JOURNAL OF TECHNIQUES

Journal homepage: http://journal.mtu.edu.iq

RESEARCH ARTICLE - MANAGEMENT

Fast Ways to Detect Outliers

Emad Obaid Merza[1*] and Nashaat Jasim AL-Anber[1]

[1] Information Technology Department, Technical College of Management-Baghdad Middle Technical University, Baghdad, Iraq.

* Corresponding author E-mail: edmistry1974@gmail.com

Article Info.

Article history:

Received
16 January 2021

Accepted
26 February 2021

Publishing
31 March 2021

Abstract

The occurrence of tremendous developments in the field of data has led to the formation of huge volumes of data, and it is normal that this leads to the presence of outliers in this data for many reasons, which may have small or large values compared to the rest of the normal data, and the presence of outliers in the data affects the statistical analysis of this data, so we must try to reduce its impact in various ways. On the other hand, the presence of outliers may be of great benefit, for example knowledge of geological activities that precede natural disasters such as (earthquakes, forest fires, floods ... etc.). Therefore, detection of outliers is of great importance in various fields. In this research, we aim to develop easy methods for detecting outliers in big data, as the problem that this research addresses is that many of the newly developed methods for detecting outliers suffer from computational complexity or are efficient when the sample size is small. An experimental approach was used in this research by suggesting three methods for detecting outliers, the first method is based on standard deviation and was tested and compared with the normal distribution method and the z-score method. The second method depends on the maximum and minimum value of the data, and the third method depends on the range between successive data points. The results of second and third methods are compared with Hample's Test method result. The accuracy of the results is measured based on the confusion matrix. The results of the proposed methods test showed the conformity of the first method with the results of the normal distribution method and the Z-Score method, as well as the superiority of the third method over the Hample's test method. In this paper, it was concluded that the Hample's test method suffers from a serious weakness when the zero values in the data constitute more than 50% of the number of elements.

Keywords: outlier; outlier detection; big data; normal distribution; Z-Score; Hample's test.

1. Introduction

The importance of analyzing outliers is increasing with the acceleration of development and broad jumps in information technology, as data volumes have become more inflated and complex, which requires converting these data into useful information for the decision-making process and data analysis, and this transformation process includes a very important matter, which is the concept of outliers[1], And that the issue of outliers has been taken up by many scientists and researchers in order to study the effect of these values on the accuracy of the results expected from the data analysis process[2], and among those prominent scientists who dealt with the concept of outliers is Hawkins and Freeman. The outlier in a particular data set that may appear in the form of one or more values. What distinguishes this value is that it is not logical in relation to the rest of the natural data, for example, it may be very large or very small compared to the mean of the data and that the existence of a unique value is of high importance. Because it has important implications in data mining as well as in analyzing medical and financial data and in the field of networks, as detection of intrusion on networks is one of the most applied topics that have gained importance in recent years [3]. The exploration of outliers or unique patterns is a very important sub-topic in data mining, as the process of detecting outliers is the exploration of unique patterns that clearly deviate from the natural path of the data [4], and the interest in classifying data and knowing their behavior, common characteristics and differences between them. It will inevitably lead to ease of study [1], and the concept of outliers is always associated with the study of the natural behavior of the data, as any behavior that does not resemble normal behavior is considered an erratic behavior. For example, unnatural geological activities that precede natural disasters. As the process of detecting outliers is applied in many sectors, for example in the health sector, as it is used to diagnose diseases, the financial sector is used to detect fraud in credit cards, and in networks, the abnormal activity in the network may mean the presence of an infiltration, attack or entry process. Unauthorized and outliers is used in image processing as well as in the industrial sector to detect industrial defects in products. On the statistical side, the idea of an outlier can be discussed on the basis that it does not share characteristics with the community or the sample. The community is defined as a group of beings who have common characteristics and the sample is part of the community and bears all of its characteristics [2]. Sometimes the

Terminology	Symbol
Local Correlation Integral	LOCI
Driving Behavior-based Trajectory Outlier Detection	DB-TOD
True Negatives	TN
True Positives	TP
Receiver Operating Characteristics	ROC
Area Under Curve	AUC
Average of a Point-to-Standard Deviation Method	AP-SDM
Average Range between Consecutive Points Method	AR-CPM
Half-Value Method	HVM

existence of outliers may not be meaningful. The first important process in statistical analysis of data is to identify and detect outliers because their presence may be misleading [5]. There are a large number of techniques for detecting outliers that can be summarized by statistical methods. Density-Based Methods, Cluster-Based Methods, Distance-Based Methods, Subspace-Based Methods and Deviation-Based Methods. Multiple outliers are detected one by one in the regression analysis model, and this may lead to misleading results due to the smearing and masking effect, and finding techniques to detect outliers at once will avoid these effects [6]. Detecting outliers is important because it affects the estimated model of regression, especially when using the least squares method. Therefore, detection of outliers is very important in practical applications [7]. There are many previous studies in which methods for detecting outliers have been proposed, some of which are mentioned: In 2002, researchers Angiulli and Pizzuti proposed a new definition of the concept of distance based on extremes by taking the sum of the distances between each of the data points with their closest points under the name (weight). Whenever the weight of a data point is large, the point is considered an outlier and the weight is calculated through Linear search of the search space using a Hilbert space filling curve [8]. In 2003, researchers Papadimitriou et al presented a new method for detecting outliers, which is Local Correlation Integral (LOCI), which automatically detects the boundaries of the outliers data without the need for a predefined threshold limit and one of the benefits of this method is that it identifies the clusters and their diameters precisely as well as calculates the distances between clusters, and in that research this method was used By determining the approximate limits of outliers under the term approximate LOCI[4]. In 2017, researchers Wu et al. introduced a new approach to detect the outliers represented by the deviating lanes of cars from their expected path, through early warning of the presence of a strange path for the car before it reaches the target without the need for complete historical data on the tracks, which is considered an expensive procedure, this new approach was called the Driving Behavior-based Trajectory Outlier Detection (DB-TOD), which is based on the statistical probability model to calculate target distractions[9]. In this paper, we have contributed in providing fast and competitive methods for detecting outliers in terms of accuracy compared to some known methods. The results of this comparison will be detailed in Section 4.

2. Outlier

The outlier, which is also known as the extreme or anomalous value, according to Hawkins's definition. This value is defined as the value that clearly differs from other values, which raises the suspicion that it was generated in a way that differs from the rest of the normal values, and it is simply a value that differs from the rest of the other values and statistically, normal values are obtained by means of a specific generation mechanism. Therefore, outliers are those that deviate significantly from this generation mechanism [1][3], that is, outliers originate from different generation sources and in many applications the data generation process It is done using one or more mechanisms and when this mechanism works abnormally it leads to the emergence of outliers [10]. So, the value is called an extreme if it deviates from the normal behavior of the data, is far from its mean, or is not similar to any other object in terms of the predominant properties of the data [11], as shown in Fig. 1.

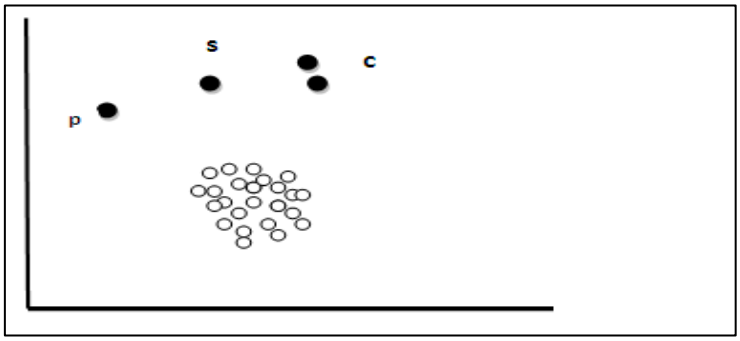

Fig. 1. shows the outliers represented by points p, s and group c

There are several types of outliers, as the data set can contain multiple types of outliers [12] and they are global outliers Figure 2, Context outliers, Collective outliers Figure 3.

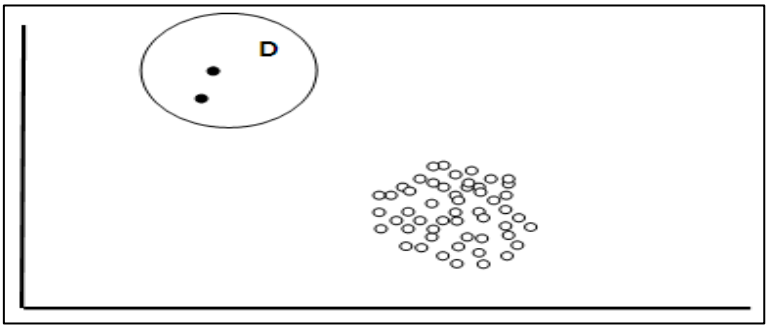

Fig. 2. The values in Zone D represent general outliers

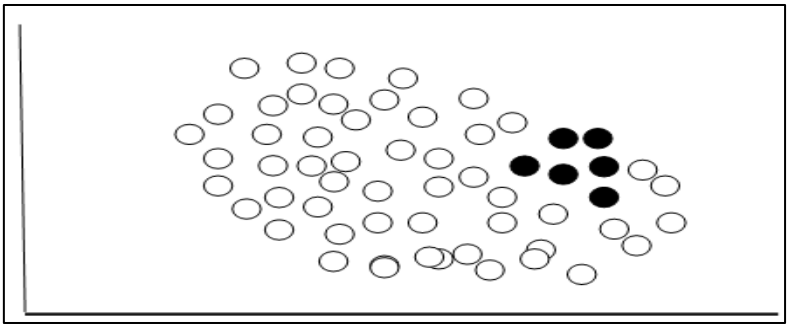

Fig. 3. The dots in bold color represent collective outliers

2.1. Causes of outliers

There are many reasons for outliers in data, and their presence is not limited to data entry errors, as in the following points [10] [12] [2].

1. The outliers are due to distributions that are different from the distribution that produced the normal values in the data set.
2. Measurement error, for example, some work and tests require collecting samples of its weight with a group of sensitive scales, and one of these scales is idle and gives a reading of its weight higher than usual.
3. The existence of a mixture of different graphic types.
4. There is an error in processing the data.

2.2. Outliers detection methods

There are several common methods for detecting outliers which are use of the item labels, statistics-based methods, proximity-based methods, and cluster-based methods [1] [12].

2.3 Measuring accuracy of detection algorithms for outliers

If we consider that the outlier's detection process is a classification process, then there are several methods to measure the accuracy of the outlier's detection algorithms, for example, the so-called Confusion Matrix is used, which is an array that contains information about the actual and expected classifications of the data, and Figure 4 shows the confusion matrix [11].

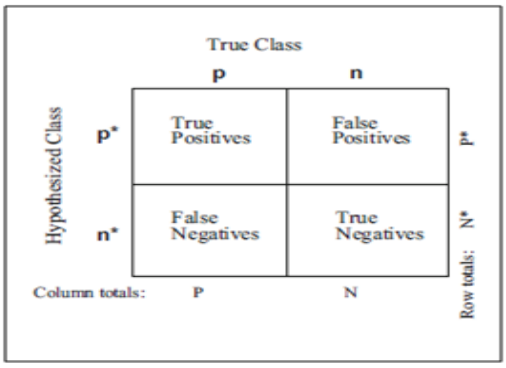

Fig. 4. Confusion Matrix

True Negatives represent the correctly predicted normal values (TN).
False Positives represent a false prediction of a number of normal values as outliers (FP).
False Negatives represents a false prediction of a number of outliers as normal values(FN).
True Positives represent correctly predicted outliers (TP).

The confusion matrix [11] is used to form a set of performance measures (accuracy), which are considered among the basic measures in evaluating the classifier's performance, as follows:

1. Accuracy measurement.

$$Accuracy = \frac{TP+TN}{P+N} \quad (1)$$

Where:

TP correctly predicted outliers

TN correctly predicted normal values

$P + N$ The total number of observations

2. Receiver Operating Characteristics (ROC) Here the FP values are represented on the **X** axis and the TP values on the **Y** axis.

3. Area Under Curve (AUC), It helps to analyze the overall performance of the classifier and the ideal classifier has AUC = 1.

3. Proposed Methods

3.1. Average of a Point-to-Standard Deviation Method (AP-SDM)

This method is characterized by its simplicity, ease of implementation and accuracy of its results as by comparing it with the use of a normal distribution within $\mu \pm 3\sigma$), it achieves a very high match, and the method aims to discover outliers regardless of the size of the data, and below we list the steps of the Average of a Point-to-Standard Deviation Method (AP-SDM):

1. Find the standard deviation (SD) of the data.
2. Dividing each data point by the standard deviation, we get a P for each point.
3. Find the arithmetic mean (AP) of paragraph product B.
4. We apply the equation as follows:

$$V = Ap \pm 3 \quad (2)$$

Any point for which the value of V is greater than Ap + 3 is considered to be a maximum outlier and any point for which the value of V is smaller than Ap-3 is considered to be a minimum outlier.

3.2. Average Range between Consecutive Points Method (AR-CPM)

This method is characterized by its ease and ability to be applied to various data volumes, and it is also a fast and free-of-complication method. The following are the steps of the Average Range between Consecutive Points Method (AR-CPM):

1. Find the range between successive data points and for each point according to the following equation:

$$R_i = p_{i+1} - p_i \quad (3)$$

Where:
R_i : Represents the range between any two consecutive points
p_i : Represents the data point
p_{i+1} : Represents the data point that follows the point of p_i

The first point, P_1, is neglected because there is no point preceding it.

2. Find the mean of the calculated AR ranges that found in step 1.
3. Any value greater than (AR × 6) is a higher outlier.
4. Any value smaller than (AR / -6) is considered a minimum outlier.

3.3. Half-Value Method (HVM)

It is one of the easiest methods as it simply depends on half the value (maximum or minimum) with the arithmetic mean and standard deviation of the data, and below we list the steps of the Half-Value Method (HVM):

1. Find the maximum value (max) and the minimum value (min) for the data.
2. Find the mean and standard deviation of the data.

3. To find the higher outlier, any data point greater than V_1 can be considered a higher extreme point since V_1 is calculated according to the equation as follows:

$$V_1 = \left(\frac{Max_p}{2}\right) + Max(\mu, \sigma) \quad (4)$$

Where:
Max_p : Represents the maximum value for the data.
$Max(\mu, \sigma)$**:** Is the largest value between the mean and the standard deviation of the data.

4. To find the minimum outlier, any data point smaller than V_2 can be considered a minimum extreme point since V_2 is calculated according to the equation as follows:

$$V_2 = \left(\frac{Min_p}{2}\right) + Min(\mu, \sigma) \quad (5)$$

Where:
Min_p : Represents the minimum value of the data.
$Min(\mu, \sigma)$**:** Represents the smallest value between the mean and the standard deviation of the data.

4. Application

The researcher used data distributed normally within ($\mu + 3\sigma$) and added to it fake outliers in order to test and compare the proposed methods. The proposed methods will be applied and their results compared with some common methods, as the (AP-SDM) method will be tested with the normal distribution method and the Z-SCORE method using different sizes of data, and then the methods will be applied to the data of the Abu Gharib factory for the year 2019, either of the two methods (AR-CPM) and (HVM) will be compared with Hample's Test method by adding fake outliers depending on the data upper limit of the normal distribution method to test the accuracy of the three methods using the Abu Gharib factory data for the year 2019 directly. As the methods (normal distribution method, the Z-SCORE and Hample's Test method) are calculated their normal limits for the data is as follows:

normal distribution method= $\mu \pm 3\sigma$ (6)
Where:
μ is the mean of data, σ is the standard deviation (SD)

Z-SCORE= $\frac{x-x}{s}$ (7)
Where:
x represents the data element, x is the mean of data, s is the standard deviation (SD)

4.1. Comparison of (AP-SDM) with normal distribution method and Z-SCORE method

This method aims to detect outliers, regardless of the sample size, with very high accuracy with the normal distribution method and the Z-SCORE method within the period (-3, +3), and the researcher tested the three methods on different data sizes (12, 50, 100, 200, 500, 1000) and the test results were as shown in table 1.

Table 1. represents the results of testing different sizes of data to detect outliers using the three methods (z-score, Normal distribution, AP-SDM)

Number of detected outliers			
AP-SDM	Normal distribution	Z-SCORE	Sample size
0	0	0	12
0	0	0	50
0	0	0	100
3	3	3	200
1	1	1	500
6	6	6	1000

Below we list the results of applying the three methods on producing a day / year (229 days representing the sample size) for all production lines for the year 2019 of the Abu Gharib production plant as shown in table 2, as well as on producing a month / year (12 months representing the sample size) For all production lines as shown in table 3.

Table 2. The number of detected outliers for three methods (z-score, normal distribution, AP-SDM) using day / year production (229 days representing the sample size) for all production lines for the year 2019 of the Abu Gharib factory

product name	Number of detected outliers		
	AP-SDM	Normal distribution	Z-SCORE
Cream	5	5	5
Cheddar cheese	1	1	1
Yoghurt	4	4	4
Laban Shanina	9	9	9
Butter	8	8	8
Free fat	10	10	10
Soft cheese	11	11	11

Table 3. The number of detected outliers for three methods (z-score, normal distribution, AP-SDM) using the production of a month / year (12 months represents the sample size) for all production lines for the year 2019 of the Abu Gharib factory

Product name	Number of detected outliers		
	AP-SDM	Normal distribution	Z-SCORE
Cream	0	0	0
Cheddar cheese	0	0	0
Yoghurt	0	0	0
Laban Shanina	0	0	0
Butter	0	0	0
Free fat	0	0	0
Soft cheese	0	0	0

4.2. Comparison of (AR-CPM) and (HVM) with (Hample's Test)

The methods (AR-CPM) and (HVM) aim to make the process of detecting outliers easy, fast, and far from complex, and the possibility of using them regardless of the size of the sample is large or small, and the researcher has added an outlier to each sample of production samples for months (March, April And August) Which belongs to the cream product, depending on the upper limit of the normal data for the normal distribution method to test the two methods and compare them with the results of the Hample's Test in terms of the possibility of discovering the outliers that were added and measuring the accuracy of the detection process and the reason for choosing the mentioned months is that the methods (normal distribution, Hample's test and AR-CPM and HVM) did not record any outliers detection. To be clear, we include below steps of the Hample's Test method:

1. Arrange the data in ascending order and calculate the median.
2. Extract the absolute value resulting from subtracting each data value from the median to produce a ri column.
3. Calculate Me | ri | Median of the absolute values of the ri column.
4. Any value in the column | ri | greater than or equal to 4.5 * Me | ri | is an outlier. Table 4 shows the results of testing the application of methods over the months (March, April, August) Which belongs to the cream product.

Table 4. Test to detect fake outliers using (AR-CPM) and (HVM) with (Hample's Test)

month	Sample size	Addition Size	Number of detected outliers		
			Hample's test	AR-CPM	HVM
March	10	1	0	0	1
April	8	1	0	0	1
August	19	1	0	0	1
sum	37	3	0	0	3

In order to measure the accuracy of both methods in according to the test in table 4, we use equation 1 as follows:

$ACCURACY_{Hample} = \frac{0+37}{43} = 86\%$

$ACCURACY_{AR\text{-}CPM} = \frac{0+37}{43} = 86\%$

$ACCURACY_{Hvm} = \frac{43}{43} = 100\%$

Below we list the results of applying the methods to production for a day / year (229 days representing the sample size) for all production lines for the year 2019 of the Abu Gharib factory as shown in table 5, as well as on the production month / year (12 months representing the sample size) for all Production lines as shown in table 6.

Table 5. The total number of detected outliers for the methods (Hample's Test, AR-CPM, HVM, normal distribution) using the day / year production (229 days representing the sample size) for all production lines for the year 2019 of the Abu Gharib factory

	Detecting outliers		
Product name	HVM	AR-CPM	Hample's Test
Cream	5	14	20
Cheddar cheese	1	1	1
Yoghurt	9	10	12
Laban Shanina	10	0	229
Butter	2	12	229
Free fat	8	8	229
Soft cheese	5	12	229
Sum	40	57	949

Table 6. The total number of detected outliers for the methods (Hample's Test, AR-CPM, HVM, normal distribution) using the month / year production (12 months representing the sample size) for all production lines for the year 2019 of the Abu Gharib factory

	Detecting outliers		
Product name	HVM	AR-CPM	Hample's Test
Cream	1	1	1
Cheddar cheese	1	0	1
Yoghurt	1	1	1
Laban Shanina	1	0	1
Butter	1	0	4
Free fat	1	0	2
Soft cheese	2	0	12
Sum	8	2	22

5. Conclusions

The AP-SDM method achieved completely identical results with the normal distribution method and the z-score method when applying it even when testing the method of normal distribution within ($\mu \pm \sigma$) and ($\mu \pm 2\sigma$) with the (AP-SDM) within ($AP \pm 1$) and (($AP \pm 2$) respectively, an exact match was obtained when the sample size was less than 200, and very close results for the sample size greater than 200. on different sizes of data (12, 50, 100, 200, 500, 1000), as well as when applying the methods to the data of the Abu Gharib factory. For all production lines day / year at sample size 229 and all production lines month / year at sample size 12, the results of the AP-SDM method were completely identical to the normal distribution method and the z-score method.

When comparing the results of the test of the two proposed methods (HVM and AR-CPM) with the results of the Hample's test method, HVM was more accurate than the AR-CPM method and the Hample's Test method. There is a weakness in the (Hample's Test method) as if the median of the data is equal to zero, then all the data values will be considered as outliers according to the condition of the method ($| ri | \geq 4.5 * Me | ri |$). Based on the conclusions, future work revolves around several main points, as follows:

1. Adoption and development of the proposed methods to be appropriate for detecting outliers in multivariate data.
2. Hybridizing the proposed methods with cluster algorithms to produce new methods for detecting local outliers.
3. Work to make the outliers detection process easier to implement, which will be positively reflected in the statistical analysis of the data.

References

[1] J. Han, M. Kamber, J. Pei, Data mining: concepts and techniques, Third Edition, 225Wyman Street,Waltham, MA 02451, USA: Elsevier, 2011.

[2] P. J. Rousseeuw and A. M. Leroy, Robust regression and outlier detection, Canada: John wiley & sons, 2005.

[3] H. P. Kriegel, P. Kröger, A. Zimek, "Outlier detection techniques," in the 2010 SIAM International Conference on Data Mining, Columbus, Ohio, 2010.

[4] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, C., Faloutsos, " Loci: Fast outlier detection using the local correlation integral," In Proceedings 19th international conference on data engineering, Bangalore, India, pp. 315-326, 2003.

[5] M. Breaban and H. Luchian, " Outlier detection with nonlinear projection pursuit," International Journal of Computers Communications & Control, 8(1):30-36, ISSN 1841-9836, February, 2013.

[6] S. Akter and M. H. Khan, " Multiple-Case Outlier Detection in Multiple Linear Regression Model Using Quantum-Inspired Evolutionary Algorithm," JOURNAL OF COMPUTERS, VOL. 5, NO. 12, DECEMBER, 2010.

[7] O. G. Alma, 'Performances Comparison of Information Criteria for Outlier Detection in Multiple Regression Models Having Multicollinearity Problems using Genetic Algorithms," Matematika, Volume 29, Number 2, pp 119–131, 2013.

[8] F. Angiulli and C. Pizzuti, " Fast outlier detection in high dimensional spaces," In European conference on principles of data mining and knowledge discovery, Springer, Berlin, Heidelberg, pp. 15-27, 2002.

[9] H. Wu, W. Sun, B. Zheng, "A fast trajectory outlier detection approach via driving behavior modeling," In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, pp. 837-846, 2017.

[10] C. C. Aggarwal. (2016, November 25). Outlier analysis (second edition) [online]. Available: http://rd.springer.com/book/10.1007/978-3-319-47578-3

[11] N. SURI, N. M, G. Athithan, Outlier detection: techniques and applications. Switzerland Springer Nature, 2019.

[12] J. Astal, "Comparison of Methods for Detecting Outliers in Medical Data", M.Sc. thesis, Al-Azhar University–Gaza, 2018.