Outlier Detection: How to Threshold Outlier Scores?

Jiawei Yang
School of Computing
University of Eastern Finland
Joensuu, Finland
jiaweiy@uef.fi

Susanto Rahardja
School of Marine Science and Technology
Northwestern Polytechnical University
Xi'an, Shaanxi, China
susantorahardja@ieee.org

Pasi Fränti*
School of Computing
University of Eastern Finland
Joensuu, Finland
franti@cs.uef.fi

ABSTRACT

Outlier detection is a fundamental issue in data mining and machine learning. Most methods calculate outlier score for each object and then threshold the scores to detect outliers. Most widely used thresholding techniques are based on statistics like standard deviation around mean, median absolute deviation and interquartile range. Unfortunately, these statistics can be significantly biased because of the presence of outliers when calculating these statistics. This makes their use inaccurate. To overcome this problem, we propose a two-stage thresholding method (2T). Most obvious outliers are first removed by using a more conservative threshold, and the same process is then repeated for the processed scores. Experiments show that this two-stage approach significantly improves the results of all the three existing thresholding techniques.

CCS CONCEPTS

• Information systems • Information systems applications • Data mining

KEYWORDS

Outlier detection, Error detection, Standard deviation, Median absolute deviation, Interquartile range, Novelty detection, MAD, SD, IQR, Anomaly detection, Two-stage threshold

1 Introduction

Outliers are objects that deviate from typical data. They can represent important information, which is critical for fraud detection, public health, and network intrusion [1], and they can affect statistical conclusions based on significance tests [2]. Outliers can also be noise that harms a data analysis process. In both cases, it is desirable to detect the outliers.

AIIPCC '19, December 19–21, 2019, Sanya, China

ACM ISBN 978-1-4503-7633-4/19/12…$15.00
https://doi.org/10.1145/3371425.3371427

Most methods calculate so-called outlier score for every object and then the objects whose scores deviate too much from the norm are marked as outliers. However, it was shown in [3] that using simple statistics can cause false positives. Similarly, it was stated in [4] that researchers misuse statistical tools like mean ± two to five times *standard deviation* (SD) as a threshold because the SD value is also affected (increased) by the outliers. Consequently, the threshold tends to be set up too high and several outliers missed. A more robust threshold called *median absolute deviation* (MAD) has been proposed in [4] to reduce the effect of the outliers.

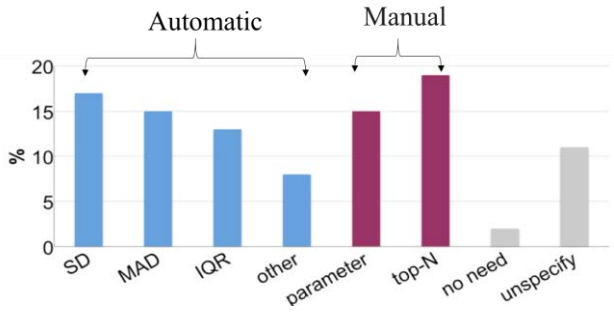

Figure 1: Frequency of Usage of Different Threshold Selection Techniques in Outlier Detection Literature between June 2016 and June 2018.

We searched from Google Scholar using keyword "*outlier detection*" and it returned 38,900 related publications during the last 2 years (June 2016 to June 2018). We randomly picked 100 publications and studied what thresholding techniques were used. The results are summarized in Figure 1. They show that standard deviation, median absolute deviation and *interquartile range* (IQR) are the most used techniques. The rest used either user-given parameter (*parameter, top-N*), did not need thresholding, or did not specify their choice of threshold technique. By *parameter*, we refer to a user-given constant for thresholding the outlier scores. *Top-N* refers to a priori knowledge of how many outliers (%) are expected to be in the data.

However, none of these techniques works satisfactorily when there are outliers very far from the true data objects. These far away outliers cause all these three statistics to have too high

values, see Figure 2. As a result, all the existing statistics over-estimate the threshold and fail to detect most of the outliers with smaller value. The main problem is that the statistics used to set the threshold are unreliable due to the outliers.

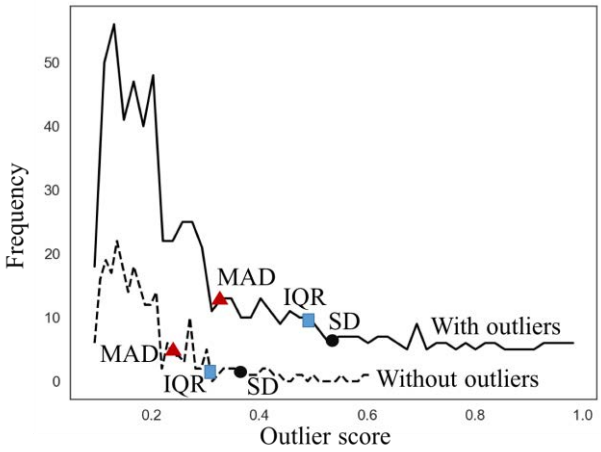

Figure 2: Effect of the Outliers for the Thresholding Techniques.

To overcome this problem, we propose a simple but significantly more robust approach to process the data iteratively. We first apply any standard technique to remove the most obvious outliers, and then repeat the same process for the processed outlier scores. As a result, the statistics calculated after the first round suffers much less from the outliers, and therefore, allows detection that is more accurate in the second round.

2 Existing Work and Their Limitations

Most used techniques in the literature are SD [5], MAD [4] and IQR [6]. Other techniques have also been introduced based on the specific outlier detection algorithm like E-value [7], percent identity [7], or bit-score [7]. In this paper, we show how all these three techniques can be improved by the proposed two-stage (2T) approach: SD, MAD and IQR.

2.1 Standard Deviation (SD)

Mean ± three times SD was first introduced in 1962 [8]. This threshold technique assumes that the data follows normal distribution and implies outlier level of 0.13%. Other authors [4, 9] suggested more aggressive choices by using mean ± 2 or 2.5 times SD, which correspond to the outlier levels of 0.62 % and 2.28%.

In general, this technique calculates the threshold as:

$$\text{Tmin} = \text{mean} - a * \text{SD}; \text{Tmax} = \text{mean} + a * \text{SD} \quad (1)$$

where mean and SD are the corresponding statistics of the outlier scores; and a is a control parameter decided by the user. The smaller the value, the more objects will be marked as outliers. Most common choice is a=3 because the amount of outliers is expected to be small [4].

There are three problems with the SD threshold technique [4, 10]. First, data, including outliers, should follow normal distribution. Second, outliers seriously affect the mean and SD. Third, there is only small percentage of outliers [1, 8], but in some cases, the amount of outliers can be even more than 50%.

2.2 Median Absolute Deviation (MAD)

Absolute deviation around the median was adopted in [10], which was motivated by the fact that median is more robust against outliers than mean [4]. Median is the object value ranked in the middle (or the average of the two central objects in case of even size dataset).

In general, MAD thresholding strategy calculates the threshold as follows:

$$\begin{cases} \text{Tmin} = \text{median}(X) - a * \text{MAD} \\ \text{Tmax} = \text{median}(X) + a * \text{MAD} \end{cases} \quad (2)$$

$$\text{MAD} = b * \text{median}(|X - \text{median}(X)|) \quad (3)$$

where median and MAD are the corresponding statistics of the outlier scores; and a is decided by the user and usually set to 3. The value b is suggested to be 1.4826 in [4]; and X is the n original observations [4]. Unfortunately, MAD is also affected by outliers especially when the amount of outliers exceeds 50%, which causes median to be located within the outliers.

2.3 Interquartile Range (IQR)

The interquartile range is the difference between the values ranking in 25% and 75% in a data set. The values are denoted as Q1 and Q3, respectively.

In general, IQR thresholding strategy calculates the threshold as follows:

$$\text{Tmin} = \text{Q1} - c * \text{IQR}; \text{Tmax} = \text{Q3} + c * \text{IQR} \quad (4)$$

$$\text{IQR} = \text{Q3} - \text{Q1} \quad (5)$$

where c is decided by the user and usually set to 1.5 [4].

Unfortunately, IQR is also affected by outliers when Q3 is located within the outliers.

2.4 Clever Standard Deviation (Clever SD)

An improvement of the SD, called *clever standard deviation* was proposed in [11]. The idea is to remove one outlier at a time and then re-calculate SD for the remaining of the scores. The process is repeated until no more outliers fall into the range mean ± 2.5*SD.

The method improves SD in case of extreme outliers. However, it tends to iterate too greedily and keep removing also normal points.

3 Two-stage Thresholding (2T)

As all the values of SD, MAD and IQR can be severely affected by the presence of outliers, these indicators are fundamentally problematic. Hence, we propose *two-stage thresholding* (2T) where the most outlying scores are excluded from the process to reduce their effect in the calculations of the threshold. In the second stage, any standard thresholding can be applied using the

remaining scores. The purpose of the first stage (cleaning step) is to make the second stage (threshold calculation) more robust in the presence of very noisy data points.

The proposed technique can be iterated multiple time if wanted. However, we found out that the simpler two-stage approach is more robust and faster in practice. Clever SD is a special case of the 2T, where the process is iterated until convergence, and only one outlier is removed at a time.

Our technique has three main differences to the above-mentioned Clever SD. First, 2T is not limited to use SD but other existing techniques such as MAD and IQR can also be applied. Second, instead of eliminating only the most extreme outlier, we remove all outliers that exceeds the preliminary threshold. Third, Clever SD iterates until to convergence whereas 2T stops after few iterations, which makes it significantly faster. The pseudo code of the technique is shown in Algorithm 1.

Algorithm 1: two-stage thresholding (2T)

Input: Outlier scores: $X \in \mathcal{R}^{1\times n}$

Output: Threshold: $T \in \mathcal{R}$

```
REPEAT
  RemoveMostOutlyingScores from X
  Calculate T over X via (1), (2) or (4)
UNTIL StopCondition
RETURN T
```

There are two ways to decide which outliers are removed. First way is to remove the biggest outlier scores manually using domain specific knowledge by the researcher, and then calculate the threshold via (1), (2) or (4). However, this can be time-consuming especially in case of big data.

Second way is to calculate threshold by (1), (2) or (4), and then remove the objects whose scores exceed the threshold. Then re-calculate the threshold value via (1), (2) or (4) again over the remaining scores. We note that the same parameter setup (a, b or c) can be used in both stages. The proposed 2T therefore does not involve any additional parameters or design choices; except to decide whether to do it only once, or iterate multiple times. Our recommendation is to do it only once.

However, if outliers are removed correctly at the first step, 2T can fail by inadvertently removing inners. Therefore, 2T can improve only when data has lots of outliers.

Similar two-stage strategies to handle outliers have also been considered in [12], [13] and [14]. In [12], mean trajectory is estimated for a set of GPS trajectories. Outlier trajectories that are far away from the tentative mean are then removed. SD strategy with fixed threshold is applied. In the second stage, new mean is calculated from the cleaned data. In [13], k-means clustering is improved by applying the algorithm twice. First k-means is applied to all data. Objects that are far away from the centroids are removed, and then k-means is run again to the remaining datasets. In [14], a two-step outlier detection method is proposed. In the first step, dataset is divided into subset based on clustering results and outliers are then detected from each subset separately.

4 Experiments and Results

To study the performance of 2T, we tested 2T with both manually generated outlier scores, and outlier scores produced by an algorithm for real-world datasets. Following [4], we first tested the manually generated outlier scores, listed in Table 1, which have varying number of outliers, and outlier scores. Since only positive values were considered, we used the Tmax as the threshold for all tests. We evaluated the results with F1 score, which is the average of precision and recall.

Table 1: Outlier Score Cases.

Case	Inners	Outliers
X0	{1, 2, 3, 6, 8}	{}
X1	{1, 2, 3, 6, 8}	{1000}
X2	{1, 2, 3, 6, 8}	{500, 1000}
X3	{1, 2, 3, 6, 8, 16, 17, 18, 18}	{60,1000}
X4	{1, 2, 3, 6, 8, 16, 17, 18, 18}	{60, 300, 500, 1000, 1500}

4.1 Comparison of SD, MAD, and IQR

Dataset X0 consists of 5 normal objects (inners) and no outliers. The objects have the following outlier scores 1, 2, 3, 6, 8, and the following statistics: mean=4.00, SD=2.61, median=3, MAD=2, IQR=4.00, Q1=2 and Q3=4. From these values, the threshold values were derived and shown in Table 2. As can be seen, all methods detect no outliers which is expected.

Table 2: Results of SD, MAD and IQR with x0 = {1, 2, 3, 6, 8}.

Method	Threshold	Outliers
SD	11.82	{}
MAD	11.90	{}
IQR	12.00	{}

We then added one outlier to the dataset with outlier score of 1000. The scores of this modified dataset are X1= {1, 2, 3, 6, 8, 1000}, and the obtained threshold values were summarized in Table 3, which shows SD fails to detect the outlier while MAD and IQR succeed.

Table 3: Results of SD, MAD and IQR with x1 = {1, 2, 3, 6, 8, 1000}.

Method	Threshold	Outliers
SD	1283.58	{}
MAD	17.84	{1000}
IQR	15.38	{1000}

The next set contains two outliers with the scores X2= {1, 2, 3, 6, 8, 500, 1000}. The resulting threshold values and the detected outliers are shown in Table 4. We can see that SD still fails,

whereas IQR can detect one of the outliers. MAD successfully detects both outliers.

Table 4: Results of SD, MAD and IQR with x2 = {1, 2, 3, 6, 8, 500, 1000}.

Method	Threshold	Outliers
SD	1304.77	{}
MAD	23.79	{500, 1000}
IQR	631.25	{1000}

Next we replaced one outlier score (500) by a smaller magnitude (50), and added several real data objects (inners). The scores are X3= {1, 2, 3, 6, 8, 16, 17, 18, 18, 60, 1000}, and the corresponding results are shown in Table 5. This time IQR detects all outliers while MAD and IQR fail with one (50).

Table 5: Results of SD, MAD and IQR with x3 = {1, 2, 3, 6, 8, 16, 17, 18, 18, 60, 1000}.

Method	Threshold	Outliers
SD	955.34	{1000}
MAD	60.48	{1000}
IQR	38.25	{60, 1000}

The last example has a large number of outliers. The outlier scores are X4= {1, 2, 3, 6, 8, 16, 17, 18, 18, 60, 300, 500, 1000, 1500}, and the corresponding results are shown in Table 6. This time none of the methods can detect all outliers correctly.

Table 6: Results of SD, MAD and IQR with x4 = {1, 2, 3, 6, 8, 16, 17, 18, 18, 60, 300, 500, 1000, 1500}.

Method	Threshold	Outliers
SD	1574.81	{}
MAD	84.22	{300, 500, 1000, 1500}
IQR	590.25	{1000, 1500}

From the five example cases, we can see that IQR and MAD indeed work better than SD. However, they also fail in several cases when there is large number of outliers.

4.2 Proposed 2T

Now, with the same setting of a and c as in Subsection 4.1, we applied the proposed 2T (Algorithm 1) with StopCondition setting until threshold convergence. The obtained threshold value results are shown in Table 7. We can see that with 2T, MAD and IQR success in all case X0, X1, X2, X3 and X4, whereas SD succeeds in case of X0 but fails with the rest. This is because the original SD value has already been converged, and cannot remove any scores unless we manually remove the biggest scores or tune the parameter a smaller.

To sum up, from the five example cases, we can see that the proposed 2T improves all SD, MAD and IQR.

Table 7: Results of SD, MAD, IQR and 2T with Cases in Table 1.

Cases	Method	2T(converge)
X0	SD	11.82
	MAD	11.90
	IQR	12.00
X1	SD	1283.58
	MAD	11.90
	IQR	12.00
X2	SD	1304.77
	MAD	11.90
	IQR	12.00
X3	SD	64.08
	MAD	39.13
	IQR	38.00
X4	SD	1574.81
	MAD	39.13
	IQR	38.00

4.3 Effect of the Tuning Parameters

It is possible to improve the results by tuning the parameters a and c in Equations (1), (2) and (4). Hence, we tested this with the case X4 by varying a in the range from 2 to 5. We always set $c=a/2$.

The results are summarized in Table 8. We can see that all the original methods (SD, MAD, IQR) fail no matter what value is chosen for a. However, after applying the proposed 2T method, all of them are improved. SD can solve this data with a=2. MAD and IQR work best as they solve this data regardless of what value is chosen for a.

Table 8: Results of Tuning Parameter a with and without Applying 2T with Case x4 = {1, 2, 3, 6, 8, 16, 17, 18, 18, 60, 300, 500, 1000, 1500}.

a	Method	Original	2T(converge)
2	SD	1131.99	23.66
	MAD	61.98	28.76
	IQR	473.50	31.00
3	SD	1574.81	1574.81
	MAD	84.22	39.13
	IQR	590.25	38.00
4	SD	2017.63	2017.63
	MAD	106.46	49.51
	IQR	707.00	45.00
5	SD	2460.44	2460.44
	MAD	128.70	59.89
	IQR	823.75	52.00

4.4 Real-world Datasets

Next, we tested the techniques with real-world datasets from [15], as summarized in Table 9. For the parameters we set a=1, and use 2 iterations in 2T. To produce the outlier scores, we employed two outlier detectors: *mean-shift outlier detection* (MOD) [5] and the detector using the distance to k-th nearest neighbor as outlier score (KNN) [16].

The results are summarized in Table 10 and Table 11 in an order of increasing noise: KDDCup99 (0.4%), Wilt (5.4%), Stamps (9.1%), PageBlocks (10.2%), Cardiotocography (22.2%), Pima (34.9%), SpamBase (39.4%), HeartDisease (44.4%), Arrhythmia (45.8%), and Parkinson (75.4%). We can see that in low noise level (<25%), the original results are better, but in high noise level (>25%), results with 2T are better. However, the proposed 2T always outperforms Clever.

Table 9: Datasets Used in the Experiments.

Dataset	Size	Outliers	Dim	Outlier Object
KDDCup99	60632	246	38	Network attack
Wilt	4839	261	5	Diseased trees
Stamps	340	31	9	Forged stamps
PageBlocks	5473	560	10	Pictures or graphics
Cardiotocography	2126	471	21	Patients
Pima	768	268	8	Patients
SpamBase	4601	1,813	57	Spam email
HeartDisease	270	120	13	Patients
Arrhythmia	450	206	259	Affected patients
Parkinson	195	147	22	Patients

Table 10: F1-score Results for MOD Detector.

Dataset	SD			MAD		IQR	
	Original	2T	Clever	Original	2T	Original	2T
KDDCup99	0.54	0.47	0.00	0.43	0.37	0.48	0.45
Wilt	0.61	0.53	0.05	0.52	0.47	0.59	0.56
Stamps	0.60	0.72	0.60	0.73	0.65	0.60	0.66
PageBlocks	0.68	0.73	0.09	0.66	0.55	0.75	0.72
Cardiotocography	0.55	0.56	0.55	0.56	0.54	0.55	0.55
Pima	0.54	0.62	0.42	0.60	0.63	0.49	0.52
SpamBase	0.55	0.49	0.38	0.56	0.52	0.42	0.43
HeartDisease	0.52	0.54	0.38	0.51	0.54	0.40	0.42
Arrhythmia	0.56	0.65	0.55	0.61	0.67	0.53	0.57
Parkinson	0.35	0.49	0.34	0.42	0.48	0.34	0.36
AVG	0.55	**0.58**	0.34	0.56	0.54	0.51	0.52

Table 11: F1-score Results for KNN Detector.

Dataset	SD			MAD		IQR	
	Original	2T	Clever	Original	2T	Original	2T
KDDCup99	0.54	0.46	0.00	0.43	0.37	0.48	0.45
Wilt	0.61	0.59	0.05	0.57	0.50	0.61	0.60
Stamps	0.58	0.72	0.08	0.75	0.61	0.59	0.69
PageBlocks	0.69	0.74	0.09	0.66	0.55	0.75	0.73
Cardiotocography	0.61	0.56	0.61	0.57	0.53	0.61	0.61
Pima	0.55	0.63	0.49	0.58	0.65	0.49	0.51
SpamBase	0.42	0.48	0.42	0.47	0.51	0.41	0.43
HeartDisease	0.52	0.59	0.38	0.53	0.59	0.42	0.46
Arrhythmia	0.55	0.65	0.31	0.65	0.67	0.54	0.59
Parkinson	0.31	0.42	0.30	0.39	0.46	0.31	0.34
AVG	0.53	**0.58**	0.33	0.57	0.54	0.51	0.54

Table 12: The Amount of Detected Outliers for Mod Detector.

Dataset	Outlier	SD	2T	Clever
KDDCup.	246	3509	9383	48105
Wilt	261	330	1073	4806
Stamps	31	35	79	334
PageB.	560	229	834	5378
Card.	471	183	522	2103
Pima	268	108	228	734
Spam.	1813	693	1047	4186
HeartD.	120	44	85	263
Arrhy.	206	63	136	428
Parki.	147	22	52	174

Table 13: Running Time (s).

Dataset (Size)	SD	2T	Clever
KDDCup. (60632)	<0.01	0.12	811.40
Wilt (4839)	<0.01	0.01	8.39
Stamps (340)	<0.01	<0.01	0.06
PageB. (5473)	<0.01	0.01	10.54
Card. (2126)	<0.01	<0.01	1.69
Pima (768)	<0.01	<0.01	0.26
Spam. (4601)	<0.01	0.01	6.29
HeartD. (270)	<0.01	<0.01	0.05
Arrhy. (450)	<0.01	<0.01	0.10
Parki. (195)	<0.01	<0.01	0.02

4.5 Comparison between 2T and Clever Mean and Clever Standard Deviation

Since 2T looks similar to Clever, we make a comparison between them based on the amount of the detected outlier, listed in Table 12, and running time, listed in Table 13. In Table 12, we can see that Clever detected too many items as outliers, compared to 2T and SD. From Table 13, we can see that Clever requires high computing time, because it removes only one outlier each time.

5 Conclusions

In this paper, a new two-stage thresholding (2T) for outlier detection is proposed. Experimental observations show that no matter which statistics is used for thresholding the outlier scores; the performance can be improved by the proposed 2T. In other words, by first excluding the scores that exceed the original threshold value and then re-calculating the statistics, the actual outlier removal can be performed by using the revised threshold. According to experimental tests shown in this paper, all tested techniques (SD, MAD, IQR) were improved by the proposed approach.

REFERENCES

[1] C Leys, C Ley, O Klein, P Bernard and L Licata (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J Exp Soc Psychol, 49(4), 764-766.

[2] M Gupta, J Gao, C Aggawal and J Han (2014). Outlier Detection for Temporal Data. Morgan & Claypool Publishers.

[3] T V Pollet and L van der Meij (2017). To remove or not to remove: the impact of outlier handling on significance testing in testosterone data. Adaptive Human Behavior and Physiology, 3(1), 43-60.

[4] J P Simmons, L D Nelson and U Simonsohn (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22(11), 1359–1366.

[5] J W Yang, S Rahardja and P Fränti (2018). Mean-shift outlier detection. Int. Conf. Fuzzy Systems and Data Mining (FSDM), In Frontiers in Artificial Intelligence and Applications (FAIA), 309(2), 208-215.

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

[7] N Shah, S F Altschul and M Pop (2018). Outlier detection in BLAST. Algorithms for Molecular Biology, 13(1), 1748-7188.

[8] H E Hawkes and J S Webb (1962). Geochemistry in mineral exploration. New York: Harper & Row.

[9] J Miller (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. The Quarterly Journal of Experimental Psychology, 43(4), 907–912.

[10] F R Hampel (1974). The influence curve and its role in robust estimation. Journal of the American Statistical Association, 69(346), 383–393.

[11] G B Ferraris and F Manenti (2011). Outlier detection in large data sets. Computers and Chemical Engineering, 35(2), 388–390.

[12] P F Marteau (2019). Estimating Road Segments Using Kernelized Averaging of GPS Trajectories. Appl. Sci., 9(13), 2736.

[13] S Cherednichenko, V Hautamäki, T Kinnunen, I Kärkkäinen and P Fränti (2005). Improving k-means by outlier removal. Scandinavian Conf. on Image Analysis (SCIA'05), 978-987.

[14] G Huang, Z Zhang and W Yang (2019). Outlier Detection Method based on Improved Two-step Clustering Algorithm and Synthetic Hypothesis Testing. 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), 915-919.

[15] G O Campos, A Zimek, J Sander, R J G B Campello, B Micenkova, E Schubert, I Assent and M E Houle (2016). On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Mining and Knowledge Discovery, 30(4), 891–927.

[16] S Ramaswamy, R Rastogi and K Shim (2000). Efficient algorithms for mining outliers from large data sets. ACM SIGMOD Record, 29(2), 427-438.