ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 5, Issue 4, April 2018

Comparative Analysis of Outlier Detection Methods

[1] Mahvish Fatima, [2] Jitendra Kurmi

[1] Pursuing M.Tech at BBAU, Lucknow, [2] Assistant Professor at BBAU, Lucknow

Abstract: - **This paper presents different types of outlier detection methods. There are many factors which generate outliers it can be a measurement error, instrumentation error etc. No matter whatever is the reason outliers can be identified using these methods effectively. The experiment was conducted on SPSS tool. The comparative analysis of these methods was performed successfully. The aim of this paper was to analyze the methods of outlier detection from the research point of view. Furthermore, we find out the impact of different methods on the dataset containing outliers.**

***Keywords*: Box and Plot, MADe, Outliers, Z-Score.**

I. INTRODUCTION

Outliers are the mismatch in the pattern of original pattern. They can provide useful information which can actually help in improving the results of analysis while it can also provide harm to the analysis of the data. There are several methods available for outlier detection. Each method has its own advantage and disadvantage. This paper will present the comparative analysis of the some known labeling methods present for outlier detection. Outlier detection is performed nowadays in almost every field for e.g., fraud detection, finance, tourism etc. [2] They play vital role in detecting and removing the negative aspects of any sector. By working on the outliers we can achieve really useful information.

II. INTRODUCTION TO OUTLIERS

Outlier is the pattern or thing which is different from others or which deviates from the normal design. In general terms, we can say that outliers are the dislocation of the original points from the normal position. There can be normal outliers as well as extreme outliers. There are distinct causes for occurrence of a outliers. It can be a measuring apparatus fault, human calculation error, fraudulent behavior, or natural deviations in the graph. It can also be the result of a wrong hypothesis generated by the researcher. A fault in datasets which generates the wrong result is also the outcome of an outlier. There are several methods available for outlier's detection but the type of method which is being applied depends on the nature of the desired outliers:

a) Point Outliers
b) Contextual Outliers
c) Collective Outliers

A. Point Outliers

If any individual data point is found to be anomalous with respect to other data points, then the point is said to be point Outlier. This is the simplest type of outlier and most of the research work is going on it.

B. Contextual Outliers

If the data point is anomalous in a specified context but not as a whole then this type of outliers are termed as contextual outliers. The context of the data point is defined by the structure of the data set. Data instance will be classified into two attributes- Behavioral Attributes and Contextual Attributes.

C. Collective Outliers

If the collected data points are found to be anomalous with respect to the entire data set, it is termed as a collective outlier. The separate data points in collective outliers may not be a outlier individually, but collectively they are anomalous.

III. DATASETS

In this paper, dataset is from Ministry of Tourism, India. It is tourism data based on medical tourism.

Methodologies There are many informal and formal methods present in the data mining for detecting outliers. But, here I am going to present few of them and conclude the comparative analysis of all of them to show the better one for detecting outliers. List of the informal methods for outlier detection are:

a) MADe (Median Absolute Deviation)
b) Z-score
c) Z-Score Modified
d) Box and Plot Method

MADe (Median Absolute Deviation)

Median Absolute Deviation is a measure of variation of the data. It measures variation by finding the value where half of the data is in proximity to the median and half of the data is distant from the median. [1] MAD is the short form for Median Absolute Deviation which is not to be confused with Mean Absolute Deviation because it also

IFERP connecting engineers... developing research
ISSN (Online) 2394-2320
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

has the same short form. [3] It is obtained by finding the median of the absolute values of the deviations of the data values from the median.

The formula for the median is

$$MAD = median(|X_i - median(X_j)|)$$

And for the normal estimated value

$$MAD = 1.486*median(|X_i - median(X_j)|)$$

The steps for finding the MAD are:

i. Find the median
ii. Subtract the Median from each Xi-value using the formula (Xj-Median)
iii. Find the median of the absolute difference.

This method is a robust method for measures of spread as compared to other methods because it is less vulnerable to the outliers.

B. Z-Scores

It is a measure of the value is how many standard deviations away from the mean or how many standard deviations below or above the population mean is. It is also termed as Standard Score. It ranges from -3standard deviation to +3 standard deviation on the normal distribution curve. It is used to compare test scores to a "normal" score. Z-score can help in comparing the test result from the available data.

The formula for calculating Z-Score is

$$Z = \frac{(x-\mu)}{\sigma}$$

Where μ= mean
σ= standard deviation
x= test value

Z-Score formula for sample mean is

$$Z = \frac{(x_i - \bar{x})}{s}$$

Where xi= test
$\bar{x}$= sample mean
s= sample standard deviation

A Z-score gives the information about where the score lies on a normal distribution curve. A zero value tells us the value is exactly average while +3 tells value is much higher than average.

C. Modified Z-Score

It is a measurement of the outlier's strength by checking the dependability of a particular score on the typical score. This method is more robust than the Z-Score method when we talk about outliers because it relies on median not mean. It is calculated from the median absolute deviation. After that they are multiplied by a constant to approximate the standard deviation.

The formula for Modified Z-Sore is

$$M_i = 0.6745(x_i - \bar{x}_i)$$

Where xi = observation
$\bar{x}_i$ = mean
MAD = median absolute deviation

D. Box and Plot

It tells us about the center and the spread of the data which allow us to identify outliers in a more appropriate manner. It is a graphical equivalent of the five number summaries. It is also known as Box Whisker Diagram. It consists of two components first a box and two whiskers. It eases out the comparisons of different variables. The ends of the box show the minimum and maximum scores. The circle (o) shows the normal outliers and the star (*) shows the extreme outliers.

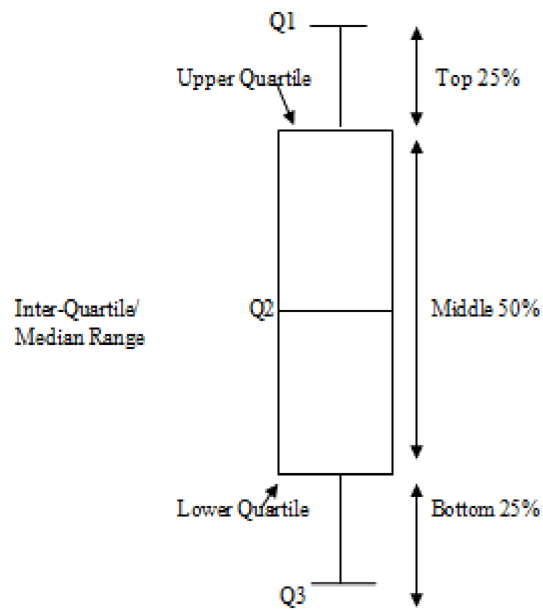

Fig. 1 Box and Whisker Plot Diagram

This box is divided into four equal parts. The top whisker is 25% and bottom is also 25% and the middle whisker is 50%. The standard formula for detecting an outlier through this method is

Q1 – (1.5×IQR)
Q3 + (1.5×IQR)
Where IQR=Q3-Q1.

IV. ANALYSIS AND RESULTS

The analysis has been performed on the dataset of medical tourism from ministry of tourism to identify outliers in the data. Sometimes just a visual scroll through the data helps us to detect the outliers easily. But this is not the case with every outlier. So for detecting outliers here we have applied four different methods on the same dataset to perform analysis and generate the results. The tool on which we have worked is SPSS. The dataset has total 4 attributes in which there are 2 nominal and 2 scale values were present. The instances having missing values have been removed for better accuracy. After completing the data cleaning phase, we applied the first outlier detection method which is Box and Whisker Plot method. In the output box it generated the descriptive statistics. By looking into statistics we find out that it has a high value of skewness and kurtosis which indicates that there are many outliers in the dataset. Fig. 2 shows the descriptive for the Box and Plot method.

Descriptives

			Statistic	Std. Error
MEDICALATTENDANT	Mean		328.59	85.914
	95% Confidence Interval for Mean	Lower Bound	158.91	
		Upper Bound	498.27	
	5% Trimmed Mean		111.87	
	Median		8.00	
	Variance		1180983.451	
	Std. Deviation		1086.731	
	Minimum		0	
	Maximum		7861	
	Range		7861	
	Interquartile Range		65	
	Skewness		4.536	.192
	Kurtosis		22.449	.381

Fig. 2 Descriptives Statistics for Box and Plot Method

Then we have generated the stem-and-leaf plot for the same attributes as shown in Fig. 3. This plot indicates whether outliers are present in the data. It shows 26 "extreme" values at the upper end of the distribution that is greater than or equal to 233.

MEDICALATTENDANT

```
MEDICALATTENDANT Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

    85.00        0 .  0000000000000000000000000011111111111111222222222222222233333333344444445557788889999
    19.00        1 .  0000011222246667899
     5.00        2 .  11444
     3.00        3 .  688
     5.00        4 .  04669
     1.00        5 .  1
     3.00        6 .  346
     4.00        7 .  3558
     2.00        8 .  18
      .00        9 .
      .00       10 .
     1.00       11 .  4
     4.00       12 .  2469
      .00       13 .
     1.00       14 .  0
     1.00       15 .  8
    26.00 Extremes    (>=233)

 Stem width:       10
 Each leaf:        1 case(s)
```

Fig. 3 Stem-and-Leaf Plot for Box and Plot Method

Second Method which we have applied is Z-Score method to detect outliers. Fig. 4 shows the descriptive values for the Z-Score. In this statistics we can see that mean, median and trimmed mean are nearly same. This shows that distribution is not skewed in one direction. But here also it has a high value of skewness and kurtosis which indicates outliers. Now let us see the stem-and-leaf plot of this descriptive. Fig. 5 shows the stem-and-leaf plot of Z-Score analysis. In this plot we can clearly see that it also has 26 "extreme" values at the upper end of the distribution that is greater than or equal to -.1132.

Descriptives

			Statistic	Std. Error
ZMEDICALATTENDANT	Mean		-.0692969	.03777360
	95% Confidence Interval for Mean	Lower Bound	-.1438996	
		Upper Bound	.0053058	
	5% Trimmed Mean		-.1268515	
	Median		-.1385635	
	Variance		.228	
	Std. Deviation		.47780244	
	Minimum		-.13947	
	Maximum		5.76345	
	Range		5.90291	
	Interquartile Range		.00727	
	Skewness		11.640	.192
	Kurtosis		142.064	.381

Fig. 4 Descriptives Statistics for Z-Score Method

ZMEDICALATTENDANT

```
ZMEDICALATTENDANT Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

    73.00     -139 .  0000000111111112222222222222222223333333333333333444444444444444444444444
    23.00     -138 .  11112233333444455566999
    10.00     -137 .  0033456668
     3.00     -136 .  777
     3.00     -135 .  114
     4.00     -134 .  2259
     2.00     -133 .  79
     3.00     -132 .  023
     3.00     -131 .  002
     2.00     -130 .  36
     1.00     -129 .  5
      .00     -128 .
      .00     -127 .
     1.00     -126 .  6
     3.00     -125 .  247
     1.00     -124 .  9
     1.00     -123 .  6
      .00     -122 .
     1.00     -121 .  6
    26.00 Extremes    (>=-.1132)

 Stem width:    .00100
 Each leaf:       1 case(s)
```

Fig. 5 Stem-and-Leaf Plot for Z-Score Method

Third method which we have applied is Modified Z-Score method. The descriptive for the following method is shown in Fig. 6 Descriptive Statistics. Here we can clearly see that the mean, median and trimmed mean is not same which indicates that skewness is not skewed in one direction while kurtosis is indicating that it has outliers. Now let us have a look at the stem-and-leaf plot of Modified Z-Score shown in Fig. 7. It is also showing 26 "extremes" value at the upper end of the distribution that is greater than or equal to 38. This is the same case as in the other methods.

Descriptives

			Statistic	Std. Error
Modified_Z_Score	Mean		54.0591	14.48718
	95% Confidence Interval for Mean	Lower Bound	25.4469	
		Upper Bound	82.6712	
	5% Trimmed Mean		17.5159	
	Median		.0000	
	Variance		33580.545	
	Std. Deviation		183.24995	
	Minimum		-1.35	
	Maximum		1324.21	
	Range		1325.56	
	Interquartile Range		10.88	
	Skewness		4.536	.192
	Kurtosis		22.449	.381

Fig. 6 Descriptives Statistics for Modified Z-Score

Modified_Z_Score

```
Modified_Z_Score Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

    78.00       -0 .  000000000000000000001111111111111111111111111111111111111111111111111111111111
    26.00        0 .  00000000000000000001111111
     5.00        0 .  22222
     4.00        0 .  4555
     5.00        0 .  66667
     3.00        0 .  999
     4.00        1 .  0111
     2.00        1 .  23
      .00        1 .
     1.00        1 .  7
     3.00        1 .  999
     1.00        2 .  0
     1.00        2 .  2
     1.00        2 .  5
    26.00 Extremes    (>=38)

 Stem width:    10.00
 Each leaf:       1 case(s)
```

Fig. 7 Stem-and-Leaf Plot for Z-Modified Method

Lastly, we have applied MADe (Median Absolute Deviation) method to detect outliers. The descriptives for the MAD is shown in Fig. 8. Here, also the skewness and kurtosis value is quite large which indicates outliers. But the stem-and-leaf plot given in the Fig. 9 makes a difference here. We can clearly see that the "extreme" values in this case are 27 at the upper end of the distribution which is greater than or equal to 150.

Descriptives

			Statistic	Std. Error
Median_Deviation	Mean		326.7500	85.76746
	95% Confidence Interval for Mean	Lower Bound	157.3596	
		Upper Bound	496.1404	
	5% Trimmed Mean		110.2431	
	Median		8.0000	
	Variance		1176969.145	
	Std. Deviation		1084.88209	
	Minimum		.00	
	Maximum		7853.00	
	Range		7853.00	
	Interquartile Range		51.50	
	Skewness		4.544	.192
	Kurtosis		22.516	.381

Fig. 8 Descriptives Statistics for Median Absolute Deviation

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Median_Deviation

```
Median_Deviation Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

   101.00        0 .  01112233444445555666666666777777788888888888888&
     8.00        1 .  136&
     1.00        2 .  &
     6.00        3 .  08&
     2.00        4 .  &
     3.00        5 .  &
     3.00        6 .  7&
     2.00        7 .  &
     1.00        8 .  &
      .00        9 .
     1.00       10 .  &
     3.00       11 .  &
     1.00       12 .  &
     1.00       13 .  &
    27.00 Extremes    (>=150)

 Stem width:      10.00
 Each leaf:        2 case(s)


 & denotes fractional leaves.
```

Fig. 9 Stem-and-Leaf Plot for Median Absolute Deviation Method

So, if we compare the experimental results of all these methods we find out that Box and Plot method, Z-Score method, and Modified Z-Score method generates the somewhat identical result. But the MAD method generated the quite different result than other methods it is because it uses median instead of mean. Therefore, it can be considered more accurate.

V. CONCLUSION

The study was conducted on datasets of tourism on which the analysis of four outlier detection labeling methods was carried out. Result concludes that every method has its own advantage. But, it will be better if we use MAD for detecting outliers. Furthermore studies can be performed on these methods to enhance the performance on datasets. A conclusion section is not required.

REFERENCES

1) Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, Laurent Licata J. U. Duncombe, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," Elsevier YJESP-03038; No. of pages: 3; 4C, March 10, 2013

2) Girish Kumar Sharma, Promila Sharma, "A Study on Data Mining algorithms for Tourism Industry," ILJET, vol. 7, issue 1, May 2016.

3) George Kurian, Hongmei Chi, "Predict Florida Tourism Trend via Using Data Mining Techniques," PEARC17, New Orleans, LA, USA, July 09-13, 2017.