar to the phenomena that occurred in sections 4.1.2 and 4.1.3, the random forest outlier detection method in both samples showed the phenomenon where the second largest value was determined as an outlier, but the maximum value was not determined as an outlier.

4.1.5 Effectiveness of different outlier detection methods on T-distributed samples

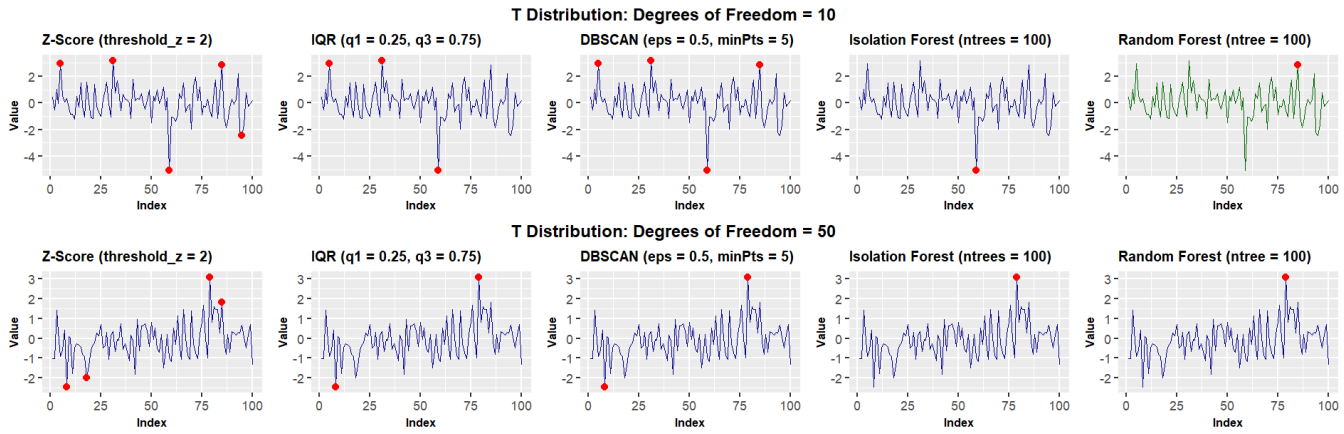

Figure 16 Effectiveness of different outlier detection methods on T-distributed samples under different parameters

In the experiments with T-distribution samples under different parameters, because the numerical span of the data is relatively small, the detection effect of DBScan is superior to its effect on samples from other continuous data distributions.

In this section, we can observe that in the Z-Score, IQR, and DBScan outlier detection methods, both upper and lower outliers can be detected simultaneously, while the Isolation Forest and Random Forest can only detect one of the upper and lower outliers at the same time.

4.2 Effectiveness of Outlier Detection Methods in Samples from Discrete Probability Distributions

This section will evaluate the performance of various outlier detection methods on samples from four discrete probability distributions. Similar to section 4.1, in order to more accurately obtain the characteristics of each outlier detection method, we will generate two sets of data with different parameters for different discrete probability distributions. The specific distribution parameters are shown in Table 2.

Table 2 Distribution Parameters of Samples from Different Discrete Probability Distributions

Distribution Type	Distribution Parameter A	Distribution Parameter B
Binomial Distribution	Size = 5, Prob = 0.5	Size = 20, Prob = 0.5
Poisson Distribution	Lambda = 5	Lambda = 20
Geometric Distribution	Prob = 0.5	Prob = 0.1
Hypergeometric Distribution	m = 50, n = 50, k = 30	m = 75, n = 100, k = 15

Because the data points of a discrete probability distribution are discrete, this means that the distances between data points may not be uniform. Therefore, using methods like IQR or Z-Score may lead to false positives or false negatives in outlier detection. Also, in discrete data, quartiles may not be unique, which makes the application of the IQR method in discrete data complex.

Hence, in the conclusions of the outlier detection method effectiveness evaluation experiment in discrete probability distribution data, we mark the Z-Score and IQR detection methods as gray.

4.2.1 Effectiveness of different outlier detection methods on binomially distributed samples

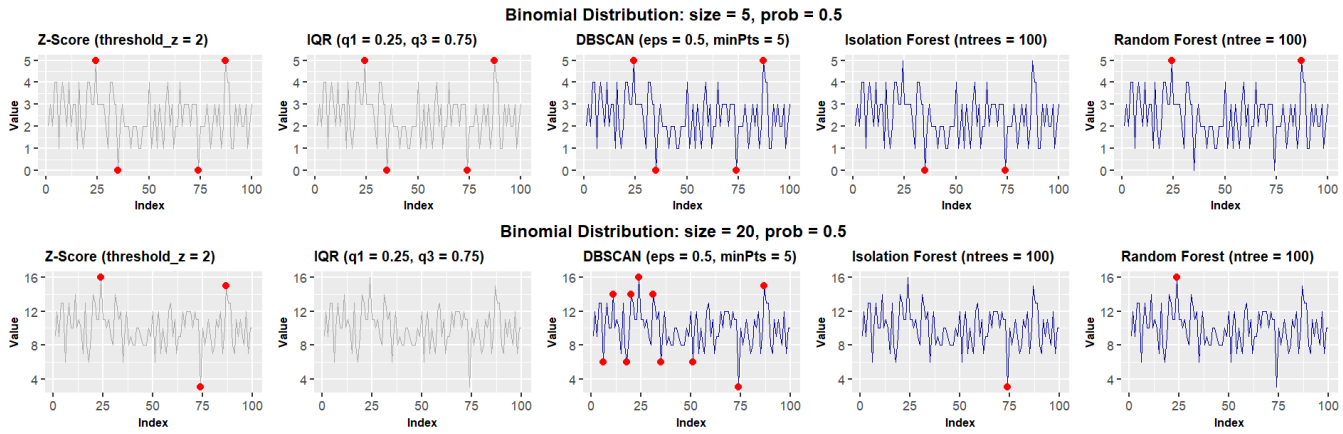

Figure 17 Effectiveness of different outlier detection methods on binomially distributed samples under different parameters

In the two binomial distribution samples, we can observe that the Z-Score method also achieves good results. Similar to the phenomenon that occurred in section 4.1.5, the Isolation Forest and Random Forest methods in the binomial distribution samples only detect one of the upper and lower outliers at the same time.

4.2.2 Effectiveness of different outlier detection methods on Poisson distributed samples

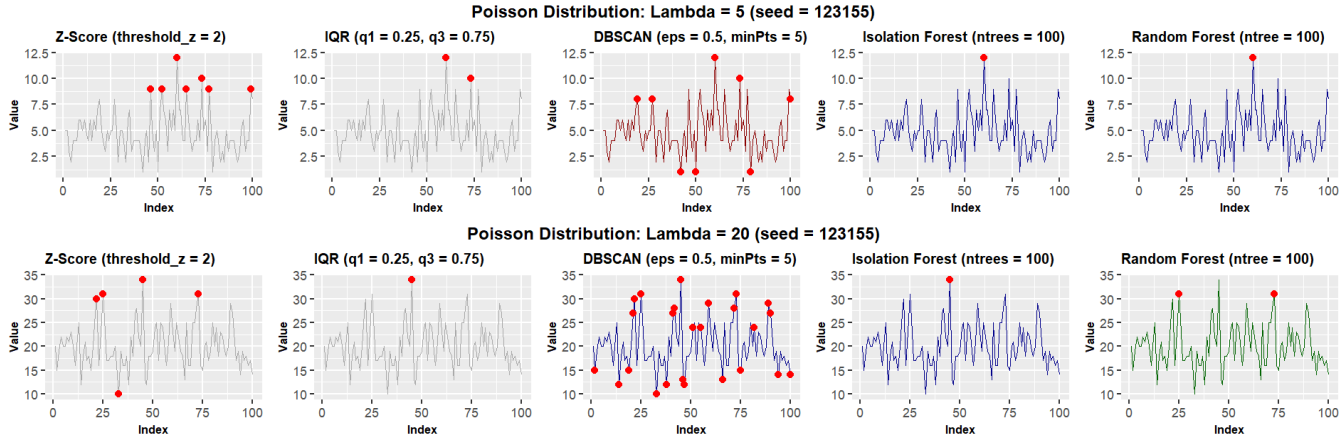

Figure 18 Effectiveness of different outlier detection methods on poisson distributed samples under different parameters

In the Poisson distribution sample with Lambda=5, a special phenomenon was observed with the DBScan outlier detection method. We found that the maximum value and the third to maximum value have been determined as outliers, however, the second to maximum value was not confirmed as an outlier.

Similarly, in the two Poisson distribution samples, we can observe that the Z-Score method also achieves good results. The Isolation Forest and Random Forest methods in the Poisson distribution samples only detect one of the upper and lower outliers at the same time. The Random Forest outlier detection method in the Poisson distribution sample with Lambda=20 showed the phenomenon where the maximum value was determined as an outlier, but the maximum value was not determined as an outlier.

4.2.3 Effectiveness of different outlier detection methods on geometrically distributed samples

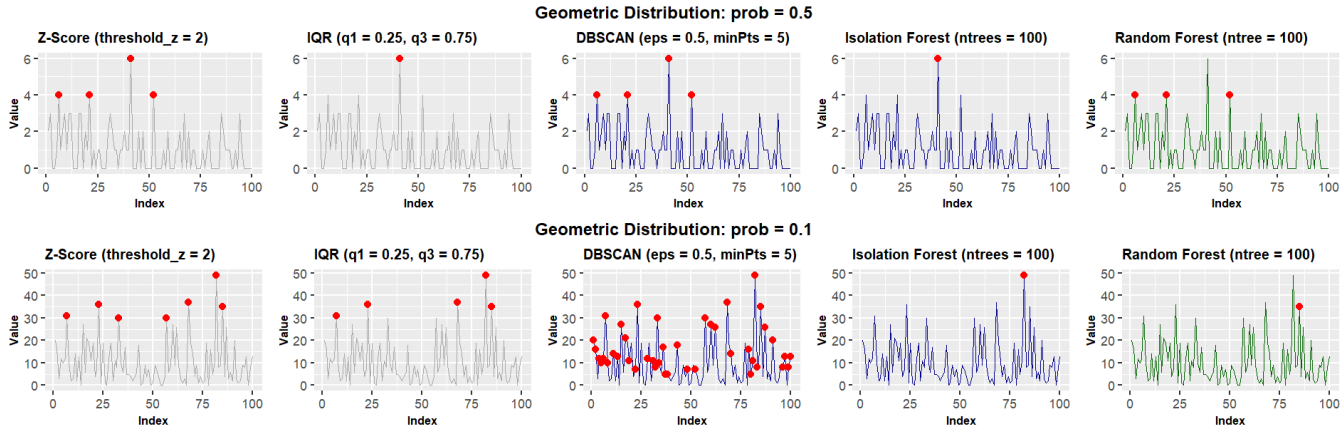

Figure 19 Effectiveness of different outlier detection methods on geometrically distributed samples under different parameters

In the two geometric distribution data samples, due to the large difference in numerical span between the data, the effect of the DBScan outlier detection method varies greatly.

Similarly, in the two geometric distribution samples, we can observe that the Z-Score method also achieves good results. The Random Forest outlier detection method in both geometric distribution samples showed the phenomenon where the second to maximum value was determined as an outlier, but the maximum value was not determined as an outlier.

4.2.4 Effectiveness of different outlier detection methods on hypergeometrically distributed samples

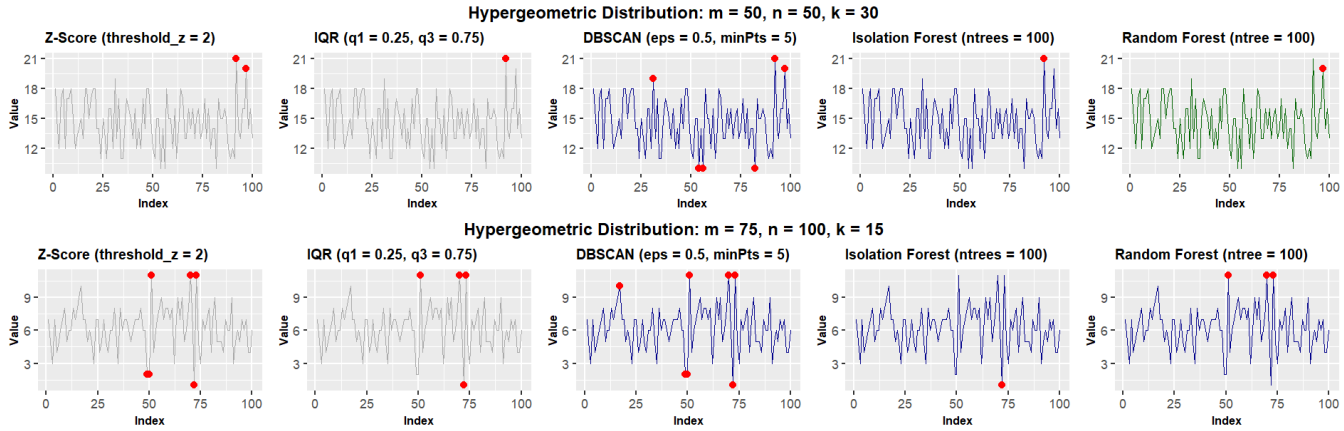

Figure 20 Effectiveness of different outlier detection methods on hypergeometrically distributed samples under different parameters

In the two hypergeometric distribution samples, we can observe that the effect of the DBScan outlier detection method is superior to its effect in other discrete probability distribution samples.

The Isolation Forest and Random Forest outlier detection methods applied to the hypergeometric distribution samples only detect one of the upper and lower outliers at the same time. The Random Forest outlier detection method in the hypergeometric distribution sample with m=50, n=50, k=30 showed the phenomenon where the second to maximum value was determined as an outlier, but the maximum value was not determined as an outlier.

5 Effectiveness Analysis

Through the evaluation of the effects of different outlier detection methods on samples from different distributions, we can find that the DBScan method is quite sensitive to the numerical span between data in the samples. It needs to adjust the eps parameter according to the numerical span to achieve better results. The Z-Score, IQR, and DBScan outlier detection methods can usually detect both upper and lower outliers simultaneously. The Z-Score and IQR outlier detection methods perform stably and excellently when applied to samples from continuous data distributions.

• **Phenomenon A:** The DBScan method may exhibit a phenomenon where the maximum value and the third to maximum value have already been determined as outliers, however, the second to maximum value has not been confirmed as an outlier. This phenomenon occurred in one out of nine sets of experimental data.

• **Phenomenon B:** The Z-Score method performs well in most of the discrete probability distribution samples.

• **Phenomenon C:** The IQR outlier detection method usually cannot achieve ideal results when applied to discrete probability distribution samples.

• **Phenomenon D:** The Isolation Forest and Random Forest outlier detection methods usually can only identify one of the groups of upper outliers and lower outliers. This phenomenon occurred in five out of nine sets of experimental data.

• **Phenomenon E:** The Random Forest outlier detection method may exhibit a phenomenon where the second to maximum value is determined as an outlier, but the maximum value is not determined as an outlier. This phenomenon occurred in seven out of nine sets of experimental data.

For Phenomenon A, it may be because the density of data points near the second to maximum value does not reach the density threshold of DBSCAN, so it is not recognized as a cluster. That is, the number of points in the neighborhood of the second to maximum value reaches *MinPts*, so the algorithm considers it as a core object, not an outlier.

For Phenomenon B, because the Z-Score method identifies outliers based on the mean and standard deviation of the data. If the number of trials in the binomial distribution data is large enough, and the probability of success is close to 0.5, then the distribution of these data may be close to the normal distribution, so the Z-Score method may have a better effect.

For Phenomenon C, because the IQR method identifies outliers based on the quartiles of the data. If there are a large number of duplicate values in your data, these duplicate values may affect the calculation of the quartiles, making the IQR method not sensitive enough when identifying outliers. But if the dispersion of the data sample is not high, that is, the difference between data points is not large, then the IQR method may still effectively identify outliers.

For Phenomenon D, both the Isolation Forest and Random Forest algorithms are designed to handle high-dimensional, continuous data. Therefore, they may not be able to identify all outliers well when dealing with discrete data or low-dimensional data. Also, if the distribution of upper and lower outliers in the data is uneven, or if the outliers on one side are more extreme, then the algorithm may only be able to detect outliers on one side.

For Phenomenon E, because the Random Forest is based on the distribution of training data to identify outliers. If the maximum value in the training data is similar to the distribution of other data points, and the second to maximum value has a large difference from the distribution of other data points, then during the model training process, the second to maximum value may be identified as an outlier, while the maximum value is considered as a normal value.

6 Discussion

This paper evaluates different outlier detection methods for samples of different distribution types, summarizing the effects of different outlier detection methods in various types of samples. However, it is worth noting that the samples in this paper are all randomly generated sample data using R language, which do not have missing values or error values. The distribution of real data may not fully comply with these theories, but is influenced by many complex factors. Therefore, when different outlier detection methods are applied to real data, they usually cannot completely coincide with the phenomena and conclusions described in this paper. When actually choosing an outlier detection method, one should first compare which distribution type the real data is closer to, and then refer to the conclusions of this paper for selection.

In addition, this paper only selected one parameter for each different outlier detection method to evaluate its effectiveness, which may lead to the phenomena and conclusions drawn in this paper being one-sided and incomplete. The selection of outlier detection methods is a topic worth exploring in depth. Summarizing the characteristics and phenomena of various outlier detection methods under different parameters can better help us choose the appropriate outlier detection method.

7 Conclusion

In this study, we conducted outlier detection on sample data from five continuous probability distributions (Normal, Chi-square, Exponential, Gamma, and T distributions) and four discrete probability distributions (Binomial, Poisson, Geometric, and Hypergeometric distributions). We employed five outlier detection methods, namely Z-Score, IQR, DBScan, Isolation Forest, and Random Forest, and evaluated their characteristics. Through comparison and analysis, we summarized the characteristics of various outlier detection methods in samples of different distribution types. These findings will assist us in making more rational method choices when facing different outlier detection scenarios.

The focus of this paper's further research is to summarize the characteristics and phenomena of various outlier detection methods under different parameters for specific types of data. When performing outlier detection on real data samples, the outlier detection methods under fixed parameters may not achieve the results we want. Therefore,

we need to delve into the methods of obtaining the characteristics of various outlier detection methods under different parameters, including but not limited to using accuracy, recall, F1 score and other indicators for performance evaluation [22, 23], and using sensitivity analysis techniques [24, 25], etc.

Acknowledgments

The author expresses gratitude for the inspiration provided by the lectures of Prof. Dr. Alexey Kharin from Belarusian State University for this paper.

Funding

Not applicable.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author Contributions

Not applicable.

About the Authors

SONG, Qingqing

Faculty of Applied Mathematics and Computer Science; Belarusian State University; 4 Nezavisimosti Avenue, Minsk 220030, Belarus; e-mails: fpm.sunC@bsu.by

XIA, Shaoliang

Faculty of Applied Mathematics and Computer Science; Belarusian State University; 4 Nezavisimosti Avenue, Minsk 220030, Belarus; e-mails: fpm.sya@bsu.by

References

[1] Boukerche, A., Zheng, L., & Alfandi, O. (2020). Outlier detection: Methods, models, and classification. ACM Computing Surveys (CSUR), 53(3), 1-37.

[2] Shiffler, R. E. (1988). Maximum Z scores and outliers. The American Statistician, 42(1), 79-80.

[3] Larson, M. G. (2006). Descriptive statistics and graphical displays. Circulation, 114(1), 76-81.

[4] Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014, February). DBSCAN: Past, present and future. In The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014) (pp. 232-238). IEEE.

[5] Al Farizi, W. S., Hidayah, I., & Rizal, M. N. (2021, September). Isolation forest based anomaly detection: A systematic literature review. In 2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE) (pp. 118-122). IEEE.

[6] Mensi, A., Cicalese, F., & Bicego, M. (2022, May). Using Random Forest Distances for Outlier Detection. In International Conference on Image Analysis and Processing (pp. 75-86). Cham: Springer International Publishing.

[7] Weisstein, E. W. (2002). Normal distribution. https://mathworld. wolfram. com/.

[8] Wilson, E. B., & Hilferty, M. M. (1931). The distribution of chi-square. Proceedings of the National Academy of Sciences, 17(12), 684-688.

[9] Balakrishnan, K. (2019). Exponential distribution: theory, methods and applications. Routledge.

[10] Nadarajah, S. (2011). The exponentiated exponential distribution: a survey. AStA Advances in Statistical Analysis, 95, 219-251.

[11] Thom, H. C. (1958). A note on the gamma distribution. Monthly weather review, 86(4), 117-122.

[12] Lukacs, E. (1955). A characterization of the gamma distribution. The Annals of Mathematical Statistics, 26(2), 319-324.

[13] Ahsanullah, M., Kibria, B. G., & Shakil, M. (2014).

Normal and student's t distributions and their applications (Vol. 4). Paris, France:: Atlantis Press.

[14] Weisstein, E. W. (2001). Student's t-Distribution. https://mathworld. wolfram. com/.

[15] Lange, K. L., Little, R. J., & Taylor, J. M. (1989). Robust statistical modeling using the t distribution. Journal of the American Statistical Association, 84(408), 881-896.

[16] Edwards, A. W. F. (1960). The meaning of binomial distribution. Nature, 186(4730), 1074-1074.

[17] Clarke, R. D. (1946). An application of the Poisson distribution. Journal of the Institute of Actuaries, 72(3), 481-481.

[18] Philippou, A. N., Georghiou, C., & Philippou, G. N. (1983). A generalized geometric distribution and some of its properties. Statistics & Probability Letters, 1(4), 171-175.

[19] Skibinsky, M. (1970). A characterization of hypergeometric distributions. Journal of the American Statistical Association, 65(330), 926-929.

[20] Harkness, W. L. (1965). Properties of the extended hypergeometric distribution. The Annals of Mathematical Statistics, 36(3), 938-945.

[21] Kemp, C. D., & Kemp, A. W. (1956). Generalized hypergeometric distributions. Journal of the Royal Statistical Society Series B: Statistical Methodology, 18(2), 202-211.

[22] Goutte, C., & Gaussier, E. (2005, March). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In European conference on information retrieval (pp. 345-359). Berlin, Heidelberg: Springer Berlin Heidelberg.

[23] Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In Proceedings of the first workshop on evaluation and comparison of NLP systems (pp. 79-91).

[24] Borgonovo, E., & Plischke, E. (2016). Sensitivity analysis: A review of recent advances. European Journal of Operational Research, 248(3), 869-887.

[25] Kleijnen, J. P. (1995). Sensitivity analysis and related analysis: A survey of statistical techniques.