Original Research Paper

Volume - 10 | Issue - 12 | December - 2020 | PRINT ISSN No. 2249 - 555X | DOI : 10.36106/ijar

Statistics

COMPARATIVE STUDY OF STATISTICAL OUTLIER LABELING METHODS

Ritul Kamal* Department Of Statistics, University Of Lucknow, Lucknow, Uttar Pradesh (India). *Corresponding Author

Sheela Misra Department Of Statistics, University Of Lucknow, Lucknow, Uttar Pradesh (India).

ABSTRACT Outliers are observations appearing inconsistent with rest of the dataset. Outlier detection methods aim to identify the data points that are significantly dissimilar, exceptional and inconsistent with majority of the dataset. Statistical outlier detection methods are some of the oldest and foremost methodologies used. These methods are based on the fundamental assumption that the dataset follows a certain probability mode or distribution, and the points not conforming to the assumed probability distribution are outliers. The statistical outlier labeling methods are evaluated in the present study. The z-scores and Tukey methods are prone to the masking effect, which reduces their sensitivity. In univariate case, the MAD is one of the most robust methods in the presence of outliers, and therefore it is mostly recommended to use the MADe method for outlier detection. The statistical outlier labeling methods are very efficient if the underlying probability distribution of the dataset is known.

KEYWORDS : Outliers, Statistical Methods, Outlier Detection Methods, Labeling Methods

INTRODUCTION

As defined by Barnett and Lewis, outliers are observations that appear inconsistent with the rest of the data (1). Outlier detection aims to find the data points that are significantly dissimilar, exceptional and inconsistent with majority of the input dataset. The outliers may be anomalies, exceptions, faults, defects, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains. Outlier detection over the years has become an integral part in various practical applications in industries, business, and engineering etc. Outlier detection can help identify suspicious fraudulent transaction for credit card companies, identify abnormal brain signals and many other practical applications. The key component in any outlier detection technique is the input dataset which is generally a collection of data points or instances. Each data point may be described using a specified set of attributes. The data instances may be binary, categorical or continuous and each data instance may consist of only one attribute or multiple attributes known as univariate or multivariate. Another key challenge for any outlier detection technique is to identify a best set of features to give the best accurate and efficient results. The paper presents a comprehensive comparison of the statistical and data-mining based outlier detection methods.

METHDOLOGY

Statistical methods are some of the oldest and foremost methodologies that can be used for outlier detection problems (1). They are used not only to detect the outlying observations but also to study the complete dataset. Statistical outlier labeling techniques reply on the fundamental assumption that the data follow a certain distribution or probability model (1). Under the assumed probability distribution, the outliers are those data points that do not conform to the underlying probability distribution of the dataset. The statistical outlier labeling methods either assumes an underlying known probability distribution of the data points (1,3) or, that they are based on statistical estimates of unknown distribution parameters (4). The outlier labeling methods flag the observations that deviate from the underlying distribution as potential outliers. However, these methods are often unsuitable for high-dimensional data or for datasets without prior knowledge of the underlying distribution (5). If prior information about the distribution model is known then these methods are highly accurate and useful. Some of the commonly used outlier labeling methods are discussed below

i.Standard Deviation (SD) method

The standard deviation method is one of the simplest classical approaches of screening outliers in a dataset. The method is defined as follows

2SD Method: $x \pm 2SD$

3SD Method: $x \pm 3SD$

where, $\bar{x}$is the sample mean and SD is the standard deviation of the dataset.

Any observation outside these intervals is labeled as an outlier. Chebyshev's inequality states that if a random variable X with mean μ and variance σ2 exists, then for any k>0,

$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$

$P(|X-\mu| \geq k\sigma) \geq 1-\frac{1}{k^2}; k \geq 0$

the inequality is used to determine the proportion of the data lying within k standard deviations of the mean (6). Though Chebyshev's inequality holds true for all the distributions, still it is limited in its scope that it only gives the smallest proportion of observations within the k standard deviations of the mean (7). For example in case of normally distributed data, any observations that lies farther than two or three SD above or below the mean may be considered as an outlier in the dataset (8).

ii. Z-Score

The z-score is defined as a statistical measure of any observation's relationship to the mean in a group of observations. A z-score of 0 coincides with the mean. The z-score may also be negative or positive, depending on whether the observation is below or above the mean and by how many standard deviations.

The z-score method also finds application in identifying outliers in a dataset using the mean and standard deviation,

$$z_{score}(i) = \frac{x_i - \bar{x}}{s}$$

where,

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

The z-score method is based on the fact that if any random variable X follows a normal distribution, N (μ, σ2), then $Z = \frac{x-\mu}{\sigma} \sim N(0,1)$ follows a standard normal distribution and any z-score exceeding 3 in the absolute value may be considered as an outlier

The interpretation of z-score is done as follows

1. z-score < 0 denotes an observation less than the mean
2. z-score > 0 denotes an observation greater than the mean
3. z-score = 0 denotes an observation equal to the mean
4. z-score = 1 denotes an observation is 1 standard deviation greater than the mean,
5. z-score = -1 denotes an observation is 1 standard deviation less than the mean

iii. Modified Z-Score

The key limitation of the z-score method is the use of two estimator viz. sample mean ($\bar{X}$) and standard deviation (s), which are adversely affected by even a single or few extreme observations in the dataset. To

overcome this limitation of the z-score, modified z-score utilizes the median and the median of the absolute deviation (MAD) instead of sample mean and sample standard deviation respectively (9).

$$MAD = median\{|x_i - \tilde{x}|\}$$

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

where, $(\bar{X})$ is the sample median and E(MAD)=0.675σ for large normal datasets.

According to Iglewicz and Hoaglin et al. (9), any observation is labeled as an outlier when (9). The Mi score is efficient for normal data in the same way as the z-score.

iv. Tukey's Method (IQR)

The method was introduced by Tukey (10), which involves making a boxplot, i.e. a simple graphical tool to display median, lower quartile upper quartile, lower extreme and upper extreme for a continuous univariate dataset. This method is less sensitive as compared to the above methods as it utilizes the quartile values which are more resistant to extreme values as compared to sample mean and standard deviation. The method uses the interquartile range for finding the outliers in the dataset. The calculation for Tukey's method is as follows:

i. Calculate the Inter Quartile Range (IQR), i.e. distance between the lower (Q1) and upper (Q3) quartile
ii. The inner extremes are located at a distance of 1.5*IQR below Q1 and above Q3 [Q1-1.5*IQR, Q3+1.5*IQR].
iii. The outer extremes are located at a distance 3*IQR below Q1 and above Q3 [Q1-3*IQR, Q3+3*IQR].
iv. Any value which lies in between the outer and inner extremes may be a considered as a possible outlier and any value beyond the outer extreme is a probable outlier.

While the earlier methods are limited to reasonably symmetric distributions such as the normal distribution (9), the Tukey's method can also be applied to skewed and non-mound shaped distributions due to the fact that no distributional assumption are required and it does not depend on the mean or standard deviation. However, the applicability of the method may be limited in case of small sample size (9).

v. MADE method

The Median Absolute Deviation (MADE) method is one of the most basic robust methods, which is largely remains unaffected by the presence of extreme values in the data set (11). The method is similar to the SD method in its approach, however in place of mean and standard deviations, median and MADE are employed.

The method is given as follows;

2 MADE Method: Median ± 2 MADE
3 MADE Method: Median ± 3 MADE,

where, MADE = 1.483*MAD for large normally distributed data and it is a measure of the spread in the data similar to standard deviation.

MAD = median {|xi – median(x)|}; i = 1, 2… n

The MAD is multiplied by a factor of 1.483 and is similar to the standard deviation in a normal distribution. MADE is the scaled value of MAD.

RESULTS AND DISCUSSION

The data for the study has been taken from Sir Francis Galton's famous height dataset, which is based on the famous 1885 study of Francis Galton exploring the relationship between the heights of adult children and the heights of their parents. The dataset records the heights of 898 people on six variables and the dataset is available for download from https://www.randomservices.org/random/data/Galton.html. The outlier labeling methods discussed in the manuscript are applied on the dataset to identify the outliers and evaluate the relative performance of the methods in detecting outliers in the dataset.

The Z-scores found the value 56'' (below) and 78'', 79'' (above) to be outliers. The Modified Z-score found no observation in the dataset to be classified as an outlier. The 2MADe method has identified 12 values as outliers viz. 56'', 57'', 57.5'', 58'', 59'', 74'', 74.2'', 75'', 76'', 76.5'', 78'', 79''. Also, the 3MADe method has identified 2 outliers viz 78'', 79''. The Tukey (boxplot) method has identified 1 outlier viz. 79'' (Figure 1).

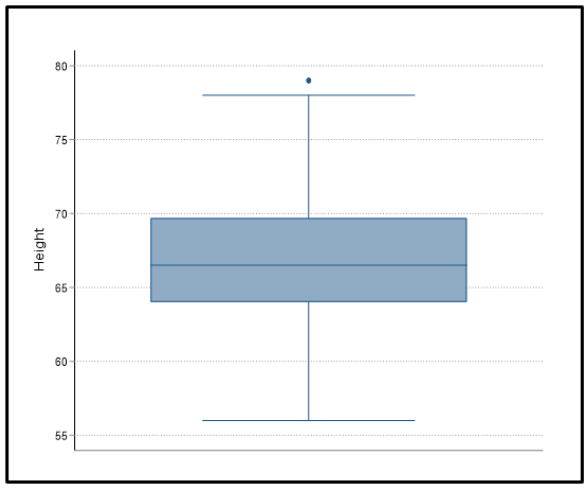

Figure 1: Tukey Boxplot for labeling outliers.

Table 1: Number of outliers detected by various methods

Method	Cut off value	Outliers detected
Z-scores	Zi>3	56", 78", 79"
Modified Z-Score	|Mi|>3.5	None
MAD	MAD>2	56",57", 57.5", 58", 59", 74", 74.2", 75", 76", 76.5", 78", 79"
	MAD>3	78", 79"
Tukey's Method	[55.45 – 78.25]	79"

Table 1 presents the number of outliers detected by various outlier labeling methods in the dataset. The outlier labeling methods have been compared using sample dataset to evaluate the performance of the methods for detecting outliers. Z-scores and Tukey methods are prone to the masking effect, which reduces their sensitivity. One of the most popular methods of outlier labeling in case of one-dimensional data is the use of Median Absolute Deviation (MADe), i.e. to identify any data point that is more than two standard deviations away. MADe and Modified Z-score are used in the MAD approach. The Z-score method has labeled 3 observations as outliers in the dataset however; modified Z-score method did not label any observation as outlier. In case of the MADe method, the MAD>2 has labeled 12 observations as potential outliers whereas the MAD>3 has labeled 2 observations as potential outliers in the dataset. In univariate case, the Median Absolute Deviation is one of the most robust methods in the presence of outliers, and therefore it is mostly recommended to use the MADe method for outlier detection.

CONCLUSIONS

Statistical models are generally suited to quantitative real-valued data sets or at the very least quantitative ordinal data distributions where the ordinal data can be transformed to suitable numerical values for statistical processing. This limits their applicability and increases the processing time if complex data transformations are necessary before processing. The statistical outlier detection methods have several advantages viz. they are mathematical justified, they are very efficient if the underlying probability distribution is known and it is possible to assess the meaning of the outlier found. However, the statistical outlier detection methods suffer from many disadvantages viz. they are often not applied to the multi-dimensional scenario as most distribution models apply typically to the univariate feature space and these methods do not perform well even in case of moderate multi-dimensional datasets.

REFERENCES

1. Barnett, V. and Lewis T. Outliers in Statistical Data. 3rd Editio. John Wiley & Sons; 1994.
2. Eskin E. Anomaly detection over noisy data using learned probability distributions. Proc Seventeenth Int Conf Mach Learn. 2000;255–62.
3. Wainer H, Rousseeuw PJ, Leroy AM. Robust Regression & Outlier Detection. J Educ Stat. 1988;13(4):358.
4. Hadi A. S. Identifying multiple outliers in multivariate data. J R Stat Soc. 1992;54:761–71.
5. Papadimitriou S, Kitawaga H, Gibbons P, Faloutsos C. LOCI: Fast Outlier Detection Using the Local Correlation Integral. 2002.
6. Bain, L., Engelhardt M. Introduction to probability and mathematical statistics. 2nd ed. Duxbury; 1992.
7. Lethen J. Chebychev and empirical rules. [Internet]. 1996. Available from: http://stat.tamu.edu/stat30x/notes/node33.html

8. Seo S. A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets. University of Pittsburgh; 2002.
9. Iglewicz, B., & Hoaglin DC. How to detect and handle outliers. Milwaukee, WI.: ASQC Quality Press; 1993.
10. Tukey J. Exploratory data analysis. Addison-Wesely; 1977.
11. Burke S. Missing Values, Outliers, Robust Statistics & Non-parametric Methods. GC Eur Online Suppl. 1998;19:22–7.