Outlier's detections and removal

Inbar Barkai and Noam Sery Levi

March 2023

1 Abstract:

In the Data Science field, the goal is to analyze large amounts of data by using scientific methods and extract relevant conclusions and knowledge from it. A common problem with datasets is detecting and removing outliers. Outliers are data points that differ significantly from other data points in the dataset. They can be caused by a variety of factors, such as measurement error, data entry errors, or they could represent a legitimate deviation from the norm. Outliers can skew the distribution of the data, making it difficult to determine the central tendency and variability of the dataset. This can result in incorrect estimates of summary statistics. In order to address this issue, we tried to identify outliers from the dataset in an efficient way. To accomplish this task, we tried to improve modified z-score, which is capable of detecting outliers in any type of dataset (in contrast to z-score that effective when analyzing normally distributed data). To evaluate the effectiveness of our proposed method, we conducted an experiment on real-world datasets. The results show that our algorithm performs better than modified z-score in certain scenarios. However, modified z-score remains more effective in detecting outliers when compared to our approach.

2 Problem description:

The modified z-score is a statistical method used for detecting outliers in datasets. It is commonly used in data science and machine learning applications to preprocess datasets before analysis. Unlike the standard z-score, which is calculated using the mean and standard deviation of the data, the modified z-score is calculated using the median absolute deviation (MAD) as a measure of dispersion.

Modified z-score calculation: subtract the median of the dataset from the data point, and then divide the result by the MAD multiplied by a scaling factor (typically 1.4826). The MAD is the median of the absolute deviations from the median. To calculate the MAD, subtract the median from each data point and take the absolute value, then calculate the median of these absolute deviations.

A modified z-score greater than a certain threshold (usually 3.5 or 4) indicates that the data point is an outlier.

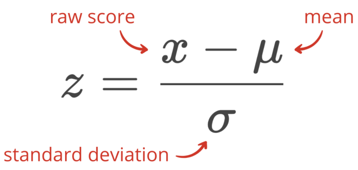


Figure 1: Z-score formula.

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$$

Figure 2: Modified z-score formula.

Modified z-score is considered to be more robust than the standard z-score as it is calculated from the median absolute deviation (MAD). However, relying on the median may not accurately represent the center of the distribution if the dataset contains extreme values or is heavily skewed.

Therefore, the modified z-score algorithm can be significantly biased in the presence of outliers, rendering it inaccurate and unreliable for analysis. This paper aims to address this method's limitations and creates a solution.

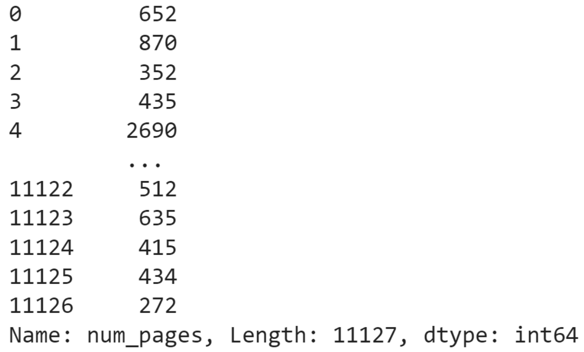
3 Solution overview:

Our proposed solution to the problem described above, consists of several steps:

1. Use a modified z-score to detect outliers in the dataset.
2. Remove the outliers from the dataset.
3. Calculate the median absolute deviation (MAD).
4. Repeat the process until the MAD converges to a certain value.

Note that the calculation of the MAD is required as part of running the modified z-score algorithm.

Explanation about the course of the algorithm: We will showcase a demonstration of the algorithm by running it on a specific example. This will provide a better understanding of how the algorithm operates. Let's look at 'num pages' column from 'books' dataset:



```
0         652
1         870
2         352
3         435
4        2690
         ...
11122     512
11123     635
11124     415
11125     434
11126     272
Name: num_pages, Length: 11127, dtype: int64
```

First iteration:

1. Calculate the **median** of the dataset: 299.0

2. Calculate the **median absolute deviation (MAD)** of the dataset: 107.0

3. Calculate the **modified z-score** for each data point, using the following formula: 0.6745 * (x - median) / 1.4826 * MAD.

4. **Identify outliers:** 87 outliers were identified.

5. Remove the outliers and repeat the steps.

Second iteration:

1. The median is 295.0

2. The MAD is 105.0

3-4. 15 outliers were identified.

5. Remove the outliers.

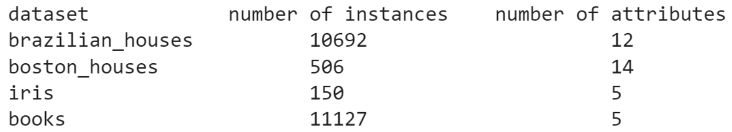
Third iteration:

1. The median is 294.0

2. The MAD value remains at 105.0, indicating that it has reached convergence. Therefore, there is no need to proceed further and the function should return. At this instance, the process was concluded after two iterations.

4 Experimental evaluation:

To evaluate the correctness of the suggested algorithm, we performed several experiments on different datasets. The details about those datasets are presented in the following table:



```
dataset            number of instances    number of attributes
brazilian_houses          10692                   12
boston_houses             506                     14
iris                      150                     5
books                     11127                   5
```

For each dataset, we trained 3 linear regression models:

1. Without removing any outliers.
2. With outliers removed by our algorithm.
3. With outliers removed by the original z- score algorithm.

We compare the performance between those models by 3 metrics: mean square error (MSE), Root Mean Squared Error (RMSE) and r2 score. The results of our analysis will present later.

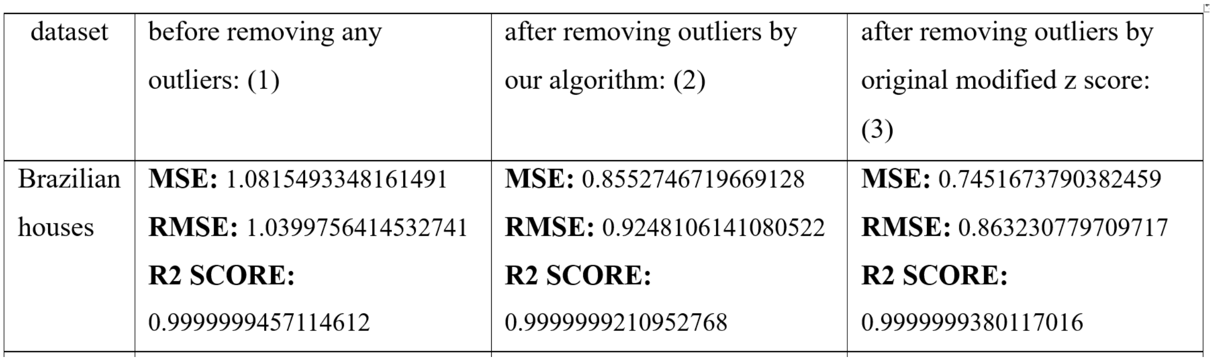
The evaluation process:

Linear regression with one dimension:

First, we look at the relationship between 'propertyTax' and 'total' columns of the 'brazilian houses' dataset (there is a graph in the notebook). We then train a linear regression model and check the fit between the model and the data by r2 score. We got a lower r2 score. This led us to train a linear regression model with more dimensions.

Linear regression with more dimensions:

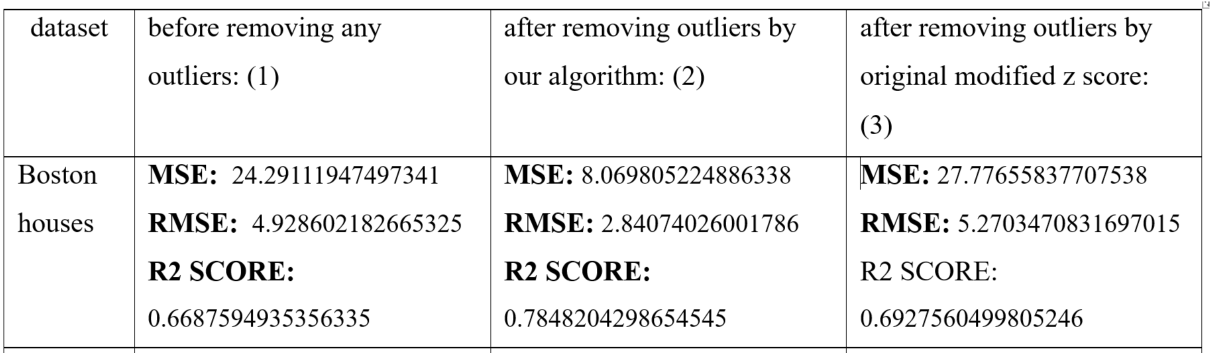
We defined Y to be 'total' column and X to be the rest of the columns. We then train 3 linear regression models as explained above. The results:



dataset	before removing any outliers: (1)	after removing outliers by our algorithm: (2)	after removing outliers by original modified z score: (3)
Brazilian houses	**MSE:** 1.0815493348161491 **RMSE:** 1.0399756414532741 **R2 SCORE:** 0.9999999457114612	**MSE:** 0.8552746719669128 **RMSE:** 0.9248106141080522 **R2 SCORE:** 0.9999999210952768	**MSE:** 0.7451673790382459 **RMSE:** 0.863230779709717 **R2 SCORE:** 0.9999999380117016

We can see that the performance of Model 2 is better than the performance of Model 1. However, the performance of Model 3 is better than the performance of Model 2. That means that the original modified z-score algorithm was better than our algorithm, although using our algorithm on the dataset was better than using the data without removing outliers at all.

The results of applying the same operations to the remaining datasets are shown in the following table:



dataset	before removing any outliers: (1)	after removing outliers by our algorithm: (2)	after removing outliers by original modified z score: (3)
Boston houses	**MSE:** 24.29111947497341 **RMSE:** 4.928602182665325 **R2 SCORE:** 0.6687594935356335	**MSE:** 8.069805224886338 **RMSE:** 2.84074026001786 **R2 SCORE:** 0.7848204298654545	**MSE:** 27.77655837707538 **RMSE:** 5.2703470831697015 R2 SCORE: 0.6927560499805246

Iris	**MSE:** 0.03723364456197505 **RMSE:** 0.19296021497183052 **R2 SCORE:** 0.9467245149351708	**MSE:** 0.03723364456197505 **RMSE:** 0.19296021497183052 **R2 SCORE:** 0.9467245149351708	**MSE:** 0.03723364456197505 **RMSE:** 0.19296021497183052 **R2 SCORE:** 0.9467245149351708
Books	**MSE:** 0.1175628888758851 **RMSE:** 0.3428744506023817 **R2 SCORE:** 0.02806642763021283	**MSE:** 0.1317994573210613 **RMSE:** 0.3630419498089185 **R2 SCORE:** 0.02116932252206183	**MSE:** 0.11433284100624533 **RMSE:** 0.3381313960670398 **R2 SCORE:** 0.018960298500731332

We observed that in the Boston houses dataset, our algorithm performed better than both the original z-score and the dataset without outlier removal. In the Iris dataset, where there were no outliers, the results remained the same. However, in the Books dataset, our algorithm performed worse than both the original modified z-score and the dataset without outlier removal. Note that the 'boston houses' dataset is the most heavily skewed of the four datasets and has extreme values (the dataset has extreme values in the 'CRIM' column, and the distribution of this column is heavily skewed to the right- An illustration for this distribution can be seen in the notebook).

5 Related work:

While choosing a topic for this project, we reviewed many articles. From which we came across the article "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median" from which we took inspiration for the project theme. While creating the algorithm, we were assisted by the article "Outlier Detection: How to Threshold Outlier Scores?". The article discusses that many methods which utilize thresholding techniques for detecting outliers, rely on statistics. However, the article argues that such statistics can be biased due to the presence of outliers within the data. To overcome this problem, the article proposes a solution called the 2T (two-stage thresholding), which is a technique for calculating threshold outlier scores by removing most of the outliers at the first stage by using a more conservative threshold, and the same process is then repeated for the processed scores. We got inspiration from 2T algorithm by writing our solution. However, the way our algorithm runs is different. For instance, within our solution, the algorithm calculates the median absolute deviation (MAD), removes the outliers, and repeats the process until the MAD converges. Whereas 2T, stops after two iterations. In addition, our solution expands modified z-score algorithm, whereas the article dealing with a solution to calculate statistics.

6 Conclusion:

Throughout our work, we were exposed to a variety of methods for detecting outliers. We focused on modified z-score and tried to improve it by Recalculation of the MAD (median absolute deviation). We had anticipated that the solution we executed would lead to improved performance by detecting and removing outliers. Based on our experimental results, it appears that in most cases, the original modified z-score algorithm performed better than our algorithm. Our algorithm seems to be effective only in specific scenarios and may not always yield the best results. We assume that our algorithm is better suited for data with extreme values and heavily skewed distributions. Therefore, it is generally advisable to use the original modified z-score algorithm.