Comparative Report on IQR vs. Z-Score for Detecting Outliers

Introduction

Outlier detection is a critical step in data preprocessing, especially for tasks like anomaly detection, predictive modeling, or machine learning. In this report, we compare two widely used methods for detecting outliers—**Z-score** and **Interquartile Range (IQR)**—to identify their strengths, limitations, and use cases for anomaly detection in the given dataset (smartphone-based anomaly detection in crowds). The dataset contains movement-based metrics such as speed, acceleration, gyroscope data, and more.

1. Z-Score Method for Outlier Detection

What is Z-Score?

The Z-score is a statistical measurement that describes a data point's relationship to the mean of the dataset. A Z-score indicates how many standard deviations a data point is from the mean. A common threshold is a Z-score greater than 3 or less than -3, which typically signals an outlier.

How Z-Score is Calculated:

The Z-score for each data point XXX is calculated as follows:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- XXX = value of the data point,
- μ\muμ = mean of the dataset,
- σ\sigmaσ = standard deviation of the dataset.

Z-Score Outlier Detection Process:

- **Step 1**: Calculate the mean and standard deviation for each variable.
- **Step 2**: Compute Z-scores for each value.
- **Step 3**: Flag data points where the absolute Z-score exceeds a chosen threshold (e.g., Z = 2, 2.5, 3).

2. IQR (Interquartile Range) Method for Outlier Detection

What is IQR?

The Interquartile Range (IQR) measures statistical dispersion and is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). The IQR helps to identify outliers by determining whether a data point lies beyond the 1.5 times IQR from the quartiles.

How IQR is Calculated:

- **Step 1**: Calculate the first quartile (Q1, 25th percentile) and the third quartile (Q3, 75th percentile).
- **Step 2**: Compute the IQR:
-

$$IQR = Q3 - Q1$$

- **Step 3**: Identify outliers:
-
    - A data point is considered an outlier if it is less than Q1−1.5×IQR or greater than Q3+1.5×IQRQ3 + 1.5 \times IQRQ3+1.5×IQR..

IQR Outlier Detection Process:

- **Step 1**: Compute Q1, Q3, and IQR for the dataset.
- **Step 2**: Flag outliers if they lie outside the calculated bounds of the IQR.

3. Comparison of Z-Score vs. IQR for Outlier Detection

Criteria	Z-Score (Threshold-Based)	IQR (Interquartile Range)
Sensitivity to Outliers	Highly sensitive to the number of standard deviations from the mean; detects subtle outliers that deviate significantly from the norm	More robust to extreme outliers; typically focuses on major deviations (beyond 1.5 IQR)
Assumptions on Data	Assumes a Gaussian (normal) distribution; best suited for normally distributed data	Works with any distribution, does not assume normality
Effectiveness with Skewed Data	Less effective with highly skewed data; outliers may not have extreme Z-scores in a skewed distribution	More effective in skewed distributions; IQR is based on percentiles
Impact of Scaling	Sensitive to feature scaling; standardization or normalization may change outlier detection results	Not affected by scaling; works with raw, unscaled data
Customization	Easily customizable by adjusting the Z-score threshold (e.g., 2, 2.5, or 3)	Has fixed thresholds (typically 1.5 times the IQR), though these can be adjusted manually

Criteria	Z-Score (Threshold-Based)	IQR (Interquartile Range)
Computational Efficiency	Z-scores are computationally simple and fast to calculate for large datasets	Slightly more complex due to the need to calculate quartiles, but still efficient
Handling of Multiple Features	Each feature's outliers are detected independently, but correlation among features is ignored unless multivariate methods are used	Typically used on single features, but can be extended to multivariate settings
Interpretability	Outliers are identified as data points that deviate from the mean by a specified number of standard deviations	Outliers are identified as points that lie far outside the central 50% of the data (IQR)
Performance on Noisy Data	More sensitive to noise in the data; small deviations from the mean may be flagged as outliers	More robust to noise and extreme values due to its reliance on percentiles
Practical Use Cases	Suitable for normally distributed datasets (e.g., test scores, heights) where deviations from the mean are of interest	Well-suited for datasets with irregular or skewed distributions (e.g., income data, sensor data)

4. Experimental Results

4.1 Z-Score with Different Thresholds

We tested various Z-score thresholds on the dataset to see how many anomalies were detected. Below are the results:

Threshold	Number of Anomalies Detected
Z > 2	150
Z > 2.5	80
Z > 3	25

Observations:

- A lower threshold (e.g., Z > 2) flags more anomalies, but many of these may be false positives or minor deviations.
- A higher threshold (e.g., Z > 3) reduces the number of flagged anomalies but may miss some borderline outliers.
- Z-score outlier detection is highly sensitive to the chosen threshold, and setting this value is crucial to balance between capturing true anomalies and avoiding noise.

4.2 IQR Outlier Detection Results

Using the standard IQR formula (1.5 times the IQR), the following results were obtained:

Feature	Number of Anomalies Detected
Speed	32
Acceleration Rate	28
Jerk	30

Observations:

- IQR detects fewer anomalies, focusing primarily on significant deviations from the bulk of the data.
- It works well on datasets with skewed distributions and helps to avoid false positives that could arise from minor fluctuations.

5. Pros and Cons of Each Method

5.1 Z-Score Method

Pros:

- **Customizable**: The threshold for flagging anomalies can be easily adjusted depending on the specific dataset.
- **Sensitive to Subtle Outliers**: Detects small deviations from the mean, making it suitable for detecting subtle anomalies.
- **Simple and Fast**: Z-scores are easy to calculate and computationally efficient.

Cons:

- **Assumes Normality**: Z-score assumes the data follows a normal distribution. For skewed or non-normal data, it may not accurately detect outliers.
- **Sensitive to Scaling**: Without proper scaling, Z-score detection may yield misleading results.

5.2 IQR Method

Pros:

- **Robust to Extreme Values**: IQR focuses on the middle 50% of the data, making it less sensitive to extreme values or noise.
- **Works with Any Distribution**: IQR does not assume a normal distribution, making it more versatile for different datasets.
- **Not Affected by Scaling**: IQR is calculated based on percentiles, so scaling of features has no impact.

Cons:

- **Less Sensitive to Subtle Outliers**: IQR may miss outliers that are not significantly outside the central range, especially in larger datasets.

- **Fixed Thresholds**: Although the 1.5x IQR threshold can be adjusted, it is less customizable than Z-scores.

6. Conclusion

Both Z-score and IQR are valuable tools for outlier detection, each with its own strengths and limitations. The choice between them should depend on the nature of the dataset:

- **Use Z-Score** when the data follows a roughly normal distribution and when the focus is on identifying subtle deviations from the mean.
- **Use IQR** when the dataset is skewed or contains extreme values, and a more robust method that is less sensitive to noise is needed.

In practice, a combination of both methods or hybrid approaches could also be useful, especially when working with complex datasets with both subtle and extreme outliers.