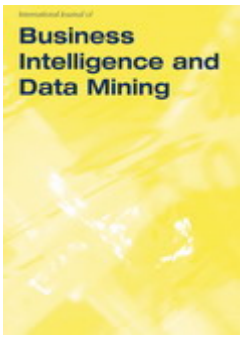
International Journal of Business Intelligence and Data Mining



Performance Evaluation of Outlier Rules for Labeling Outliers in Multidimensional Dataset

Kelly C. Ramos da Silva, Helder L. Costa de Oliveira, André C.P.L.F. de Carvalho

Published in: International Journal of Business Intelligence and Data Mining

View online: https://doi.org/10.1504/IJBIDM.2021.117111

Published online: 21 Jul 2021.

Please cite this paper as: da Silva, K.C.R., de Oliveira, H.L.C. and de Carvalho, A.C.P.L.F. (2021) 'Performance evaluation of outlier rules for labelling outliers in multidimensional dataset', Int. J. Business Intelligence and Data Mining, Vol. 19, No. 2, pp.135–152.

Performance Evaluation of Outlier Rules for Labelling Outliers in Multidimensional Dataset

Kelly C. Ramos da Silva

Institute of Mathematical and Computer Sciences,
University of São Paulo,
São Carlos, SP, Brazil
E-mail: kelly.ramos.silva@usp.br

Helder L. Costa de Oliveira

Institute of Mathematical and Computer Sciences,
University of São Paulo,
São Carlos, SP, Brazil
E-mail: helder.luiz.oliveira@usp.br

André C. P. L. F. de Carvalho

Institute of Mathematical and Computer Sciences,
University of São Paulo,
São Carlos, SP, Brazil
E-mail: andre@icmc.usp.br

Abstract: The output of outlier detection algorithm applied to multidimensional dataset usually consists of scores defining the level of abnormality of each instance. However, this process per se does not identify the outlying instances. For this purpose, it is common to use an outlier rule to convert outlier scores into labels. The problem is therefore to determine an appropriate outlier rule, based on certain patterns of the scores alone. In order to deal with this problem, we studied and evaluated several traditional robust outlier rules following a pragmatic approach. The analysis of the results was facilitated by an evaluation measure developed by us. This measure was proved to be more effective than traditional measures involving only true positive and true negative rates. By using this measure, we were been able to study the behaviour of different outlier rules whose performances were evaluated under varying skewness and contamination level.

Keywords: outlier detection; outlier rule; evaluation measure; boxplot; adjusted boxplot; k-NN.

Reference to this paper should be made as follows: Ramos da Silva, Kelly C. ; Costa de Oliveira, Helder L., and de Carvalho, André C. P. L. F. (201X) 'Performance Evaluation of Outlier Rules for Labelling Outliers in Multidimensional Dataset', *International Journal of Business Intelligence and Data Mining*, Vol. x, No. x, pp.xxx–xxx.

1 Introduction

Although there is no formal definition of an outlier, we can use Hawkins' definition to say: *"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"* [1]. Outlier detection turns out to be very important for many applications, including credit-card fraud detection, intrusion detection computer system, detection of anomalous events in sensors, and medical diagnosis [2].

For the outlier detection task itself, a variety of methods have been proposed [3]. These methods are classified according to the requirement of information in the form of examples of data instances previously classified as outliers or inliers (non-outlier data points). Methods which require such information are classified as supervised, while methods which do not require this information are classified as unsupervised. Methods which make use of both labelled and unlabelled data are classified as semi-supervised. In this paper, we focus on unsupervised outlier detection only.

Most outlier detection algorithms output scores by measuring the level of abnormality of each data instance [4]. The problem with this approach is that scores do not define a concise summary for outliers [2]. Thus, many methods have been proposed in the literature to simplify the identification of outliers when outlier scores are used. For example, the work in [5], the authors convert the scores into probabilities estimates by using a bimodal mixture of normal and exponential distributions. For such, the outliers are assumed to follow the normal distribution, while the inliers are assumed to follow the exponential distribution. In [6], a mixture model, more specifically a beta mixture model, is also used. Nevertheless, instead of converting the scores into probabilities, the aim is to separate the outlier scores into several components, so that the beta component that corresponds to outliers is identified. The main problem with these two approaches lies in the fact that the dataset must contain a sufficient amount of outliers so that the scores can be effectively modelled.

Other very common solutions for the identification of outliers involve to take a manual approach. For example, the popular top-n approach considers the n data instances with the highest scores as outliers. This approach obviously considers an unrealistic scenario, where the exact number of outliers in a dataset is assumed to be known in advance [7]. Other manual approach consists of plotting the sorted outlier scores and then choosing the knee point of the curve as a threshold in such a way that any instance whose score is larger than this threshold is declared as an outlier [5]. Note, however, that such a choice may be highly subjective and significantly affect the final result, since an inappropriate threshold could either flag a very large number of inliers as outliers or label many outliers as inliers.

Alternatively, a threshold could be automatically determined by using an outlier rule. It must be observed that the determination of the optimal outlier rule might be very dependent on the underlying distribution of the scores. Nevertheless, in real world applications, the underlying distribution of the data (scores in this case) can be completely unknown. In this case, a reasonable solution is to use a generic (non-parametric) outlier rule. This raises an important question: which is the most appropriate non-parametric outlier rule? In order to answer this question, this paper focuses on the study of the properties and performance evaluation of several commonly used outlier rules, considering both synthesised and real world datasets. Note that by non-parametric, we mean a procedure which is supposed to be used for a broad, not parametrised, set of underlying distributions [8].

The main contributions of this paper are twofold: first, it studies and evaluates performance and behaviour of popular outlier rules when applied to outlier scores in order

to automatically identify the outliers. Second, it introduces an evaluation measure to assess, in a pragmatic way, performance of thresholds yielded by outlier rules.

2 Outlier Rules

A common technique for establishing an automatic threshold for outlier scores of multidimensional data, in order to convert them into labels, is the use of an outlier rule.

An outlier rule, also known as an outlier identifier [9], is a rule that establishes thresholds so as to identify the tails of a distribution. This type of rule usually defines two thresholds (one lower and one upper). Values that are above the upper threshold or below the lower threshold are labelled as outliers. An outlier identified in this way belongs to a very specific category of outliers, which in this case is called an extreme value [10].

Outlier rules are typically developed to be applied to univariate data. Since the outlier scores of multidimensional data are also typically univariate, the application of an outlier rule to these scores would ideally have the points marked as outliers in the one-dimensional space corresponding to the outliers in the multidimensional space. In a real situation, false positive or false negative errors are expected to occur.

In this paper, we shall focus on the study and evaluation of six outlier rules addressing the problem of automatic labelling of outliers in multidimensional data. For generating outlier scores, we shall use the k-NN (k-nearest neighbours) algorithm[11, 12]. This algorithm outputs outlier scores by summarizing the k-nearest neighbours distances of each data point. As a summary measure, maximum [11] and mean [12] are the most commonly used. However, any distance metric can be used by k-NN to measure the distance between a pair of points. In this paper we use k-NN with the mean as a summary measure and euclidean distance as the distance metric.

Although we use k-NN for outlier scoring, in theory, it is expected that the analysis and conclusions obtained should not change if another outlier scoring method satisfying certain requirements were used. These requirements are that the method must output outlier scores with the following characteristics: enough contrast for a good distinction between inliers and outliers; unimodality; outliers lie in the upper tail. If outliers lie in the lower tail, the lower threshold of the considered outlier rule should be used instead of the upper threshold. It is also desirable that the underlying distribution of the score be symmetric (or approximately symmetric) or right skewed, since detection of outliers lying in the upper tail of left skewed distributions is not focus of this study. Note that one or more of these requirements are due to requirements or limitations of the outlier rules selected for this study. It is worth mentioning that although one of the datasets used in this study exhibits scores whose distribution is left-skewed, we are regarding it as approximately symmetric, as it is only slightly left-skewed.

The outlier rules selected for this study are listed in Table 1. It is worth mentioning that, although all these rules originally include the determination of lower and upper thresholds, for this work, only the upper thresholds will be the focus of interest, since only the points with scores higher than a given threshold will be regarded as outliers. Thus, we assume the outliers are located only in the upper tail of the outlier score distribution. This assumption is very reasonable, since the k-NN algorithm outputs scores in such a way that outliers are characterised by high scores while inliers are characterised by low scores.

The selected rules include the most common non-parametric rules, such as Tukey's boxplot (Section 2.1), the SIQR boxplot (Section 2.2) and the MAD rule (Section 2.3).

In addition to these rules, we consider a more recently developed outlier rule, known as the adjusted boxplot (Section 2.4). It is worth pointing out that we are considering only resistant/robust rules, since non resistant rules encounter problems of masking effect [9, 13]. Masking effect refers to the problem where some or even all outliers are not detected because they are masked by the presence of other outliers. This problem is related to the concept of breakdown point, which is defined as *the smallest amount of contamination that may cause an estimator to take on arbitrarily large aberrant values* [14]. Because of this issue, we did not consider using outlier rules that rely on the mean and the standard deviation, since these estimators are not robust, i.e they have breakdown point of 0%.

In addition to being non-parametric and robust, the selected outlier rules have other good characteristics, such as simplicity, low computational cost, high breakdown point and no need of prior knowledge of the number of outliers present in the dataset.

2.1 Tukey's Method (Classical Boxplot)

A popular outlier rule is based on a graphical tool for univariate data analysis known as boxplot. The boxplot, proposed by Tukey [15], defines four thresholds based only on Q_1 (first quartile) and Q_3 (third quartile). The two most internal thresholds (called inner fences) are defined as extreme ends of the following interval:

$$[Q_1 - 1.5IQR; \quad Q_3 + 1.5IQR].$$

While the other two thresholds (called outer fences) are given by the extremes of the following interval:

$$[Q_1 - 3IQR; \quad Q_3 + 3IQR].$$

With

$$IQR = Q_3 - Q_1 \quad \text{interquartile range.}$$

A common criterion, when using these rules, states that values outside the inner fences, but inside the outer fences, are defined as weak outliers, while values outside the outer fences are defined as strong outliers. Note that there is no requirement for the quartiles to be equidistant with respect to some location representing the centre of the distribution. Therefore, apparently, there is also no assumption of symmetry, and thus we could conclude that Tukey's method would be suitable for both symmetric and skewed distributions. In fact, there is an implicit assumption of symmetry and it is found in the constant multiplying the IQR. For example, for the internal fences, which have 1.5 as a constant, it would be expected a false positive rate of 0.70% when applied to a normal distribution (symmetric distribution). While for the exponential distribution (skewed distribution), the expected false positive rate would be much higher: 4.81%.

Note that the boxplot has breakdown point of 25% since this is the breakdown point of quartiles. In addition, this rule can be computed in $O(n)$ time, since quartiles can be computed in $O(n)$ time [16].

For our study, we shall use only the upper boundary of each interval and independently of each other.

2.2 The SIQR Boxplot

Kimber [17], after realizing that Tukey's boxplot would not adjust itself well for skewed distributions, proposed, without any formal demonstration, the following rule:

$$[Q_1 - 3SIQR_L; \quad Q_3 + 3SIQR_U].$$

With

$$SIQR_L = Q_2 - Q_1; \quad SIQR_U = Q_3 - Q_1.$$

Kimber's rule splits the interquartile range into two dispersion measures ($SIQR_L$ and $SIQR_U$). Unlike the classical boxplot, the lower and upper thresholds take into account dispersions that may be asymmetric. In the symmetrical case, the SIQR boxplot becomes exactly the classical boxplot. Despite these modifications, the SIQR boxplot is known to produce only slightly better results than those yielded by Tukey's boxplot for skewed distributions.

2.3 The MAD Rule

The MAD (Median Absolute Deviation), promoted by Hampel (1974) [18], but attributed to Carl Friedrich Gauss, is a robust scale/dispersion estimator. By combining the MAD as a dispersion estimator and the median as a location estimator, an outlier rule can be obtained. This rule has a breakdown point of approximately 50%, since the breakdown point of the median is 50% and the MAD itself is approximately 50% [19]. The rule in question, also known as the Hampel identifier [9], has the form $Q_2 + aMAD$ for the upper threshold and $Q2 - aMAD$ for the lower threshold. $Q2$ (second quartile) is the sample median and the constant a is usually employed between 2 and 3.
Consider $X = \{x_1, x_2, \cdots, x_n\}$. The MAD can be calculated as follows:

$$MAD = bMed(|xi - Med(X)|).$$

Where $Med(X)$ is the median of X.

For a given distribution in its standard form, it is common to use the constant b as $1/Q_3$ [20]. When the distribution is the standard normal distribution, $Q_3 = 0.6745$. Thus, b results in 1.4826. The use of b = 1.4826 has the property to make the MAD a robust standard deviation estimator of a normal distribution population.

An important point to mention about the MAD is its implicit dependence on the distribution symmetry [19]. This implies that the MAD should not produce good results for skewed distributions.

Just like the boxplot, the MAD rule can be computed in $O(n)$ time.

2.4 The Adjusted Boxplot

The adjusted Boxplot [21], proposed by Hubert and Vandervieren, is exactly the inner fences of Tukey's boxplot adjusted to take into account the distribution skewness.

When the boxplot is applied to skewed distributions, the skewness tends to increase the amount of false positives in the detection of outliers for the upper tail and the amount of

false negatives for the lower tail of a right skewed distribution. The reverse is true for left skewed distributions. In order to solve this problem, Hubert and Vandervieren proposed the use of the medcouple (robust measure of skewness) [22] to obtain a correction factor for the boxplot when the distribution is skewed.

Let $S = \{x_1, x_2, \cdots, x_n\}$ be a sample from a continuous unimodal distribution. The medcouple can be determined as follows:

$$MC = Med\frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i}.$$

Where Q_2 is the median of S; $\quad x_i \leq Q_2 \leq x_j \quad x_i \neq x_j$.

The values of MC range from -1 to 1. $MC > 0$ indicates that the distribution is right skewed, while $MC < 0$ indicates that the distribution is left skewed. The value $MC = 0$ indicates that the distribution is symmetric.

Using the MC with various models to adjust the boxplot rule, Hubert and Vandervieren found that the best model for this adjustment was an exponential model. To sum up, the inner fences of the classical boxplot are redefined as:

If $MC \geq 0$

$$[Q_1 - 1.5e^{-4MC}IQR; \quad Q_3 + 1.5e^{3MC}IQR]$$

If $MC < 0$

$$[Q_1 - 1.5e^{-3MC}IQR; \quad Q_3 + 1.5e^{4MC}IQR].$$

Note that when the distribution is symmetric ($MC = 0$), the adjusted boxplot rule becomes the classical boxplot rule. An important observation to be made regarding the adjustment of the boxplot rule is that it was constructed based on distributions with moderate skewness ($|MC| \leq 0.6$). This indicates that this adjustment might not produce as good results for highly skewed distributions as those produced in the Vandervieren and Hubert's simulations for distributions with moderate skewness.

Just like the boxplot, the adjusted boxplot has breakdown point of 25%, since the MC and quartiles have breakdown point of 25% [22]. The MC is also responsible for increasing, in relation to the boxplot, the computational complexity of the adjusted boxplot, which is $O(n\ log\ n)$.

3 Performance Evaluation of Outlier Rules

When evaluating and comparing thresholds produced by outlier rules, one key question that arises is: Is it better to have a threshold that is closer to the end of the distribution of the inliers or closer to the beginning of the distribution of the outliers? The answer to this question can be purely subjective. One person, for example, might argue that a threshold closer to the outliers may be more appropriate, since the distribution of the inliers may consist of missing data (for example, due to censorship or truncation). Whereas another person could think the opposite and consider that the threshold that is closer to the end of the distribution of the inliers is better, since such a person would have the view that the

Table 1 Summary of the outlier rules (upper thresholds only)

Rule	Threshold	Type
Boxplot 1	$Q_3 + 1.5(Q_3 - Q_1)$	Boxplot
Boxplot 2	$Q_3 + 3(Q_3 - Q_1)$	
MAD 1	$Q_2 + 2\,MAD$	MAD
MAD 2	$Q_2 + 3\,MAD$	
Adj. Boxplot	$Q_3 + 1.5\,e^{bMC}(Q_3 - Q_1) \quad b = \begin{cases} 3 & : \quad MC \geq 0 \\ 4 & : \quad MC < 0 \end{cases}$	Adjusted Boxplot
SIQR Boxplot	$Q_3 + 3(Q_3 - Q_2)$	SIQR Boxplot

Note: MAD 1 and MAD 2 rules use MAD with b = 1.4826 (see Section 2.3)

data form a closed world and therefore any instance outside this world should be flagged as outlier. Although both arguments are valid, the first approach may be more risky since, in this case, even if there is no incidence of false negative for a given configuration of outliers, inherently the threshold closer to the outliers will be associated with a higher probability of false negative than that of a threshold closer to the inliers. Thus, in this paper we shall adopt the second view in the evaluation of the threshold obtained through a given outlier rule, especially taking into consideration that for most applications a false negative has a much greater weight than a false positive to signal a decrease in performance.

3.1 Performance Evaluation Measure

Sometimes, outliers reveal themselves very separated from inliers. As a result, a large gap may exist between inliers and outliers.

The problem is that this large gap makes it more difficult to distinguish performance between outlier rules whose thresholds fall within it, as, in this case, an evaluation measure involving only TP_r (True Positive rate) and TN_r (True Negative rate) would yield the same measurement for these thresholds, regardless of whether a given threshold is closer to the beginning of the distribution of outliers or closer to the end of the distribution of inliers. Even in situations where the thresholds are not in the gap there could still be a problem with the distinction of performance between outlier rules. One example for this situation is when two given thresholds, which are in the same region as the inliers, exhibit almost the same true negative rate when there is in fact a significant distance between these thresholds. Since in this case the evaluation measurement would be dominated by TN_r, the outlier rules associated with these thresholds would be assigned to almost undistinguishable performances, although the distance between these thresholds demonstrates a clear difference of performance. So as to resolve these issues, we shall consider an evaluation measure that takes into account not only true positive and true negative rates, but also the difference between the threshold value and the highest score of the inliers. Obviously, we must also establish a scale for such a measurement. For this purpose, we could somehow take into account the variability of the data for the side of the tail in question, that is, variability above the median or bellow the median. Since in this

paper our interest lies in the upper tail, we shall compute the variability above the median. Thus, by following this reasoning we would have a measure as follows:

$$DE = \left| \frac{value - max_N}{max_N - median_N} \right|. \tag{1}$$

With

max_N: the highest score of the inliers.
$median_N$:the median score of the inliers.
$value$: the threshold value.

In case the threshold lies in the gap, this measure is related to a probability of occurrence of FN (false negatives), in such a way that the value zero is related to a 0% probability of FN occurrence, while a value tending to infinity is asymptotically related to a 100% probability of FN occurrence. One problem is that this measure is not normalised. Then to normalise it to values between 0 and 1, we can apply a transformation such as: $1/(1+x)$. Thus, we can define a normalised measure relating the distance between the highest score of the inliers and the threshold value as follows:

$$NDE = \frac{1}{1 + 0.4 \left| \frac{value - max_N}{max_N - median_N} \right|}. \tag{2}$$

Finally, as an evaluation measure relating TP_r, TN_r and NDE, we shall use the following geometric mean (inspired by the G$mean$ [23]):

$$GME = \sqrt[3]{TP_r.TN_r.NDE}. \tag{3}$$

With

$$TP_r = \frac{TP+1}{P+1}; \quad TN_r = \frac{TN}{N}. \tag{4}$$

It is worth mentioning that the factor 0.4 in Equation 2 is due to the criterion of making the result of GME approximately the same as that of G$mean$ when the threshold value coincides with the median of the scores of the inliers. Another point to note is that, since we are using the data variability for the upper tail by taking the median as the reference point for the separation between the lower and upper tails, this measure is appropriate only for values of threshold greater than or equal to the median. For this work, this is not a problem, since all the rules we are evaluating satisfy this condition.

Note that both numerator and denominator of TP_r in Equation 4 are added by 1. This is necessary to take into account the case of 0% outliers, since in this case P is zero. This "1" can also be interpreted as the existence of an additional outlier with an infinite score.

4 Experimental Evaluation

The experiments have been conducted by adding varying percentage of outliers to uncontaminated data. The outliers have been generated following a process of generating

random points outside the regions determined by the points of the classes in a given dataset. For this process, we consider that the points of each class are delimited by an imaginary inner hyper-parallelepiped. This hyper-parallelepiped is inflated to form a forbidden region for outliers. Thus, the minimum and maximum in each dimension of this hyper-parallelepiped is, respectively, decreased and increased by 10%. We also consider that all the points of the dataset are delimited by an imaginary outer hyper-parallelepiped, which is also inflated in the same way as described earlier, but 25% is used instead of 10%. Thus, points are randomly generated in such a way that if a point is generated inside the outer hyper-parallelepiped, but outside all inner hyper-parallelepipeds, this point will be a valid outlier, otherwise this point will be disregarded and then a new one will be randomly generated.

It is worth pointing out that the process described earlier considers uncontaminated dataset. Furthermore, R software (version 3.3.0) with the robustbase add-on package (version 0.92-7) has been used to implement all the experiments. The robustbase package has been used since it implements the medcouple as well as the adjusted boxplot.

4.1 Experiments with Real World Datasets

All the real world datasets used in this paper have been retrieved from UCI machine learning repository [24].

4.1.1 Methodology

As our methodology requires clean data, it was necessary to perform a pre-processing stage in order to cleanse the data. This stage consists of the following steps:

1. Normalise the data; a min-max normalization has been used.
2. Apply the k-NN algorithm.
3. Mark as suspicious the instances whose percentile is at least 75%.
4. Repeat the steps 2 and 3 for $k = 5 \ldots 10$.
5. Remove from the original dataset (non normalised dataset) instances marked as suspicious at least 4 times.
6. Normalise the cleansed data.

For performance evaluation of each outlier rule, the k-NN method has been applied to the data varying k and percentage of outliers. Thus, starting from a given clean dataset, versions of this dataset with different percentages of outliers were generated to be used to evaluate each outlier rule for each k. This whole process was performed 100 times, so that, for a same percentage of outliers, 100 different configurations of outliers were generated. These different configurations for a same percentage of outliers are necessary due to the completely random aspect of the outlier generation, which might eventually generate a configuration of outliers that could favour some outlier rules and significantly harm others. $k = 5, 6, \cdots 20$, and percentages of outliers: $0\%, 1\%, 2\%, \cdots, 20\%$ were used. At the end of all the repetitions for all combinations of k and percentages of outliers, there were, for each dataset, 33600 evaluations for each outlier rule. Note that the GME was the evaluation measure used at this stage.

As a summary of performance for a given outlier rule, considering the set of GME measurements (obtained for all combinations of k and percentages of outliers in question), we shall use the following efficiency measure:

$$eff = \frac{median(X)}{1 + MAD(X)}. \tag{5}$$

Where X is the set of GME measurements.

The idea behind this measure is to penalise rules that deliver unstable performance, marked by great variability of GME measurements, i.e. great value of $MAD(X)$.

4.1.2 Organization of the Experimental Results

For a given dataset, the set of GME results was partitioned into ranges of contamination levels and then the efficiency measure (see Equation 5) was applied independently to each range to summarise the results in one table (see e.g. Table 2). Note that in each table the ordering of rows is carried out from the rule associated with the best performance (top) to the rule associated with the worst performance (bottom). The sorting criterion is based on the (geometric) average of the efficiency measurements for the first three ranges of contamination levels. The last range is disregarded for such a calculation, being in the table only as additional information, for the reason that many researchers do not consider contamination levels above 15%. Furthermore, the performance (Best eff) achieved by the best rule for each range is also being shown for the range in question for a given table. The percentages in the cells refer to how many percent below this highest efficiency value the performance of a given rule was. The number in parentheses to the right of the percentage refers to the position of a given rule in the ranking of performance for the range in question. Mean and standard deviation of the medcouple (MC) for the uncontaminated data considering all k values are also being reported in the tables.

In addition, the thresholds produced by the rules are represented in figures (see e.g. Figure 1) showing plots of (normalised) dataset row index versus outlier scores. Note that, for a better visualization, the scores have been transformed by applying power law with exponent equal to 0.02. For each plot, for a given dataset, the number of neighbours (k) equal to 10 has been selected as it has produced results that match the average result of the outlier rule ranked first for the first range of contamination levels ($0\% \ldots 5\%$).

Another visualization we are considering for a given dataset is the plot showing the influence on each rule caused by varying contamination levels (see e.g. Figure 2). For this kind of plot, the threshold variations have been calculated for contaminated data with respect to the threshold for uncontaminated data.

4.1.3 Shuttle Dataset

The shuttle dataset contains measurements of radiator positions in a NASA space shuttle with 9 numeric attributes and 43499 instances.

As can be seen from Table 2, the adjusted boxplot achieved the best performance with significant differences in relation to the performances of the other rules.

When inspecting Figure 1, it becomes very clear that these differences are due to the fact that the adjusted boxplot presents a much lower number of false positives than the other rules, since the classical boxplot, MAD (1 and 2) and SIQR boxplot rules fail to follow the skewness of the distribution of the scores.

Table 2 Experimental results for the shuttle dataset.

Rule	$MC = 0.41 \pm 0.06$			
	0%...5% (Best $eff = 0.906$)	5%...10% (Best $eff = 0.913$)	10%...15% (Best $eff = 0.925$)	15%...20% (Best $eff = 0.950$)
Adj. Boxplot	0% (1)	0% (1)	0% (1)	0% (1)
Boxplot 2	-1.195% (2)	-1.515% (2)	-2.149% (2)	-3.951% (2)
SIQR Boxplot	-2.308% (3)	-2.489% (3)	-2.899% (3)	-4.577% (3)
Boxplot 1	-3.081% (4)	-3.201% (4)	-3.615% (4)	-5.266% (4)
MAD 2	-4.025% (5)	-4.218% (5)	-4.718% (5)	-6.688% (5)
MAD 1	-6.051% (6)	-6.050% (6)	-6.340% (6)	-8.147% (6)

Note: see Section 4.1.2 for explanation of the organization of the results in the table.

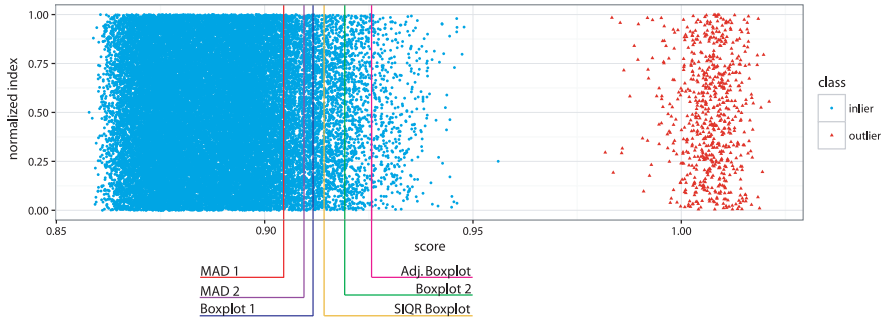


Figure 1 Plot of index versus outlier score for the shuttle dataset with 2% outliers.

From Figure 2, it can be seen that the adjusted boxplot is more heavily affected by contamination (especially above 10% outliers) than the rest of the rules. By analysing, once again, Table 2 and Figure 1 we realise that this larger influence favours the adjusted boxplot's performance since the number of false positives becomes lower and lower as the proportion of outliers increases. Furthermore, in this case, up to 20% contamination was not enough to cause any outlier rule to set its threshold in the gap.

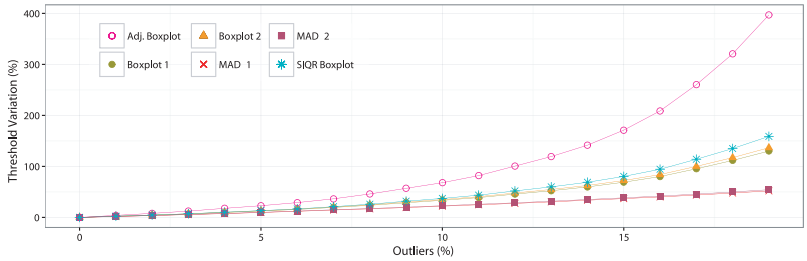


Figure 2 Plot showing the influence of varying contamination levels on each outlier rule for the shuttle dataset.

Table 3 Experimental results for the abalone dataset

Rule	$MC = 0.16 \pm 0.02$			
	0%...5% (Best $eff = 0.923$)	5%...10% (Best $eff = 0.935$)	10%...15% (Best $eff = 0.951$)	15%...20% (Best $eff = 0.956$)
Adj. Boxplot	-0.6794% (2)	-0.07312% (2)	0% (1)	0% (1)
Boxplot 2	0% (1)	0% (1)	-1.047% (2)	-0.397% (2)
SIQR Boxplot	-2.062% (3)	-1.947% (3)	-3.022% (3)	-2.453% (3)
MAD 2	-2.241% (4)	-2.388% (4)	-3.551% (5)	-3.572% (5)
Boxplot 1	-2.416% (5)	-2.436% (5)	-3.543% (4)	-3.217% (4)
MAD 1	-4.406% (6)	-4.328% (6)	-5.329% (6)	-5.172% (6)

Note: see Section 4.1.2 for explanation of the organization of the results in the table.

4.1.4 Abalone Dataset

The abalone dataset consists of data from a study concerning the physical measurements of edible sea snails belonging to the Haliotis species. This dataset consists of 4177 instances with 7 numeric attributes, one categorical and one attribute with the classes. Note that the non numeric attributes have been disregarded.

As can be seen from Table 3, the distribution of the scores for the abalone dataset presents lower skewness than for the Shuttle dataset. Not surprisingly, the difference of performance between the adjusted boxplot and any other rule is also smaller. Furthermore, despite exhibiting the best performance for the first two percentage ranges, the boxplot 2 was not very superior to the adjusted boxplot in these two ranges. Taking into account that the adjusted boxplot has a more significant difference of performance in the third range, this result gives the adjusted boxplot the first overall position in terms of performance.

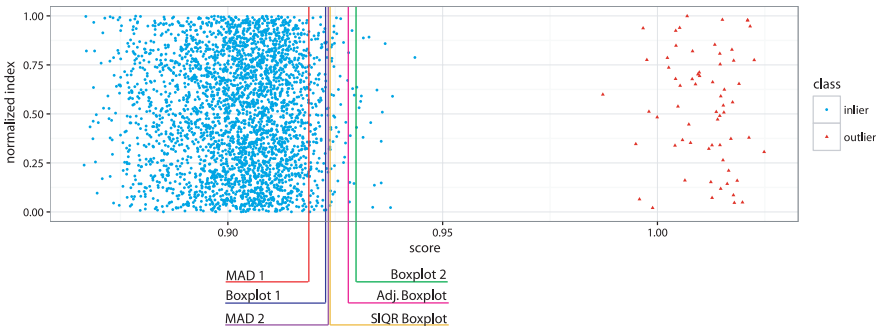


Figure 3 Plot of index versus outlier score for the abalone dataset with 2% outliers.

In both Figure 3 and Table 3, we can see that the difference of performance between MAD 2, boxplot 1 and SIQR boxplot is not significant.

From Figure 4, we can see that the adjusted boxplot is, again, the rule that is the most affected by the variation of percentage of outliers. Furthermore, due to the same reason as that for the case of the Shuttle dataset, this greater influence plays a role in favour of the adjusted boxplot since its performance becomes higher and higher as the percentage of outliers increases (up to 20%).

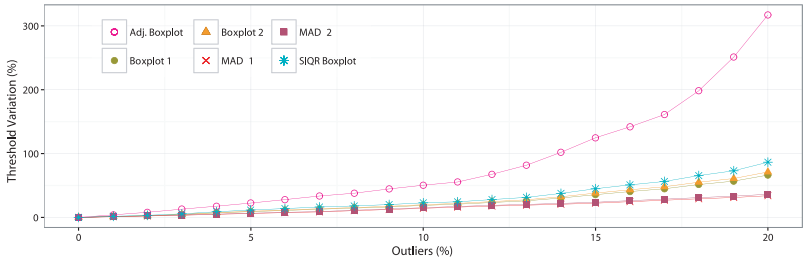


Figure 4 Plot showing the influence of varying contamination levels on each outlier rule for the abalone dataset.

4.1.5 Waveform Dataset

The waveform dataset is a dataset for the study of the problem of three classes of triangular waveforms. It consists of 4999 instances with 21 continuous attributes.

Table 4 Experimental results for the waveform dataset

Rule	$MC = -0.02 \pm 0.01$			
	0%...5% (Best eff = 0.980)	5%...10% (Best eff = 0.984)	10%...15% (Best eff = 0.991)	15%...20% (Best eff = 0.980)
MAD 1	-0.6358% (4)	0% (1)	0% (1)	0% (1)
Boxplot 1	0% (1)	-1.661% (2)	-4.102% (2)	-4.865% (2)
SIQR Boxplot	-0.1479% (2)	-2.076% (3)	-4.754% (3)	-5.980% (4)
MAD 2	-1.525% (5)	-3.102% (4)	-5.221% (4)	-5.707% (3)
Adj. Boxplot	-0.3452% (3)	-3.748% (5)	-9.792% (5)	-17.570% (6)
Boxplot 2	-8.731% (6)	-10.430% (6)	-12.870% (6)	-13.790% (5)

Note: see Section 4.1.2 for explanation of the organization of the results in the table.

By analysing the Table 4, we see completely different results from the results for the previous data. This time the result is almost the opposite, since the MAD 1, MAD 2, boxplot 1 and SIQR boxplot rules are now better ranked than the adjusted boxplot. However, when considering low percentage of outliers, these rules do not present a large difference of performance compared to the adjusted boxplot, as can be seen from Figure 5. Indeed, it is not even visually noticeable from the figure that the adjusted boxplot is slightly ahead of the SIQR boxplot and that the boxplot 1 is slightly behind the SIQR boxplot. This result can be better understood when we observe that the (mean) MC is approximately -0.02 for the distribution of the scores of the inliers. This value indicates that such a distribution is approximately symmetric. Thus, in this case, the adjusted boxplot tends to reach performance very close to that of the boxplot 1 (and consequently the SIQR boxplot). In spite of this fact, as the percentage of outliers increases, the difference of performance between the adjusted boxplot and the boxplot 1 becomes increasingly significant, as can be seen from Figure 6. From the figure, it becomes clear that the adjusted boxplot is the rule most affected by the presence of outliers. As can also be seen from Table 4, the boxplot 2 exhibits significantly lower performance than the other rules. From Figure 5, we can see that this rule is very conservative for the Waveform dataset. Again, this result has to do with the fact that the distribution of the scores is approximately symmetric for this dataset.

Figure 5 Plot of index versus outlier score for the waveform dataset with 2% outliers.

Note that unlike what has been observed for the previous datasets, this time we see that the influence of outlier plays a role against the performance of the rules. The reason for this behaviour is that the rules set thresholds in the gap. Thus, as the percentage of outlier increases the thresholds in the gap becomes increasingly distant from the inlier with the highest score.

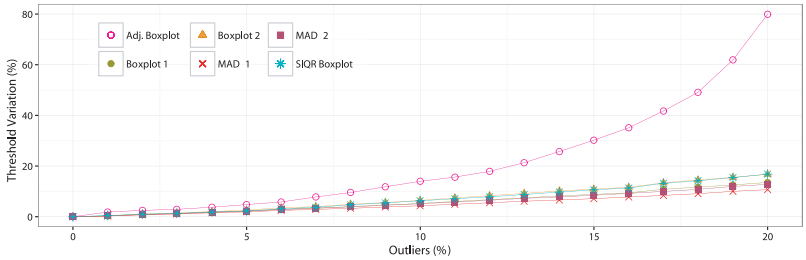


Figure 6 Plot showing the influence of varying percentages of outliers on each outlier rule for the waveform dataset.

4.2 Effect of Skewness on the Outlier Rules

In order to evaluate the rules from the perspective of varying skewness, we have used the lognormal distribution to generate 80 datasets of 1000 instances each. For each one of these datasets, the logarithm of the instances have presented mean zero and standard deviation between 0.01 and 4. Unlike the case of real world data, the outliers rules were applied directly to these generated datasets, since they were unidimensional data with unimodality. We could also think of these datasets as being synthesised outlier scores.

By evaluating the results of the GME measure for each rule, we constructed GME versus MC plots for a given percentage of outliers. The MC, however, refers to the data of inliers, since the MC value of contaminated data is influenced by the outliers. Note that the outliers have been generated in the upper tail only.

As can be seen from Figure 7, although the boxplot 2, due to its more conservative behaviour, generally achieves better performance for the asymmetric case than the boxplot 1, it still turns out to be significantly inferior to the adjusted boxplot. Moreover, this better

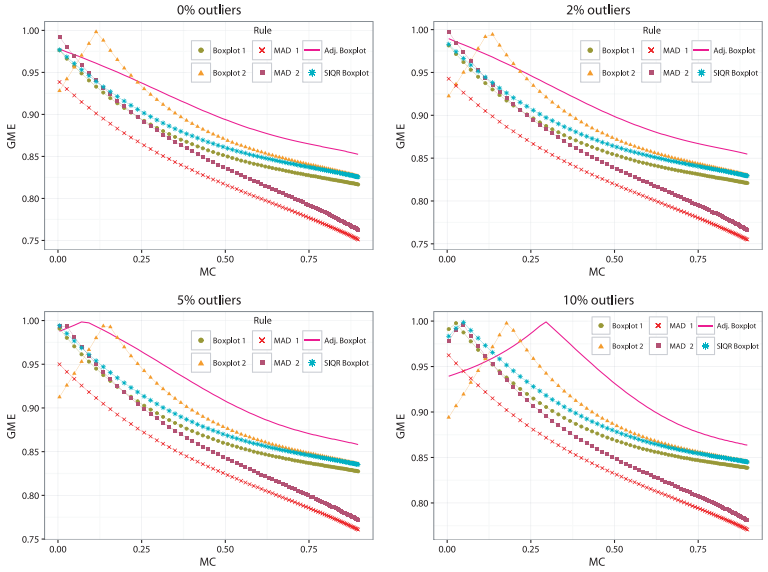


Figure 7 Results for the lognormal datasets generated with different skewness and percentage of outliers.

performance for the asymmetric case is accompanied by a worse performance for the symmetric case. With respect to the symmetric case itself, it is seen that, as the percentage of outliers increases, the adjusted boxplot has a considerable performance drop relative to the other rules, reaching the point that at 10% contamination its performance is lower than that of the MAD 1 rule.

As could be seen for the case of the Waveform dataset, this behaviour has to do with the fact that the adjusted boxplot is more significantly affected by the presence of outliers than the other rules. However, as the skewness increases the adjusted boxplot achieves the best overall performance (when considering all contamination levels in question).

It can also be noted that although the SIQR performs better than the boxplot 1, it is even inferior to the boxplot 2 when asymmetric cases are being considered. The MAD 2 is the rule that achieves the best performance for the symmetric case, but just like the other rules, in relation to the adjusted boxplot, it has low performance for asymmetric cases. Note that, for certain percentages of outliers, some rules reach a performance peak for a very specific skewness and then begin to exhibit increasingly low performance as the skewness increases. The reason for this behaviour lies in the fact that before the peak is reached, the threshold is in the gap. However, as the skewness increases the threshold becomes closer and closer to the inlier with the highest score, which leads to an increasingly high GME measurement. After reaching the performance peak, it is possible to observe, with respect to performance, that the opposite occurs, given that the false positive rate becomes increasingly high as the skewness increases. These situations demonstrate that the GME was able to provide performance contrast between rules when it would not be possible if another evaluation measure involving only true positive and false negative rates had been used.

It is also remarkable that, in terms of ranking and proximity of performance, these results considering lognormal distributions agree to a large extent with the results of the abalone and shuttle real world datasets. However, when comparing these results of synthesised data

with the waveform dataset results, we can see that for both these cases the results are very similar for some rules, but significantly different for others. For example, the MAD 2 clearly achieves better performance for the lognormal datasets than the MAD 1. Nevertheless, with respect to the waveform dataset results, the opposite is observed, as the MAD 1 achieves better overall performance than the MAD 2. The main reason for this difference of results lies in the fact that the distribution of outlier scores for the waveform dataset turns out to be more truncated than for the case of the lognormal datasets.

5 Conclusion

With regard to the problem of setting threshold on outlier scores yielded by the k-NN, the results presented in this paper clearly show that the adjusted boxplot rule achieves significantly better overall performance than the other tested rules, especially for outlier scores with underlying asymmetric distribution. Even for the approximately symmetrical case with up to 5% contamination, the performance of the adjusted boxplot was almost the same as that of the rule with the best performance for this case. Thus, it is natural to recommend the use of the adjusted boxplot when there is no prior knowledge on the underlying distribution of the scores of the uncontaminated data, especially in relation to the asymmetry. When there is knowledge that the distribution is symmetrical or approximately symmetrical, the recommendation is the use of an outlier rule based on the MAD or even the classical boxplot (boxplot 1), since the adjusted boxplot turned out to be the outlier rule whose overall performance is the most affected by the presence of outliers.

It should also be noted that the clarity of the results was facilitated by the proposed evaluation measure, which allowed performance contrast between thresholds in situations where no contrast would be possible with measures that take into account only the relationship between TP_r (true positive rate) and TN_r (true negative rate).

In future research, we intend to explore some ideas that we believe can improve the identification of outliers, such as combining the outputs of outlier rules, as well as transforming outlier scores before applying an outlier rule.

Acknowledgements

The authors are grateful to Be Mundus - Brazil Europe Erasmus Mundus and PROEX (Programa de Excelência Acadêmica) - CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for the sponsorship of this research work.

References

[1] Douglas M Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

[2] Charu C Aggarwal. *Outlier Analysis*. Springer, 2016.

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[4] Shebuti Rayana, Wen Zhong, and Leman Akoglu. Sequential ensemble learning for outlier detection: A bias-variance perspective. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1167–1172. IEEE, 2016.

[5] Jing Gao and Pang-Ning Tan. Converting output scores from outlier detection algorithms into probability estimates. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 212–221. IEEE, 2006.

[6] Mohamed Bouguessa. Modeling outlier score distributions. In *International Conference on Advanced Data Mining and Applications*, pages 713–725. Springer, 2012.

[7] Hans-Peter Kriegel, Peer Kroger, Erich Schubert, and Arthur Zimek. Interpreting and unifying outlier scores. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 13–24. SIAM, 2011.

[8] Elvezio M. Ronchetti Peter J. Huber. *Robust Statistics, Second Edition*. Wiley Series in Probability and Statistics. 2009.

[9] Laurie Davies and Ursula Gather. The identification of multiple outliers. *Journal of the American Statistical Association*, 88(423):782–792, 1993.

[10] James Pickands III. Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131, 1975.

[11] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *ACM Sigmod Record*, volume 29, pages 427–438. ACM, 2000.

[12] Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 15–27. Springer, 2002.

[13] David C Hoaglin, Boris Iglewicz, and John W Tukey. Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 81(396):991–999, 1986.

[14] David L Donoho and Peter J Huber. The notion of breakdown point. *A festschrift for Erich L. Lehmann*, 157184, 1983.

[15] John W Tukey. *Exploratory data analysis*. Reading, Mass., 1977.

[16] Manuel Blum, Robert W Floyd, Vaughan Pratt, Ronald L Rivest, and Robert E Tarjan. Time bounds for selection. *Journal of computer and system sciences*, 7(4):448–461, 1973.

[17] AC Kimber. Exploratory data analysis for possibly censored data from skewed distributions. *Applied statistics*, 39(1):21–30, 1990.

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

[19] Peter J Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.

[20] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.

[21] Mia Hubert and Ellen Vandervieren. An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, 52(12):5186–5201, 2008.

[22] Guy Brys, Mia Hubert, and Anja Struyf. A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 13(4):996–1017, 2004.

[23] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *ICML*, volume 97, pages 179–186. Nashville, USA, 1997.

[24] Andrew Frank and Arthur Asuncion. Uci machine learning repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of california. *School of Information and Computer Science*, 213, 2010.