A Review of various statestical methods for Outlier Detection

#Dhwani Dave,*Tanvi Varma

#Student
*Asst.Professor
Computer Science and Engineering
PIT
Vadodara, India

Abstract— Outlier Detection has many applications. Outlier is instance of the data set which has exceptional characteristics compare to other instance of data and exhibits unusual behavior of the system. There are many methods used for detecting outliers. Every method has its advantages and limitations. In this review paper a relative comparison of few statistical methods is carried out. This shows which method is more efficient in detecting outlier.

Keywords-outlier detection methods; mean; standard deviation; median absolute deviation; clever variance and clever mean

I. INTRODUCTION

Detecting outlier has always been challenging problem for researchers. Many methods have been used for detecting outliers. The popular methods used for detecting outliers are mean and standard deviation.

In next section the advantage and disadvantage of each method is described. Also there are other more efficient and robust methods such as (a) Median Absolute deviation and (b) Clever Mean and Clever Variance are also described. These methods give more efficient results compare to mean and standard deviation.

II. MEAN AND STANDARD DEVIATION

A. Mean

Mean of any given data set is derived as follows:

$$\overset{n}{\underset{i=1}{}} \quad {}_{i} / n \qquad (1)$$

It is the average value of any given data set. The reasons why it is considered a nonrobust estimator are as follows:

1) Mean value is highly biased even if there is a single outlier and 2) in a large data sets a mean value can be changed even though an outlier is removed. So, while using a mean value for detecting an outlier an outlier can be considered as a normal data point. This reduces the efficiency of the method and makes it a nonrobust estimator.

B. Standard Deviartion

The standard deviation (represented by the Greek letter sigma, **σ**) shows how much variation or dispersion from the average/mean exist [1].

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad (2)$$

A low standard deviation indicates that the data points tend to be very close to the mean (also called expected value); a high standard deviation indicates that the data points are spread out over a large range of values [2].

For this outlier detection method, the mean and standard deviation of the residuals are calculated and compared. If a value is a certain number of standard deviations away from the mean, that data point is identified as an outlier. The specified number of standard deviations is called the threshold. The default value is 3.

This method can fail to detect outliers because the outliers increase the standard deviation. The more extreme the outlier, the more the standard deviation is affected [3].

There are three problems to face using mean. 1) It assumes that the distribution is normal including outliers. 2) The mean and standard deviation are strongly influenced by outliers. 3) As stated by Cousineau and Chartier (2010), this method is very unlikely to detect outliers in small samples [4].

Reference [5] illustrates this problem with example given below:

Accordingly, this indicator is fundamentally problematic:

It is supposed to guide our outlier detection but, at the same time, the indicator itself is altered by the presence of outlying values. In order to appreciate this fact, consider a small set of n = 8 observations with values 1, 3, 3, 6, 8, 10, 10, and 1000. Obviously, one observation is an outlier (and we made it particularly salient for the argument). The mean is 130.13 and the uncorrected standard deviation is 328.80.Therefore, using the criterion of 3 standard deviations to be conservative, we could remove the values between− 856.27 and 1116.52. The distribution is clearly not normal (Kurtosis = 8.00; Skewness = 2.83), and the mean is inconsistent with the 7first values. Nevertheless, the va lue 1000 is not identfied as an outlier, which clearly demonstrates the limitations of the mean plus/minus three standard deviations method [5].

III. THE MEDIAN ABSOLUTE DEVIATION (MAD)

The MAD overcomes these problems. In [5], authors have illustrated the efficiency of MAD over mean and standard deviation with example. Which is given here as follows: The median (M) is, like the mean, a measure of central tendency but offers the advantage of being very insensitive to the presence of outliers. One indicator of this insensitivity is the “breakdown point” [6]. The estimator's breakdown point is the maximum proportion of observations that can be contaminated (i.e., set to infinity) without forcing the estimator to result in a false value (infinite or null in the case of an estimator of scale). For example, when a single observation has an infinite value, the mean of all observations becomes infinite; hence the mean's breakdown point is 0. By contrast, the median value remains unchanged. The median becomes absurd only when more than 50% of the observations are infinite. With a breakdown point of 0.5, the median is the location estimator that has the highest breakdown point.

Exactly the same can be said about the Median Absolute Deviation as an estimator of scale (see the formula below for a definition). Moreover, the MAD is totally immune to the sample size. These two properties have led [7] to describe the MAD as the “single most useful ancillary estimate of scale” (p. 107). It is for example more robust than the classical inter quartile range [8], which has a breakdown point of 25% only.

To calculate the median, observation has to be sorted in ascending order. Let us consider the previous statistical series: 1, 3, 3, 6, 8, 10, 10, and 1000. The average rank can be calculated as equal to (n + 1) / 2 (i.e., 4.5 in our example). The median is therefore between the fourth and the fifth value, that is, between six and eight (i.e., seven).

Calculating the MAD involves finding the median of absolute deviations from the median. the MAD is defined as follows [7]:

$$MAD = b\ M_i\left(\left|x_i - M_j\left(x_j\right)\right|\right) \tag{3}$$

where the x_j is the n original observations and M_i is the median of the series. Usually, b = 1.4826, a constant linked to the assumption of normality of the data, disregarding the abnormality induced by outliers (Rousseeuw & Croux, 1993).

Calculating the MAD implies the following steps:

(a) the series in which the median is subtracted of each observation becomes the series of absolute values of (1–7), (3–7), (3–7), (6–7), (8–7), (10–7), (10–7), and (1000–7), that is, 6, 4, 4, 1, 1, 3, 3, and 993; (b) when ranked, we obtain: 1, 1, 3, 3, 4, 4, 6, and 993; (c) and (d) the median equals 3.5 and will be multiplied by 1.4826 to find a MAD of 5.1891.

Let us briefly consider the case of a fictional series in Fig. 1, which includes a larger number of observations. Fig. 1a shows a normal distribution and reports the mean, SD and median. Fig. 1b shows the same distribution but with one value (=0.37) changed into an outlier (=3). The same indicators are reported and we can see that the mean and SD have drastically changed whereas the median remains the same.

To calculate MAD all the observations has to be sorted first. This can be a huge overhead in large data set.

It requires preprocessing and it can become time consuming and in highly dynamic data it may become more difficult to hold a correct value.

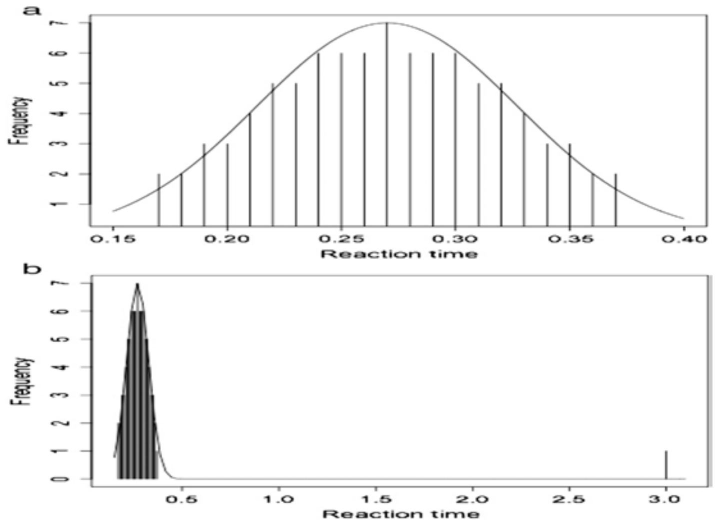

Fig. 1 Outlier generating asymmetry. a) Normal distribution, n = 91, mean = 0.27, median = 0.27, standard deviation = 0.06. b) Asymmetry due to an outlier, n = 91,mean = 0.39, median = 0.27, standard deviation = 0.59.

IV. Clever Mean and Clever Variance

Mean and standard deviation do not serve as a robust estimators. But a novel approach has been proposed that allows the mean and variance to be estimated while at the same time detecting possible outliers. These new estimators are called the clever mean and clever variance, respectively [9]. The method as described by [9] is as follows:

$$\text{sun}\ \ \underset{i=1}{\underline{n}}\ \ {}_{i} \qquad (4)$$

$$\text{s}\ \ \underset{i=1}{n}\ \ {}_{i} \qquad (5)$$

and a predetermined number of maximum and minimum values is collected

Let:

$$cm_0 = \frac{\sum^n_{i=1} y_i}{n} = \bar{y} \qquad (6)$$

denote the zeroth-order clever mean and

$$cv_0 = \frac{\sum^n_{i=1} (y_i - cm_0)}{n - 1} = s^2 \qquad (7)$$

the zeroth-order clever variance.

The selection of a possible outlier y_k*, is very simple indeed: it is whichever one that minimizes cv_k between the two observations that currently represent the minimum and the maximum values after the removal of previous outliers.

The clever mean might maintain its value while outliers are progressively removed for two reasons. Firstly, when there is a particularly large set of data, the arithmetic mean can change slightly even though an outlier is removed. Secondly, if two outliers that are symmetric with respect to the expected value are removed, the clever mean remains unchanged. It would be an error to check for outliers only by looking at the value of the clever mean.

The clever variance, on the other hand, has a monotone decreasing trend as outliers are gradually removed; it is, moreover, practically unvaried or further increases when the observation removed is not a real outlier.

The authors [9] has given an example to illustrate the efficiency of CM and CV.

As a simple comparative example, let us consider the following set of measures (the data set does not contain any outlier): **y** = 31.1, 31.2, 31.2, 31.3, 31.3, 31.1, 31.4, 31.3, 31.0

The arithmetic mean is 31.2111 and the variance 0.01611.

If three outliers are added to the original data set , giving:

y=_31.1, 31.6, 31.2, 31.2, 31.3, 311.1, 31.3, 31.1, 31.4, 31.3, **32.1**, 31.0

the following results are obtained:

- y¯ = 54.642
- Median: 31.3
- cm_0 = 54.642, cm_1 = 31.327, cm_2 = 31.25, cm_3 = 31.2111
- s^2 = 6522.8
- $(mad)^2$: 0.049457
- cv_0 = 6522.8, cv_1 = 0.09218, cv_2 = 0.02944, cv_3 = 0.01611

Outliers sorted by relevance are the no. 6, 11, and 2.

Standard residuals obtained using the mean ¯y and the variance s^2 highlight only the point no. 6 as outlier. Robust residuals obtained by the median and the mad identify the points no. 6 and no. 11 as outliers, but not the point no. 2. This happens since mad is inaccurate.

Note that the clever mean and clever variance preserve their values in both the cases and coincide with the traditional variance estimation of the data set without any outliers.

V. CONCLUSION

There deferent methods can be used to detect outlier in any given data set. Mean and standard deviations are the methods which are very sensitive to outliers and can produce inefficient results. While MAD makes use of absolute median values which is not affected by the outliers. But when outliers are present in one of the extreme end of the data set, it can be biased; and it also requires sorting of data values. On the other hand, clever mean and clever variance are more robust and provide most efficient results.

REFERENCES

[1] Bland, J.M.; Altman, D.G. (1996). "Statistics notes: measurement error.". Bmj, 312(7047), 1654.

[2] http://en.wikipedia.org/wiki/Standard_deviation

[3] http://docs.oracle.com/cd/E17236_01/epm.1112/cb_statistical/frameset.htm?ch07s02s10s01.html

[4] Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: A review. International Journal of Psychological Research ,3(1),Pages 58–67

[5] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, Laurent Licata, Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median Journal of Experimental Social Psychology, Volume 49, Issue 4, July 2013, Pages 764-766

[6] Donoho, D. L., & Huber, P. J. (1983). In Bickel, Doksum, & Hodges (Eds.), The notion of breakdown point. California: Wadsworth

[7] Huber, P. J. (1981). Robust statistics. New York: John Wiley

[8] Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation.Journal of the American Statistical Association, 88(424), 1273–1283

[9] Guido Buzzi-Ferraris∗ , Flavio Manenti, Outlier detection in large data sets, Computers and Chemical Engineering 35 (2011) ,Pages 388–390