Journal of Big Data

RESEARCH Open Access

Unsupervised outlier detection in multidimensional data

Atiq ur Rehman* and Samir Brahim Belhaouari

*Correspondence:
atrehman2@hbku.edu.qa;
atiqjadoon@gmail.com
ICT Division, College
of Science and Engineering,
Hamad Bin Khalifa University,
Doha, Qatar

Abstract

Detection and removal of outliers in a dataset is a fundamental preprocessing task without which the analysis of the data can be misleading. Furthermore, the existence of anomalies in the data can heavily degrade the performance of machine learning algorithms. In order to detect the anomalies in a dataset in an unsupervised manner, some novel statistical techniques are proposed in this paper. The proposed techniques are based on statistical methods considering data compactness and other properties. The newly proposed ideas are found efficient in terms of performance, ease of implementation, and computational complexity. Furthermore, two proposed techniques presented in this paper use transformation of data to a unidimensional distance space to detect the outliers, so irrespective of the data's high dimensions, the techniques remain computationally inexpensive and feasible. Comprehensive performance analysis of the proposed anomaly detection schemes is presented in the paper, and the newly proposed schemes are found better than the state-of-the-art methods when tested on several benchmark datasets.

Keywords: Anomaly/outliers detection, Advanced statistical methods, Computationally inexpensive methods, High dimensional data

Introduction

An observation in a dataset is considered an outlier if it differs significantly from the rest of the observations. The problem of finding patterns in data that deviate from the expected behavior is called the anomaly detection or the outliers' detection problem. Outliers in data can occur due to the variability in measurements, experimental errors, or noise [1], and the existence of outliers in data makes the analysis of data misleading and degrades the performance of machine learning algorithms [2, 3].

Several techniques have been developed in the past to detect outliers in data [4–6]. The techniques for outlier detection can be broadly classified as methods based on: (i) Clustering [7], (ii) Classification [8], (iii) Neighbor based [9], (iv) Statistical [10], (v) Information-Theoretic [11], and (vi) Spectral methods [12]. The working of classification-based methods mostly relies on a confidence score, which is calculated by the classifier while making a prediction for the test observation. If the score is not high enough, the observation is not assigned any label and is considered an outlier. Some clustering-based methods identify the outliers by not forcing every observation to belong to a

cluster, and the observations that are not assigned to any cluster are identified as outliers. The nearest neighbor techniques are mostly based on a calculation of the distance or similarity measure between the observation and its neighboring observations. Suppose the calculation is greater than a certain threshold, that means that the observation lies far apart from the rest of the observations and is considered as an outlier. Statistical methods usually fit a statistical distribution (mostly normal distribution) to the data and conduct a statistical inference test to see if the observation belongs to the same distribution or not. If not, the observation is marked as an outlier. Information-theoretic techniques use different information theoretic measures for example entropy, relative entropy, etc., to analyze the information content of the data. These techniques are based on an assumption that the outliers or anomalies in the data induce irregularities in the information content. Spectral methods transform the data to a new dimensional space such that the outliers are easily identified and separated from the data in the new space. Furthermore, some outlier detection techniques are also based on geometric methods [13] and neural networks [14].

All the techniques mentioned above are based on some assumptions and all the techniques have some pros and cons as described in Table 1. The ideas proposed in this work are based on the novel statistical methods considering data properties like compactness. Where, the compactness of data is estimated either by the interactions between different kernel function tails, new and adapted kernel functions are proposed, or by the variance of independent gaussian distributions over different regions that are captured as a new clustering idea. Moreover, contrary to the existing approaches, statistical methods are modeled based on the transformation of data into a unidimensional distance space. The newly proposed methods are based on boxplot adjustment, kernel based probability density estimation and neighborhood information. The aim here is to utilize the power of some statistical methods and to enhance the performance of the outlier detection algorithms in an unsupervised way, while keeping their implementation easy and computationally efficient. The proposed methods are evaluated using both the synthetic and the real datasets and are found better in scenarios where the traditional approaches fail to perform well. Especially the cases where the data is contaminated with a mixture of different noise distributions.

The rest of the paper is organized as follows: "Background" section, "Proposed methods" section, Evaluation on synthetic datasets "Evaluation using synthetic examples" section, Evaluation on real data "Evaluation of boxplot adjustments using a real example" section, "Comparison with State-of-Art" section and "Conclusions" section.

Background

After the initial pivotal works for outlier detection based on the distance measure [15, 16], several new methods based on the distance measure were also proposed by different authors in literature [17, 18]. The difference between the latterly proposed methods and the previous studies is the use of nearest neighbors in distance calculation. Among the variants of the actual work, either a single distance based on the *kth* closest neighbor is calculated [19] or the aggregate of the distance of k closets points is calculated [20]. Among other unsupervised outlier detection algorithms are the *local* approaches originated from the concept of Local Outlier Factor [21].

Table 1 Assumptions, advantages, and disadvantages of different outlier detection approaches

Method	Assumption(s)	Advantages	Disadvantages
Classification	A classifier with an ability to distinguish among the inlier and the outlier classes can be learnt utilizing the given feature space	1. Multi-class approaches can make use of powerful algorithms to discriminate between instances belonging to different classes 2. Since each test instance needs to be compared against the pre-computed model, therefore, the testing phase is faster	1. Dependent on the availability of accurate labels which is often not possible
Nearest neighbor	Inliers occur in the dense neighborhoods, while the outliers occur far from their closest neighbors	1. Unsupervised in nature and purely data driven 2. Their adaptability to a different data type is straightforward, and primarily requires a proper distance measure for the given data	1. The performance of these methods degrade if the assumption made is not true 2. The computational complexity of these algorithms is a significant challenge 3. Performance greatly relies on the distance measure used
Clustering	1. Outliers do not belong to any cluster while the inliers belong to a cluster in the data 2. Outliers are far away from their closest cluster centroid while inliers lie close to their closest cluster centroid 3. Outliers either belong to small or sparse clusters while the inliers belong to large and dense clusters	1. Ability to operate in an unsupervised mode 2. Ability to adapt to other complex data types by simply using a clustering algorithm that can handle the particular data type 3. Since the number of clusters against which every test instance needs to be compared is a small constant, therefore, the testing phase is fast	1. Highly dependent on the effectiveness of clustering algorithm in capturing the cluster structure of inlier data points 2. Many techniques detect anomalies as a by-product of clustering, and hence are not optimized for anomaly detection 3. Several clustering based techniques are effective only when the outliers do not form significant clusters among themselves
Statistical	Outliers occur in the low probability regions of the stochastic model, while the inliers occur in high probability regions of a stochastic model	1. Provide a statistically justifiable solution if the assumptions regarding the underlying data distribution hold true 2. The anomaly score is associated with a confidence interval which can be used as additional information while making a decision regarding any test instance 3. If the distribution estimation step is robust to outliers in data, these techniques can operate in a unsupervised setting	1. Rely on the assumption that the data is generated from a particular distribution which is often not true especially for high dimensional real data sets 2. Constructing hypothesis tests for complex distributions that are required to fit high dimensional data sets is nontrivial

Table 1 (continued)

Method	Assumption(s)	Advantages	Disadvantages
Information theoretic	Outliers in data induce irregularities in the information content of the data set	1. Ability to operate in an unsupervised setting 2. They do not make any assumptions about the underlying statistical distribution for the data	1. Sensitive towards the choice of the information theoretic measure. Often, these measures can only detect the presence of outliers when there are significantly large number of outliers present in the data 2. When applied to spatial and sequential data sets, these techniques rely on the size of substructure, which is often nontrivial to obtain 3. It is difficult to associate an outlier score with a test instance
Spectral	Data can be transformed into a lower dimensional subspace, where the inliers and the outliers appear significantly different	1. Automatically perform dimensionality reduction, therefore, these are suitable for high dimensional datasets 2. Can be used in an unsupervised setting	1. Useful only if the inliers and outliers are separable in the lower dimensional transformation of the data 2. Typically have high computational complexity

Furthermore, boxplot outlier detection scheme is also one of the fundamental unsupervised approach and the concept of univariate boxplot analysis was first proposed by Tukey et. al. [22]. In a univariate boxplot, there are five parameters specified as: (i) the upper extreme bound (UE), (ii) the lower extreme bound (LE), (iii) the upper quartile Q3 (75th percentile), (iv) the lower quartile Q1 (25th percentile) and (v) the median Q2 (50th percentile). The best way to estimate the extreme boundaries is to estimate Probability Density Function (PDF), $f(x)$, at first step from where the boundaries will be defined, as follows:

$$\begin{cases} UE : \frac{\tau}{2} = P(X > UE) = \int_{UE}^{+\infty} f(x)dx \\ LE : \frac{\tau}{2} = P(X < LE) = \int_{-\infty}^{LE} f(x)dx \end{cases}. \tag{1}$$

where τ is the significance level, the region of suspected outliers is defined for $\tau = 0.05$ and the region of extremely suspected outliers is defined for $\tau = 0.01$. The Eq. (1) estimates well the boundaries only if the distribution is unimodal, i.e., a distribution that has single peak or at most one frequent value.

However, in a standard boxplot the UE and LE values are computed and well estimated only under the assumption that the PDF is symmetric, as:

$$\begin{cases} LE = Q1 - 1.5(IQR), \\ UE = Q3 + 1.5(IQR). \end{cases} \tag{2}$$

where the term IQR is defined as the Inter Quartile Range and is given by:

$$IQR = Q3 - Q1. \tag{3}$$

A common practice to identify the outliers in a dataset using a boxplot is to mark the points that lie outside the extreme values, that is, the points greater than UE and less than LE are identified as outliers. This version of outlier detection scheme works well for the symmetric data. However, for skewed data different other schemes are proposed in the literature. For example, different authors have used the semi-interquartile range i.e. $Q3 - Q2$ and $Q2 - Q1$ to define the extreme values as:

$$\begin{cases} LE = Q1 - c_1(Q2 - Q1), \\ UE = Q3 + c_2(Q3 - Q2). \end{cases} \tag{4}$$

where c_1 and c_2 are the constants and different authors have adjusted their values differently for example, $c_1 = c_2 = 1.5$ [23] $c_1 = c_2 = 3$ [24] or calculation based on the expected values of the quartiles [25] and few more adjustments to the boxplot for outliers detection are also available, for example [26].

The traditional methods of boxplot for detecting the outliers sometimes fails in situations where the noise in the data is a mixture of distributions, multimodal distribution, or in the presence of small outlier clusters. In this paper, some novel statistical schemes based on (i) the boxplot adjustments, and (ii) a new probability density estimation using k-nearest neighbors' distance vector are proposed to overcome the problem faced by traditional methods. These proposed methods are described in detail in the next section.

Proposed methods

Boxplot adjustments using D-k-NN (BADk)

Traditional boxplot identifies the outliers from unique dimensions. One useful idea other than using all the dimensions for identifying outliers is to transform the data into a unidimensional distance space and to identify the outliers in new space. This can simply be done by measuring the distance between data points considering all the dimensions and calculating the resulting distance vector. The idea of using a single dimension distance vector is useful not only for avoiding problem of sorting data in high dimension but also in terms of computational cost, and can be further enhanced in terms of performance by extending it to consider k number of neighbors in the distance calculation. This idea of boxplot adjustment based on the Distance vector considering k number of Nearest Neighbors (D-k-NN) is presented here and the resulting extreme values estimation from the modified boxplot are found to be quite useful in identifying the right outliers. Furthermore, the proposed scheme is useful in the cases where the distribution of the noise is not normal or is a mixture of different distributions, and can identify small outlier clusters in the data.

Suppose a dataset in $\mathbb{R}^N$, this dataset is transformed from N dimensional space to a unidimensional *distance* space by using a distance metric such as '*Euclidian distance*'. This is done by computing the distance of each observation in N dimensional space to its *kth* closest neighbor. This transformation results in a set that contains the distance of each observation to its *kth* closest neighbor, and the resulting set is represented as $d_k \in \mathbb{R}$. This transformation can be represented as:

$$d_k : \mathbb{R}^N \to \mathbb{R} \tag{5}$$

The set d_k is used for computing the extreme value of the boxplot as follows:

$$\begin{cases} LE_{d_k} = Q1_{d_k} - c_1\left(Q2_{d_k} - Q1_{d_k}\right), \\ UE_{d_k} = Q3_{d_k} + c_2\left(Q3_{d_k} - Q2_{d_k}\right). \end{cases} \tag{6}$$

From the extreme values defined in (6), the outliers are identified as points those lie outside the boundaries of LE_{d_k} and UE_{d_k}. The two constant values c_1 and c_2 are adjustable with respect to the dataset under consideration and selection of smaller values for these constants results in more points being marked as outliers. The suggested values of these constants by different authors in the literature are $c_1 = c_2 = 1.5$ or $c_1 = c_2 = 3$.

Furthermore, another useful idea to identify the outliers in a data is to adjust the UE and LE values of a boxplot as follows:

$$\begin{cases} LE = Q1_{d_k} - c_1 \times \sqrt{var\left(X.1_{X<Q2_{d_k}}\right)}, \\ UE = Q3_{d_k} + c_2 \times \sqrt{var\left(X.1_{X\geq Q2_{d_k}}\right)}. \end{cases} \tag{7a}$$

or

$$\begin{cases} LE = Q1_{d_k} - c_1 \times \sqrt{var\left(X.1_{X<Q1_{d_k}}\right)}, \\ UE = Q3_{d_k} + c_2 \times \sqrt{var\left(X.1_{X\geq Q3_{d_k}}\right)}. \end{cases} \tag{7b}$$

where *var* is defined as the variance and the quartiles are computed from the set $d_k \in \mathbb{R}$. The extreme values can also be estimated based on the calculation of the separation threshold between centers of two variances, see Eq. (9), as:

$$\begin{cases} LE = M - c_1 \times var(X.1_{X<M}), \\ UE = M + c_2 \times var\big(X.1_{X\geq M}\big). \end{cases} \tag{8}$$

where M is a value that separate the one-dimension region in order to calculate the variance of two centers. Let $x \epsilon \mathbb{R}$ be any random variable with PDF $f(x)$; and the values of μ_1 and μ_2 are calculated such that:

$$\begin{cases} \big(\mu_1^*, \mu_2^*\big) = \underset{(\mu_1,\mu_2)}{\arg} \left[\min_{M,\mu_1,\mu_2} \left[\int_{-\infty}^{M} (x-\mu_1)^2 f(x)dx + \int_{M}^{\infty} (x-\mu_2)^2 f(x)dx\right]\right], \\ Var_1 = \int_{-\infty}^{M} (x-\mu_1)^2 f(x)dx, \\ Var_2 = \int_{M}^{\infty} (x-\mu_2)^2 f(x)dx. \end{cases} \tag{9}$$

Both, Var_1 and Var_2 can be partially differentiated with respect to μ_1 and μ_2 respectively, to find the minimum. After simplification the minimization occurs when:

$$\begin{cases} \mu_1 = \frac{E(X_-)}{P(X_-)}, \\ \mu_2 = \frac{E(X_+)}{P(X_+)}, \end{cases} \tag{10}$$

where $X_- = X.1_{X<M}$ and $X_+ = X.1_{X\geq M.}$ and the value of M is calculated as:

$$M = \left[\frac{E(X_-)}{P(X_-)} + \frac{E(X_+)}{P(X_+)}\right].\frac{1}{2} \tag{11}$$

For further details on the idea proposed in (8)–(11), the readers are referred to [27].

Detecting outliers based on Boxplot is efficient only if the data is unimodal distribution. To overcome the drawbacks of the boxplot estimation, some other statistical methods based on the probability density estimation computed from either the set $d_k \in \mathbb{R}$ or the actual data $D \in \mathbb{R}^N$ are also proposed for outlier's detection, which are discussed below.

Joint probability density estimation using D-k-NN

The methods proposed in this section compute the set d_k from the actual data and utilize it for estimating some parameters of the joint distribution function. Three different schemes are proposed here which are described as follows:

Scheme 1: Normal distributions are often used for representing the real value random variables with unknown distributions [28, 29]. The joint probability density function of independent and identically normal distribution is given as:

$$f(x_1, \ldots, x_N) = \frac{1}{\left(\zeta\sqrt{2\pi}\right)^N} e^{\sum_{i=1}^{N} -\frac{1}{2}\left(\frac{x_i-\mu_i}{\zeta}\right)^2} \tag{12}$$

where ζ is the standard deviation modeled differently in (15) and (17), μ is the mean of the random variable and N is the dimension of the data. Here, some functions based on the normal distribution to identify the outliers in a dataset are proposed. Suppose a two-dimensional dataset $D(x, y)$, we can define a separation threshold T based on the normal distribution for detecting the outliers such that:

$$\begin{cases} Z = \sum_{i=1}^{n} f(x_i, y_i), \\ T = \alpha max(Z). \end{cases} \tag{13}$$

where Z is joint probability distribution function after normalization, n is the total number of observations and the function $f(x, y)$ can be defined as:

$$f(x, y) = \frac{1}{2\pi\zeta^I} e^{-\left(\frac{(x-x_i)^2+(y-y_i)^2}{2\zeta^2}\right)}; i = 1, 2, \ldots, n.; I = 0, 1, 2. \tag{14}$$

The σ in Eq. (14) can be computed as:

$$\zeta = \beta Q3_{d_k}. \tag{15}$$

where $Q3_{d_k}$ is the third quartile computed from the set d_k as defined in Eq. (5) and β is a constant value. The points below the threshold value T defined in Eq. (13) are considered as outliers and the points above T are considered normal inlier data points. The α used in Eq. (13) is the significance value and it can be used to control the percentage amount of data to be removed as outliers.

Scheme 2: To better detect the outliers, a better function $f(x, y)$ needs to be constructed in order to weaken the position of the outlier in terms of support and amplitude of the function. Furthermore, another scenario can be defined to detect the outliers based on the threshold defined in (13) by using the below function:

$$f(x, y) = \frac{\zeta^2}{\pi} e^{-\left(\frac{(x-x_i)^2+(y-y_i)^2}{\zeta^2}\right)}; i = 1, 2, \ldots, N. \tag{16}$$

$$\zeta = \frac{\gamma}{(1 + d_k)^2} \tag{17}$$

where k defines the *kth* closest neighbor for the distance metric and γ is a constant whose value can be adjusted to control the smoothness of the gaussian distribution. The concept is demonstrated in Fig. 1, where (a) shows the effect of traditional gaussian approach on compression and (b) shows the effect of proposed scheme 2 on compression.

Both of the above schemes proposed in this section are based on a single gaussian distribution and are expected to work well for the datasets which can be well approximated using a single gaussian distribution. However, if a dataset can be better approximated using multiple gaussians then a better idea is to use a model based on the variable number of gaussians. A new and robust estimation of multiple gaussian distribution is proposed in the next subsection.

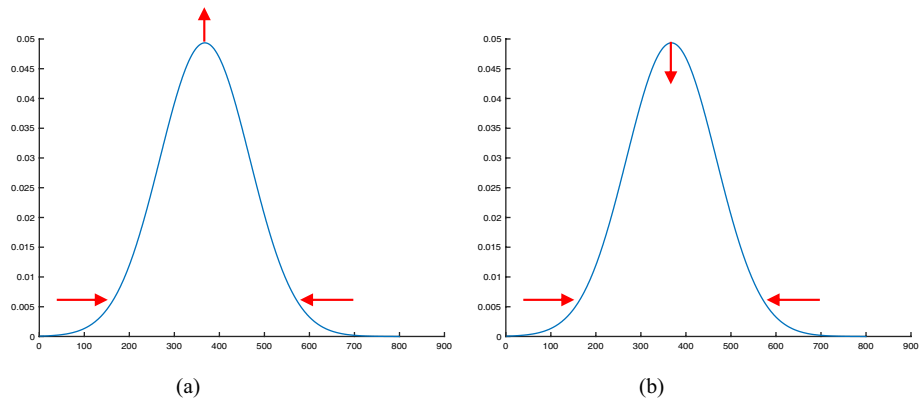

Fig. 1 **a** Effect of traditional gaussian approach on compression. **b** Effect of proposed scheme 2 approach on compression in x and y axes

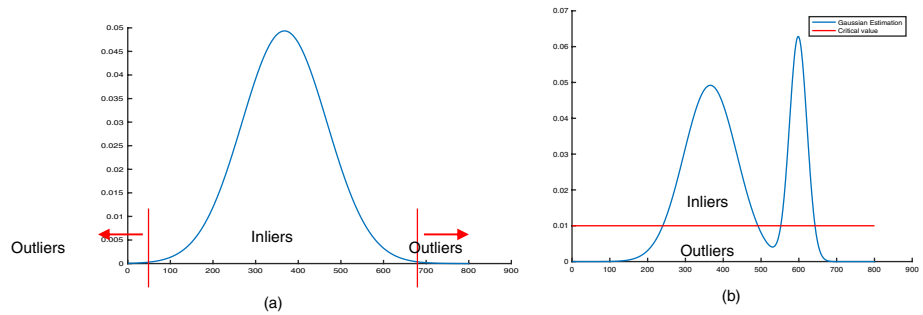

Fig. 2 An Example of Gaussian estimation and marking of critical value for outlier detection. **a** Points those lie outside the red boundaries are considered outliers. **b** The points below the critical value are identified as outliers

Scheme 3: The scenarios where the data is estimated using a gaussian distribution, the outliers are identified as the points lying on the extreme tails of the gaussian distribution, as shown in Fig. 2a. However, if the better estimation of underlying data is possible through multiple gaussians, the outliers located at the connecting points of different gaussians might remain unidentified using a single gaussian estimation. In order to identify the outliers existing at the connecting points of the multiple gaussians, an idea based on multiple gaussian estimation is proposed, where a Rejection Area (RA) is defined and computed as:

$$\begin{cases} RA = \{\vec{x} : f(\vec{x}) \leq Cv\}, \\ \quad f(x \in RA) = \tau. \end{cases} \tag{18}$$

where Cv is defined as a critical value or a threshold value below which is the rejection area or where the outliers are identified, and τ is the significance level. The concept is shown in Fig. 2b, where as an example a single dimensional data is estimated using two gaussians and the outliers can be identified as the points below Cv.

In order to find the optimum number of gaussians that better approximate the joint probability distribution for a given dataset the sorted values of the vector d_k can be

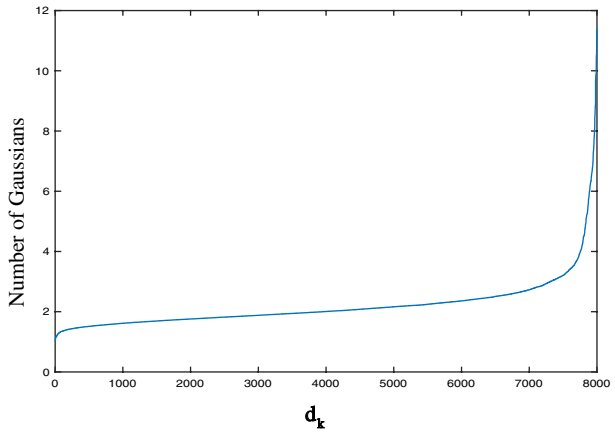

Fig. 3 Plot of sorted values of the vector $\boldsymbol{d_k}$

utilized. For example, in Fig. 3 the graph of sorted values of the vector d_k is shown and the best value of number of gaussians can be estimated by taking the value where the graph takes off sharply.

Each estimated gaussian represent a region and for each gaussian inside a region the values of mean and variance can be computed as:

$$\mu_i = \frac{\sum_{i \in R_j} x_i}{n_i}, \mathrm{j} = 1, 2, \ldots., \mathrm{m} \tag{19}$$

$$Var(x) = \frac{\sum_{i \in R_j} (x_i - \mu_i)^2}{n_i - 1}, \mathrm{j} = 1, 2, ..., \mathrm{m} \tag{20}$$

where R_j represents the jth region, m is the total number of estimated gaussians and n_i is the total number of elements in the respective region. The combined multiple gaussians model is then estimated by:

$$x \sim \sum_{i=1}^{m} \alpha_i N(\mu_i, C_i) \tag{21}$$

where

$$\alpha_i = \frac{Card(R_i)}{N} \text{ and } \sum \alpha_i = 1. \tag{22}$$

In order to determine the regions, lets define an application S_U that sorts any given sequence $U_i, i = 1, \ldots, N$ such that $U_{S_U(1)} \leq U_{S_U(2)} \leq \cdots \leq U_{S_U(N)}$. For any given data $\overrightarrow{X}$, the sorted data can be represented as $\overrightarrow{X}_{S_U(i)}$ and suppose that $\overrightarrow{\Delta X}_{S_U(i)}$ represents the difference between two consecutive elements of $\overrightarrow{X}_{S_U(i)}$. Similarly, the sorted difference can be represented as $\overrightarrow{\Delta X}_{S_{\Delta X}(S_U(i))}$. In order to define the regions, the elements are grouped together sequentially until$\Delta X_{S_{\Delta X}(S_U(i))} \leq \overrightarrow{\Delta X}_{S_{\Delta X}(S_U(N-m+1))}$, once this condition is not true, start grouping the remaining elements as a new region until all the elements are assigned to a region.

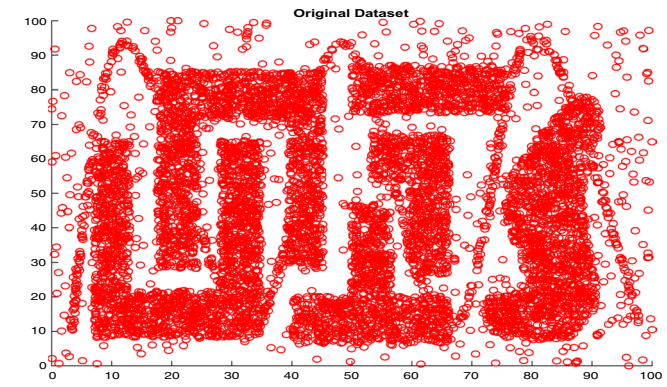

Fig. 4 **a** Original data representing clusters with different shapes and a mixture of sinusoidal and gaussian noise. **b** Traditional boxplot showing no outliers/anomalies

Evaluation using synthetic examples

The ability of proposed methods is demonstrated here by the use of some two-dimensional synthetic datasets. The results for each of the proposed method are discussed in the following subsections.

Evaluation of boxplot adjustments using D-k-NN (BADk)

Figure 4a shows an example of the data used for evaluating the proposed methods to detect the outliers. From the data shown in Fig. 4a, it can be seen that the actual data is composed of different clusters with different shapes and is contaminated by a mixture of sinusoidal and gaussian distribution of noise. The aim here is to detect the noise as outliers and different shape clusters as inliers. The traditional boxplot is used to detect the outliers from the data and the resulting boxplot is shown in Fig. 4b. It can be seen from the boxplot in Fig. 4b that the traditional boxplot is unable to identify any outliers in the data.

The same dataset is used to evaluate the proposed adjusted boxplot with extreme values define in Eq. (6) and the results are shown in Fig. 5. Different values of k are used to see how it effects the outcome in identifying the outliers. It can be observed from the results shown in Fig. 5 that for the smaller values of k only the gaussian

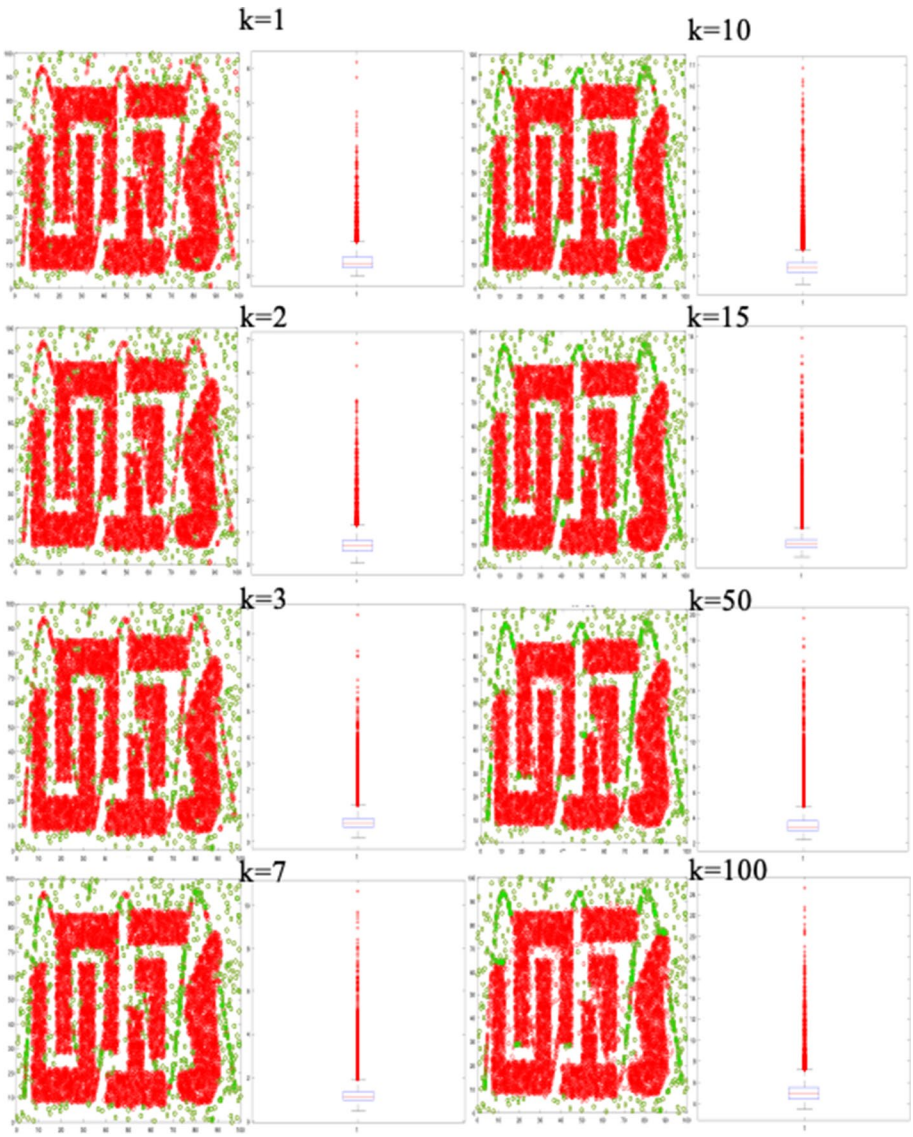

Fig. 5 Outliers detection from the data shown in Fig. 4a, using the proposed boxplot with extreme values defined in Eq. (6) for different values of k. The data in red is identified as the inliers while the data in green is identified as the outliers

noise is identified and while we keep on increasing the value of k the outliers with sinusoidal distribution are also identified.

However, after a certain value of k the data points from the actual clusters (inliers) are also marked as the outliers, while the outliers started to reappear as the inliers. This shows that although the selection of value of k is flexible in this case, still an optimum value of k has to be selected for the optimum performance based on the data. Another example shown in Fig. 6a with a different distribution of noise is also tested for evaluating the ability of the proposed method in (6) for outlier detection. The results for this example are shown in Fig. 7 using three different values of k. It can be seen from the results in Fig. 7 that the selection of value of k is very flexible and

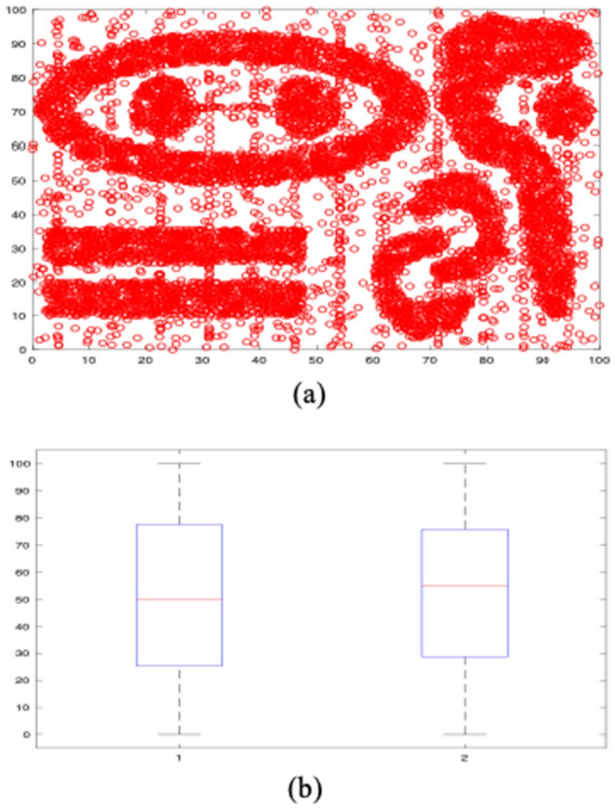

Fig. 6 **a** An example of dataset having different shapes of inlier clusters contaminated with noise. **b** Traditional boxplot showing no outliers/ anomalies

still the proposed method performs well in terms of outlier's detection. During all the experiments performed, the values of constants are fixed to $c_1 = c_2 = 1.5$.

Evaluation (joint probability density estimation) scheme 1

The density estimation refers to estimation of an unobservable Probability Density Function (PDF) associated with an observable data. The PDF gives an estimate of the density according to which a large population in a data is distributed. In this proposed method, the PDF is computed by placing a gaussian density function at each data point, and then summing the density functions over the range of data, and a threshold value α defines the margin between the inlier data and the outliers. The value of α is computed as a percentage amount of the maximum value of the PDF. The value of σ in Eq. (14) is computed utilizing the d_k vector as defined in Eq. (15).

The results for scheme 1 when evaluated using the same example data as shown in Fig. 4a are given in Fig. 8. The example is evaluated using only two different values of α and a fixed value of β. The outliers are shown in green color and the inlier data is shown in red color. Figure 8 also shows the associated 3D plots of the probability density estimations computed using Eq. (14). For $\alpha = 0.1$ the proposed method is able to identify the outliers having gaussian distribution only while placing $\alpha = 0.3$ the proposed method has identified both the gaussian and the sinusoidal outliers in the data. Figure 9 shows the

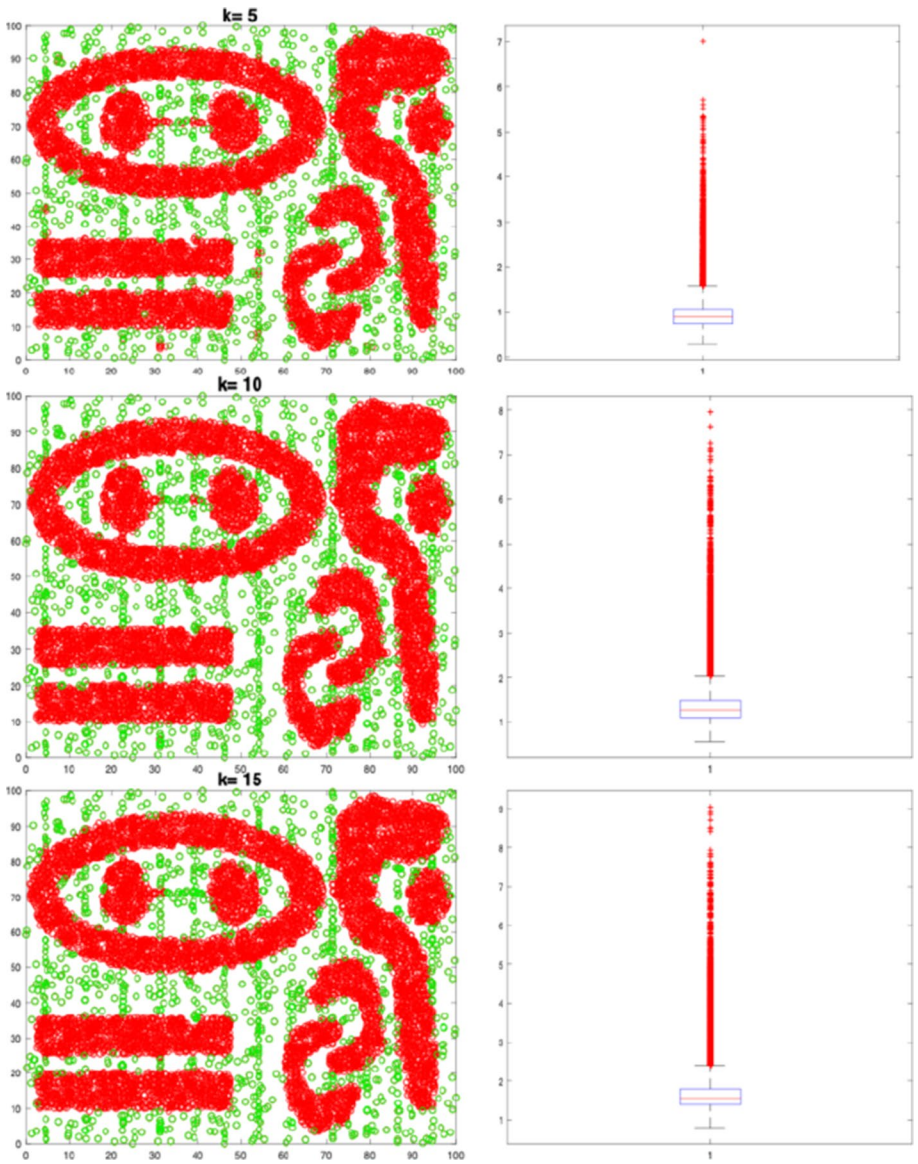

Fig. 7 Outliers detection from the data shown in Fig. 6a, using the proposed boxplot with extreme values defined in Eq. (6) for different values of k. The data in red is identified as the inliers while the data in green is identified as the outliers

results for the second example with a different noise distribution with fixed values of α and β using Eq. (14).

Evaluation (joint probability density estimation) scheme 2

The results for scheme 2 proposed in Eqs. (16)–(17) with different value of parameters are shown in Fig. 10. It can be observed from the results in Fig. 10 that the small value of γ produces sharp density distribution while a larger value of γ produced a smoother distribution. For a small value of γ the inlier data points are also identified as the outliers which is not the case with a comparatively larger value of γ for this particular dataset.

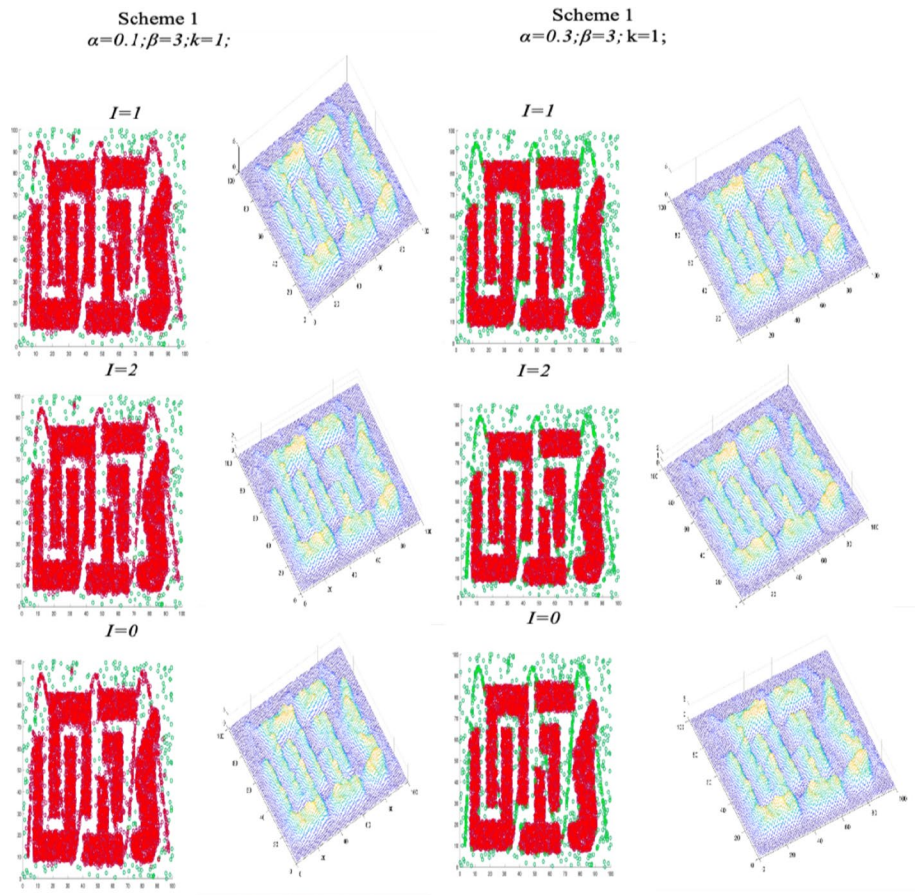

Fig. 8 Outlier Detection results using scheme 1 with two different values of α = 0.1 and α = 0.3. Outliers are shown in green and the inlier data in red

However, optimum values for the parameters need to be tuned to get the optimum results using this scheme.

Evaluation (joint probability density estimation) scheme 3

The idea proposed in Eq. (18) based on the different ways of estimation of gaussians of the distance vector d_k is evaluated on three different synthetic examples having different distribution of noise and the results are shown in Fig. 11. It can be seen from the visual results depicted in Fig. 11 that this scheme is successful in identifying the outliers of different distributions and even the noisy data that lies in close proximity to the inlier data. The value of G represents the number of gaussians estimated from the d_k.

Evaluation of boxplot adjustments using a real example

The ideas proposed for boxplot adjustments in Eq. (6) and Eq. (7) are also evaluated on a real dataset. The dataset used is a subset of the original KDD Cup 1999 dataset from the UCI machine learning repository, the subset used is still a large data containing 95,156

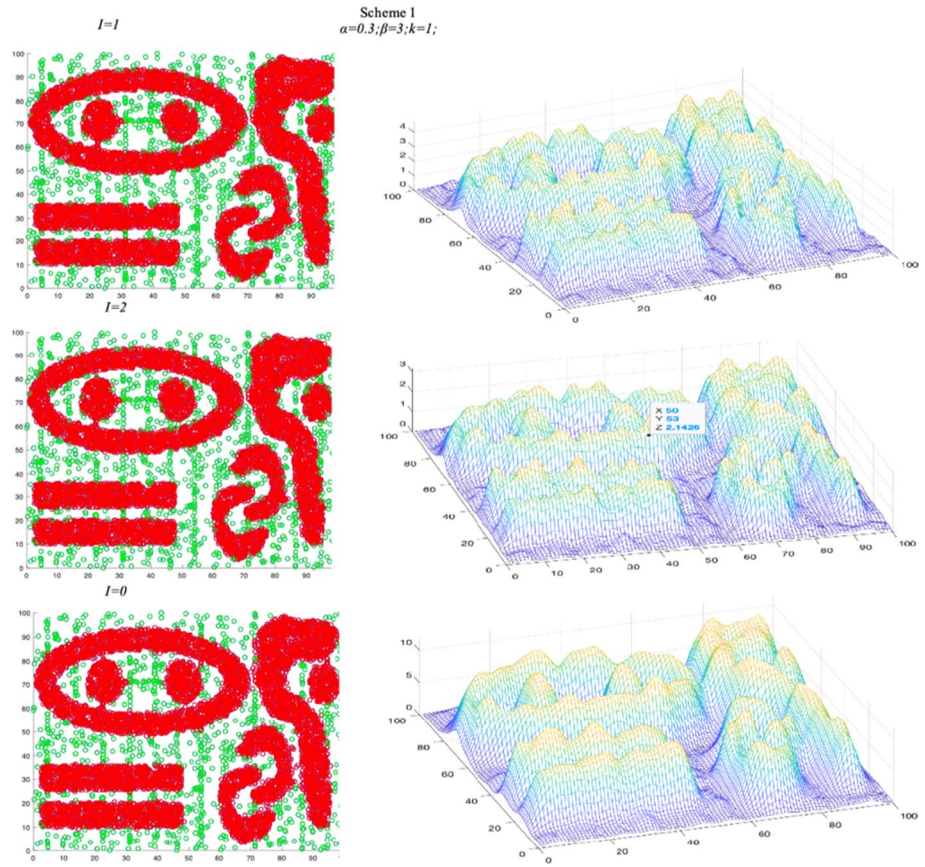

Fig. 9 Second example of Outlier Detection results using scheme 1 with α = 0.3, β = 3 and k = 1. Outliers are shown in green and the inlier data in red

observations and three attributes. The dataset is publicly available online.[1] The ground truth of the dataset used is shown in Fig. 12a, where the blue data points represent the actual inliers and the yellow points represent the actual outliers. The results for boxplot using extreme values defined in Eq. (6) are shown in Fig. 12b and the achieved value for Area Under Curve (AUC) evaluation parameter is 0.83 for this dataset.

The results achieved for the proposed idea in (7) are shown in Fig. 13a, b, respectively for Eqs. 7a and 7b. The detected outliers are shown in green color and the inliers are shown in red color. The achieved value of AUC using both Eq. 7a and 7b is 0.833.

Comparison with State-of-Art

The proposed schemes are compared with several state of the art unsupervised outlier detection algorithms of similar kind, using a variety of benchmark datasets reported in [30]. The details of the benchmark datasets used for comparison are given in Table 2. The algorithms used for comparison include kNN [19], kNN-weight (kNNW) [20, 31], Outlier detection using Indegree Number (ODIN) [32], Local Outlier Factor (LOF) [21],

[1] http://odds.cs.stonybrook.edu/smtp-kddcup99-dataset/.

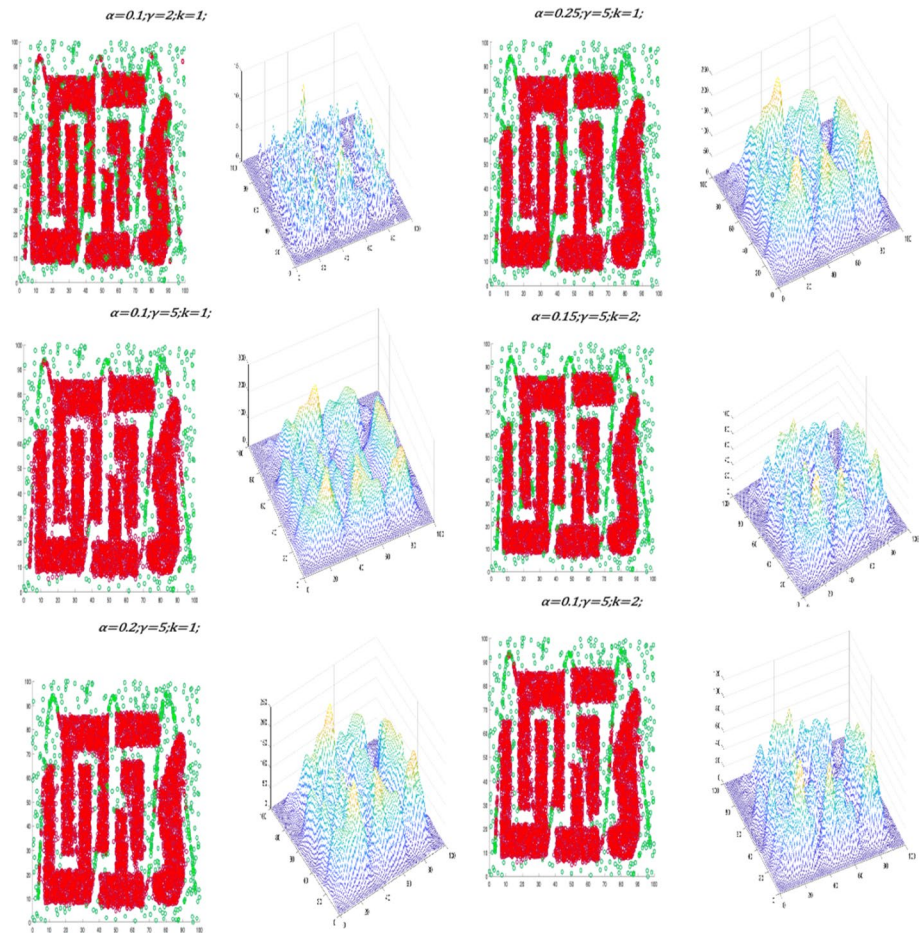

Fig. 10 Outlier detection results using scheme 2 with different values of α, γ and k. Green points represent outliers and red points represent the normal data points

Simplified LOF (SLOF) [33], Connectivity based Outlier Factor (COF) [34], Influenced Outlierness (INFLO) [35], Local Outlier Probabilities (LoOP) [36], Local Distance-based Outlier Factor (LDOF) [37], Local Density Factor (LDF) [38], Kernel Density Estimation Outlier Score (KDEOS) [39], Multi-Objective Generative Adversarial Active Learning (MO-GAAL) [40], Single-Objective Generative Adversarial Active Learning (SO-GAAL) [40], Artificially Generating Potential Outliers (AGPO), Active-Outlier method (AO) [41], Gaussian mixture model (GMM) [42], Parzen [43], One-Class Support Vector Machine (OC_SVM) [44], and Fast Angle-Based Outlier Detection (FastABOD) [45].

Initially, some of the fundamental outlier detection algorithms are compared with the proposed algorithms using the same three synthetic datasets. To test the unsupervised outlier detection methods, the most popular evaluation measure proposed in literature is based on the Receiver Operating Characteristics (ROC) and is computed as the Area Under Curve (AUC) [30]. The ROC AUC is computed for these three datasets using the proposed schemes and are compared with some of the fundamental state-of-art algorithms in Table 3. Hyperparameters of all the methods are tuned and the best results are reported. It can be seen from the results in Table 3

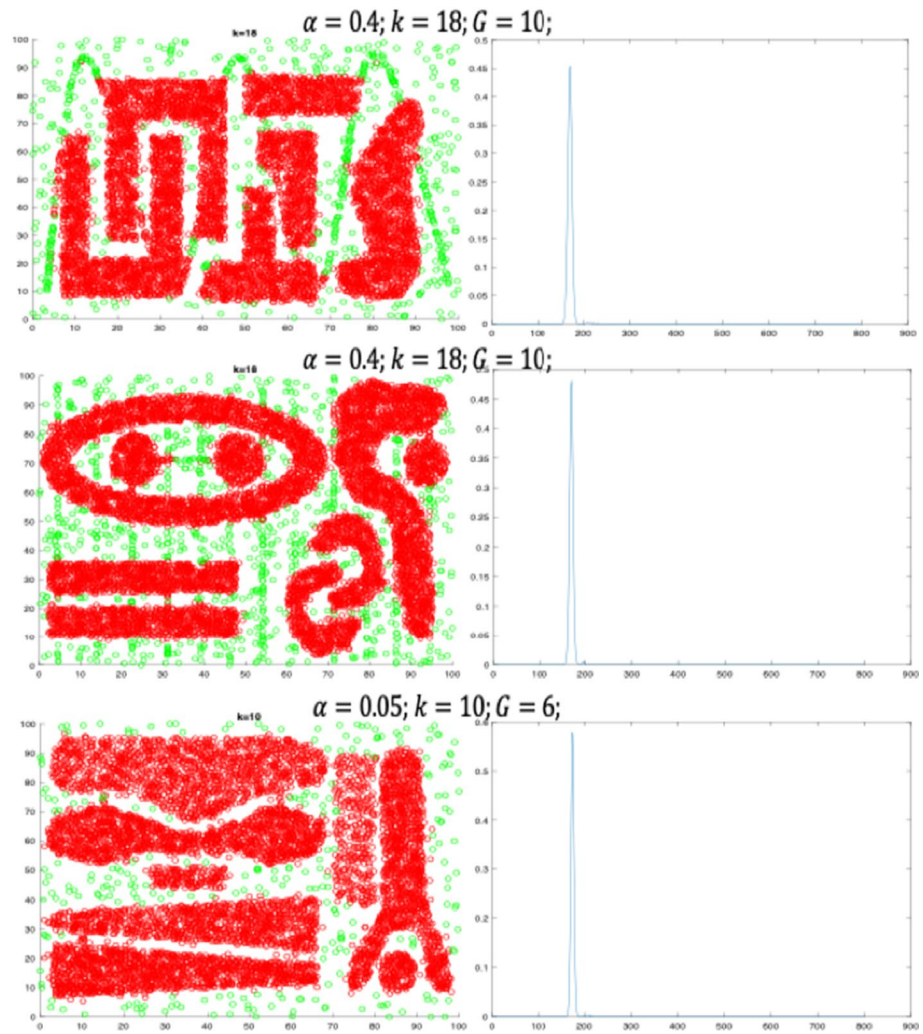

Fig. 11 Results achieved on three different datasets with different distribution of noise using the proposed GMM estimation of d_k. The data points in red are the inliers and the green data points are identified as the outliers

that the proposed schemes are performing better than the existing algorithms. As these three datasets are only two dimensional the visual comparison is also possible which is provided in Fig. 14. From the visual inspection it is clearer that the newly proposed methods are better than the existing ones in identifying the outliers lying in close proximity to the inliers. Although, all the proposed schemes out-performed the existing approaches when tested on two dimensional synthetic datasets, only BADk and scheme 3 are recommended for large scale datasets. This is because the computational complexity of scheme 1 and scheme 2 is relatively higher than the BADk and scheme 3, as given in Table 4. Therefore, we recommend scheme 1 and scheme 2 only for dataset with dimensions less than or equal to 3 and for small datasets. For large scale datasets we recommend using scheme 3 and BADk. However, with a compromise on the computational complexity, scheme 1 and scheme 2 have the ability to

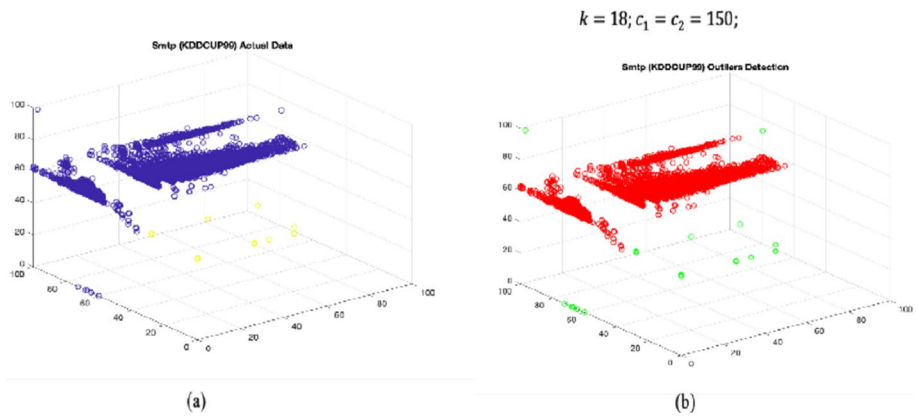

Fig. 12 **a** Ground truth of the real data used to evaluate the proposed methods, blue data points are inliers and yellow are the outliers. **b** Results achieved using the boxplot with extreme values proposed in Eq. 6. Red points are inliers and green are outliers

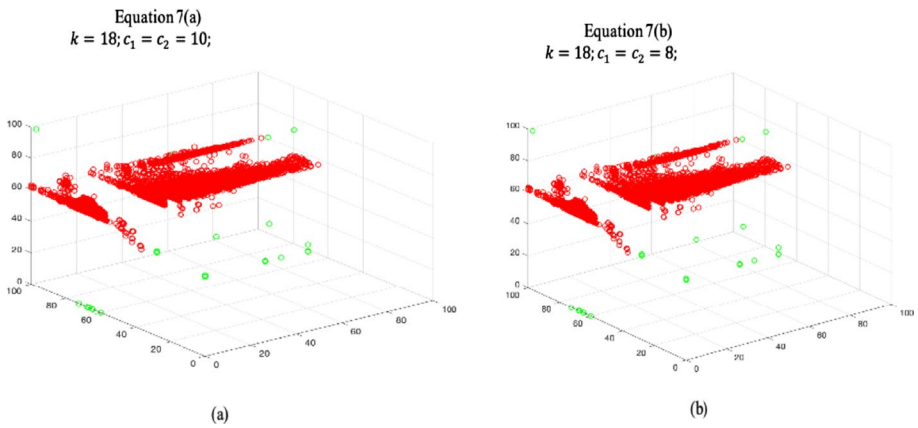

Fig. 13 **a** Results achieved using the boxplot with extreme values proposed in Eq. 7(a). Red points are inliers and green are outliers. **b** Ress achieved using the boxplot with extreme values proposed in Eq. 7(b). Red points are inliers and green are outliers

perform better in terms of ROC AUC on individual datasets. Although, the best running time complexity is achieved by LOF but the proposed methods are performing much better than LOF in terms of ROC AUC values on individual datasets.

Furthermore, the proposed methods are evaluated and compared with eight existing approaches using five real benchmark datasets and the results are reported in Table 5. From the results in Table 5 it can be seen that the proposed schemes outperformed the existing approaches when tested on five real benchmark datasets. As, the proposed BADk and scheme 3 performed better than the proposed scheme 1 and scheme 2 when tested on synthetic datasets in terms of average ROC AUC value and computational time, therefore only these two methods are included for comparison with existing approaches using the real datasets.

In order to perform more comprehensive comparison, ten more benchmark datasets and twelve state-of-art methods reported in [30] are also used and are compared with

Table 2 Details of benchmark datasets used for evaluation and comparison with State-of-art

Dataset	Type	Description	# observations	# dimensions	# Outliers
T48K*	Synthetic	Six multi-shape clusters with two types of noise	8000	2	764
Complex9*	Synthetic	Nine multi-shape clusters with noise	10,000	2	792
Cluto*	Synthetic	Eight multi-shape multi-density clusters with noise	8000	2	323
Arrhythmia**	Real	Patient records: normal vs cardiac arrhythmia	450	259	206
Heartdisease**	Real	Medical data on heart problems: healthy vs sick	270	13	120
Hepatits**	Real	Medical data on hepatitis: patient will die (outliers), survive (inliers)	80	19	13
Parkinson**	Real	Medical data: healthy people vs Parkinson's disease	195	22	147
Spambase40**	Real	Emails classified as spam (outliers) or non-spam	4207	57	1679
Glass**	Real	A forensic dataset describing types of glass	214	7	9
Pendigits**	Real	Different handwriting digits from 0 to 9	9868	16	20
Shuttle**	Real	Space Shuttle Data	1013	9	13
WBC**	Real	Cancer types, benign or malignant	454	9	10
WPBC**	Real	Wisconsin Prognostic Breast Cancer dataset	198	33	47
Pima**	Real	Medical data on diabetes	768	8	268

* Available at: https://github.com/deric/clustering-benchmark

** Available at: https://www.dbs.ifi.lmu.de/research/outlier-evaluation/DAMI/

Table 3 ROC AUC comparison using three synthetic datasets

Dataset	State-of -the-Art					Proposed			
	KNN	ABOD	FastABOD	COF	LOF	BADk	Scheme 1	Scheme 2	Scheme 3
T48K	0.6222	0.7692	0.7431	0.9081	0.9240	0.9466	**0.9655**	0.9524	0.9574
Complex9	0.5527	0.8526	0.8368	0.9047	0.9527	0.9776	**0.9816**	0.9779	0.9799
Cluto	0.7281	0.8903	0.8951	0.8839	0.9435	0.9514	0.9025	0.9030	**0.9520**
Average	0.6343	0.8373	0.8250	0.8989	0.9400	0.9585	0.9498	0.9444	**0.9631**

Bold values indicate are best-achieved

newly proposed unsupervised outlier detection methods. The results for these state-of-art methods and the newly proposed methods for the ten benchmark datasets are given in Table 6. It can be seen from the results in Table 6 that the newly proposed methods are clearly outperforming the existing algorithms in most of the cases. However, for two of the datasets (Hepatits and Parkinson) the proposed scheme 3 is performing equally well as compared to the LDF and both these methods are performing best as compared to the existing state-of-art approaches. For the Shuttle and the WPBC datasets the LDF approach is outperforming all other methods. However, the difference between the performance of LDF with proposed schemes on these two datasets is marginal, especially on WPBC. Furthermore, the two newly proposed methods BADk and scheme 3 make use of the distance vector for detection of outliers, so irrespective of the increasing

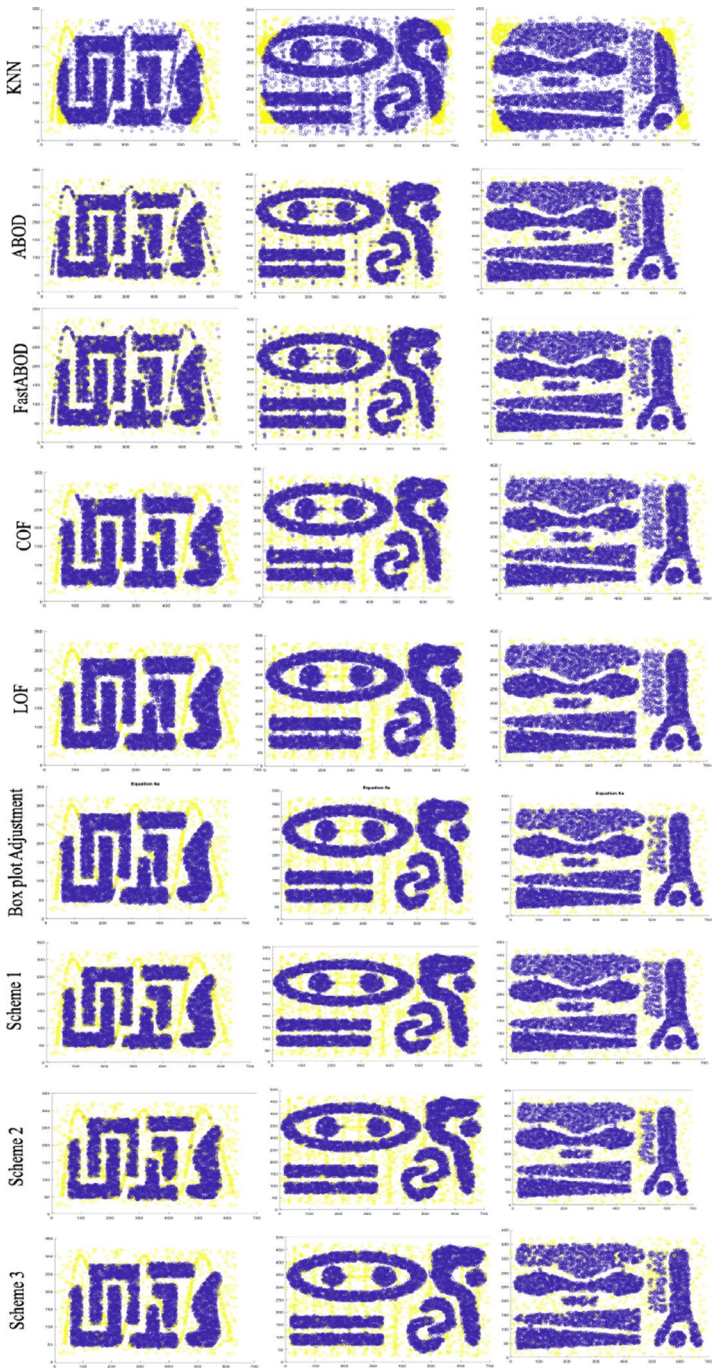

Fig. 14 Visual comparison for outlier detection using three synthetic datasets. Row 1–5: state-of-art methods and Row 6–9: the newly proposed methods

dimensions of the input data the computational complexity of these proposed algorithms remains low.

A visual comparison of proposed methods with the existing state-of-art methods is also provided in Fig. 15. From the results in Fig. 15, it is clearer that the newly

Table 4 Comparison based on computational time using three synthetic datasets

	Computational time in seconds								
Dataset	**KNN**	**ABOD**	**FastABOD**	**COF**	**LOF**	**BADk**	**Scheme 1**	**Scheme 2**	**Scheme 3**
T48K	1.000425	≈57,600	2.122058	1.620226	**0.178415**	1.149709	300.935130	310.405923	1.103109
Complex9	1.475390	≈ 650 00	3.040610	1.910642	**0.194985**	1.814219	495.625886	496.596672	1.689002
Cluto	4.135310	≈57,600	2.186437	1.611008	**0.163346**	1.156311	314.445905	315.213935	1.108947
Average	2.203708	≈60,067	2.449701	1.713958	**0.178915**	1.373413	370.335640	374.072177	1.300352

Bold values indicate are best-achieved

proposed scheme 3 is outperforming rest of the methods in terms of AUC. Furthermore, the required computational cost for the proposed method is also low because of using the d_k vector for outlier detection, instead of using the entire input data dimensions. As the proposed method is using only a single dimension distance vector of outlier detection, this makes it independent of the dimensions of the input data in terms of computational cost, which in turn makes it more feasible for high dimensional data.

Conclusions

Outlier detection is one of the most important preprocessing steps in data analytics, and for best performance consideration, it is considered a vital step for machine learning algorithms. Different methods are presented in this paper, keeping in view the need for a robust and easy-to-implement outlier detection algorithm. The newly proposed methods are based on novel statistical techniques considering data compactness, which resulted in an added advantage of easy implementation, improved accuracy, and low computational cost. Furthermore, to demonstrate the proposed ideas' performance, several benchmark multidimensional datasets and three complex synthetic two-dimensional datasets containing the different shapes of clusters contaminated with a mixture of varying noise distributions are used. The proposed methods are found accurate and better in terms of outlier detection as compared to the state-of-art. It is also an observation that some of the fundamental state-of-art methods cannot detect the outliers in scenarios where the outliers are a mixture of two different distributions. Moreover, two of the newly proposed schemes use only a single dimension distance-vector instead of utilizing the entire data dimensions for outlier detection. This makes the proposed methods more feasible and computationally inexpensive, irrespective of the input data's large sizes and growing dimensions.

Moreover, the evaluation of proposed unsupervised outlier detection methods on several benchmark real datasets reveal the usefulness of the proposed methods in detection of multivariate outliers in real datasets. The work can be extending by performing optimization on distance calculation method for the proposed scheme 3 and BADk. This will further enhance the computational complexity of these methods. Moreover, investigation of other distance metrics other than Euclidian can also be studied in future, as this metric suffers a lot for high dimensions.

Table 5 Comparison with existing approaches using some real datasets

Dataset	MO-GAAL	SO-GAAL	AGPO	AO	GMM	Parzen	OC_SVM	*k*-means	BADk	Scheme 3
Shuttle	0.907	0.902	0.273	0.701	0.964	0.970	0.672	0.969	0.9865	**0.9895**
PenDigits	0.976	0.934	0.810	0.768	0.808	0.969	0.365	0.977	**0.9826**	0.9819
Arrhythmia	0.751	0.729	0.743	0.636	0.473	0.751	0.707	0.746	0.8074	**0.8770**
Pima	0.758	0.669	0.588	0.575	0.674	0.729	0.569	0.681	0.6950	**0.7680**
SpamBase	0.627	0.380	0.616	0.599	0.549	0.599	0.590	0.578	**0.7143**	0.6472

Bold values indicate are best-achieved

Table 6 ROC AUC values computed on different benchmark datasets using State-Of-Art algorithms and the proposed schemes

Dataset	State-of-art												Proposed	
	KNN	KNNW	LOF	SLOF	LoOP	LDOF	ODIN	FABOD	KDEOS	LDF	INFLO	COF	BADk	Scheme 3
Arrhythmia	0.7930	0.7674	0.7674	0.7500	0.7500	0.7551	0.7663	0.7715	0.6680	0.8565	0.7950	0.7663	0.8074	**0.8770**
Heartdisease	0.8644	0.8311	0.8533	0.7800	0.7977	0.7577	0.8544	0.8911	0.6844	0.8933	0.8333	0.8955	0.8467	**0.9400**
Hepatits	0.8955	0.8706	0.9452	0.8806	0.8905	0.8706	0.9228	0.8408	0.8706	**0.9403**	0.9005	0.8955	0.8393	**0.9403**
Parkinson	0.9895	0.9895	0.9895	0.9895	0.9895	0.9895	0.9895	0.9895	0.9895	**1**	0.9895	0.9791	0.9583	**1**
Spambase40	0.5734	0.5661	0.4738	0.5011	0.4965	0.4796	0.5191	0.4372	0.4766	0.5364	0.4738	0.4994	0.6034	**0.6125**
Glass	0.8748	0.8832	0.8666	0.8650	0.8395	0.7788	0.7292	0.8579	0.7420	0.9035	0.8037	0.8953	0.9122	**0.9293**
Pendigit	**0.9868**	0.9854	0.9168	0.9107	0.9039	0.7181	0.9225	0.9610	0.8113	0.9545	0.8908	0.9609	0.9826	0.9819
Shuttle	0.9890	0.9861	0.9896	0.9869	0.9869	0.9775	0.9888	0.8381	0.9810	**0.9922**	0.9863	0.9920	0.9865	0.9895
WBC	0.9929	0.9920	0.9906	0.9835	0.9737	0.9582	0.9563	0.9892	0.6023	0.9929	0.9821	0.9863	0.9842	**0.9989**
WPBC	0.5409	0.5319	0.5254	0.5018	0.5018	0.5034	0.5072	0.5341	0.5185	**0.5829**	0.4957	0.5568	0.5427	0.5801

Bold values indicate are best-achieved

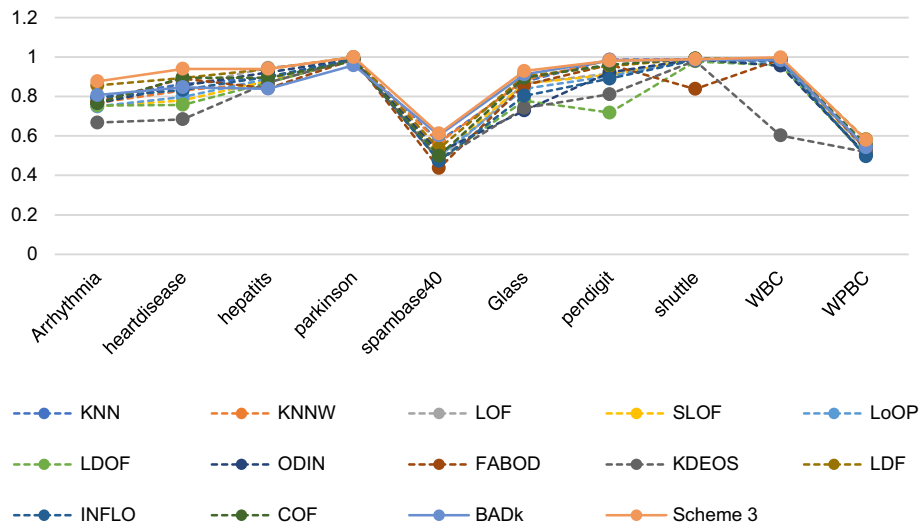

Fig. 15 Comparison of the proposed schemes with the state-of-art methods using 10 benchmark datasets for outlier detection. Y-axis represents the computed ROC AUC values

Abbreviations

ABOD: Angle-Based Outlier Detection; AUC: Area Under Curve; BADk: Boxplot adjustments using D-k-NN; COF: Connectivity based Outlier Factor; D-k-NN: Distance vector considering k number of Nearest Neighbors; INFLO: Influenced Outlierness; KDEOS: Kernel Density Estimation Outlier Score; LDF: Local Density Factor; LDOF: Local Distance-based Outlier Factor; LE: Lower extreme bound; LOF: Local Outlier Factor; LoOP: Local Outlier Probabilities; ODIN: Outlier Detection using Indegree Number; PDF: Probability Density Function; ROC: Receiver Operating Characteristics; SLOF: Simplified LOF; UE: Upper extreme bound.

Acknowledgements

The publication of this article was funded by the Qatar National Library. The authors would like to thank Qatar National Library (QNL) for supporting the publication charges of this article.

Authors' contributions

Both the authors have equally contributed in this work. All authors read and approved the final manuscript.

Authors' informations

Atiq Ur Rehman received the master's degree in computer engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2013 and PhD degree in Computer Science and Engineering from Hamad Bin Khalifa University, Qatar in 2019. He is currently working as a Post doc researcher with the College of Science and Engineering, Hamad Bin Khalifa University, Qatar. His research interests include the development of pattern recognition and machine learning algorithms.

Samir Brahim Belhaouri received the master's degree in telecommunications from the National Polytechnic Institute (ENSEEIHT) of Toulouse, France, in 2000, and the Ph.D. degree in Applied Mathematics from the Federal Polytechnic School of Lausanne (EPFL), in 2006. He is currently an associate professor in the Division of Information and Communication Technologies, College of Science and Engineering, HBKU. He also holds and leads several academic and administrator positions, Vice Dean for Academic & Student Affairs at College of Science and General Studies and University Preparatory Program at ALFAISAL university (KSA), University of Sharjah (UAE), Innopolis University (Russia), Petronas University (Malaysia), and EPFL Federal Swiss school (Switzerland). His main research interests include Stochastic Processes, Machine Learning, and Number Theory. He is now working actively on developing algorithms in machine learning applied to visual surveillance and biomedical data, with the support of several international fund for research in Russia, Malaysia, and in GCC.

Funding

Open access funding provided by the Qatar National Library.

Availability of data and materials

The datasets analysed during the current study are publicly available at http://odds.cs.stonybrook.edu/smtp-kddcup99-dataset/; https://www.dbs.ifi.lmu.de/research/outlier-evaluation/DAMI/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Authors declare no competing interest.

Received: 17 February 2021 Accepted: 15 May 2021
Published online: 02 June 2021

References

1. Zhu J, Ge Z, Song Z, Gao F. Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data. Annu Rev Control. 2018;46:107–33.
2. McClelland GH. Nasty data: unruly, ill-mannered observations can ruin your analysis. In: Handbook of research methods in social and personality psychology. Cambridge: Cambridge University Press; 2000.
3. Frénay B, Verleysen M. Reinforced extreme learning machines for fast robust regression in the presence of outliers. IEEE Trans Cybern. 2015;46(12):3351–63.
4. Wang X, Wang X, Wilkes M, Wang X, Wang X, Wilkes M. Developments in unsupervised outlier detection research. In: New Developments unsupervised outlier detection. Springer: Singapore; 2021. p. 13–36.
5. Zimek A, Filzmoser P. There and back again: outlier detection between statistical reasoning and data mining algorithms. Wiley Interdiscip Rev Data Min Knowl Discov. 2018;8(6):e1280.
6. Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. ACM Comput Surv. 2009;41(3):1–58.
7. Angelin B, Geetha A. Outlier detection using clustering techniques-K-means and K-median. In: Proceedings of the international conference on intelligent computing control system. ICICCS 2020; 2020. p. 373–8.
8. Bergman L, Hoshen Y. Classification-based anomaly detection for general data. arXiv; 2020.
9. Wahid A, Annavarapu CSR. NaNOD: a natural neighbour-based outlier detection algorithm. Neural Comput Appl. 2020;33:2107–23.
10. Domański PD. Study on statistical outlier detection and labelling. Int J Autom Comput. 2020;17:788–811.
11. Dong Y, Hopkins SB, Li J. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. arXiv; 2019.
12. Shetta O, Niranjan M. Robust subspace methods for outlier detection in genomic data circumvents the curse of dimensionality. R Soc Open Sci. 2020;7(2):190714.
13. Li P, Niggemann O. Non-convex hull based anomaly detection in CPPS. Eng Appl Artif Intell. 2020;87:103301.
14. Borghesi A, Bartolini A, Lombardi M, Milano M, Benini L. Anomaly detection using autoencoders in high performance computing systems. CEUR Workshop Proc. 2019;2495:24–32.
15. Knorr E, Ng R. A unified notion of outliers: properties and computation. In: Proceedings of the 3rd ACM international conference on knowledge discovery and data mining (KDD), Newport Beach; 1997, p. 219–22.
16. Knorr E, Ng R. Algorithms for mining distance-based outliers in large datasets. In: Proceedings of the 24th international conference on very large data bases (VLDB), New York; 1998, p. 392–403.
17. Wu G et al. A fast kNN-based approach for time sensitive anomaly detection over data streams. In: International conference on computational science; 2019, p. 59–74.
18. Zhu R, et al. KNN-based approximate outlier detection algorithm over IoT streaming data. IEEE Access. 2020;8:42749–59.
19. Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. In: Proceedings of the ACM international conference on management of data (SIGMOD), Dallas; 2000, p. 427–38.
20. Angiulli F, Pizzuti C. Outlier mining in large high-dimensional data sets. IEEE Trans Knowl Data Eng. 2005;17(2):203–15.
21. Breunig M, Kriegel H, Ng R, Sander J. LOF: identifying density-based local outliers. In: Proceedings of the ACM international conference on management of data (SIGMOD), Dallas; 2000, p. 93–104.
22. Tukey JW. Exploratoy data analysis. Addison-Wesley Ser Behav Sci; 1977.
23. Kimber AC. Exploratory data analysis for possibly censored data from skewed distributions. Appl Stat. 1990;39:21–30.
24. Aucremanne L, Brys G, Hubert M, Rousseeuw PJ, Struyf A. A study of belgian inflation, relative prices and nominal rigidities using new robust measures of skewness and tail weight. In: Theory and applications of recent robust methods. Basel: Birkhäuser; 2004. p. 13–25.
25. Schwertman NC, Owens MA, Adnan R. A simple more general boxplot method for identifying outliers. Comput Stat Data Anal. 2004;47:165–74.
26. Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. Comput Stat Data Anal. 2008;52(12):5186–201.
27. Belhaouari SB, Ahmed S, Mansour S. Optimized K-means algorithm. Math Probl Eng. 2014; 2014.
28. N. Distribution. Encyclopedia.com: https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/distribution-normal. Gale encyclopedia of psychology.
29. Casella G, Berger RL. Statistical inference, 2nd edn. Duxbury. ISBN 978-0-534-24312-8; 2001.
30. Campos GO, et al. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Min Knowl Discov. 2016;30:891–927.

31. Angiulli F, Pizzuti C. Fast outlier detection in high dimensional spaces. In: Proceedings of the 6th European conference on principles of data mining and knowledge discovery (PKDD), Helsinki; 2002, p. 15–26.
32. Hautamäki V, Kärkkäinen I, Fränti P. Outlier detection using k-nearest neighbor graph. In: Proceedings of the 17th international conference on pattern recognition (ICPR), Cambridge; 2004, p. 430–3.
33. Schubert E, Zimek A, Kriegel H. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. Data Min Knowl Discov. 2014;28(1):190–237.
34. Tang J, Chen Z, Fu A, Cheung D. Enhancing effectiveness of outlier detections for low density patterns. In: Proceedings of the 6th Pacific-Asia conference on knowledge discovery and data mining (PAKDD), Taipei; 2002, p. 535–48.
35. Jin W, Tung A, Han J, Wang W. Ranking outliers using symmetric neighborhood relationship. In: Proceedings of the 10th Pacific-Asia conference on knowledge discovery and data mining (PAKDD), Singapore; 2006, p. 577–93.
36. Kriegel H, Kröger P, Schubert E, Zimek A. LoOP: local outlier probabilities. In: Proceedings of the 18th ACM conference on information and knowledge management (CIKM), Hong Kong; 2009, p. 1649–52.
37. Zhang K, Hutter M, Jin H. A new local distance-based outlier detection approach for scattered real- world data. In: Proceedings of the 13th Pacific-Asia conference on knowledge discovery and data mining (PAKDD), Bangkok; 2009, p. 813–22.
38. Latecki L, Lazarevic A, Pokrajac D. Outlier detection with kernel density functions. In: Proceedings of the 5th international conference on machine learning and data mining in pattern recognition (MLDM), Leipzig; 2007, p. 61–75.
39. Schubert E, Zimek A, Kriegel H. Generalized outlier detection with flexible kernel density estimates. In: Proceedings of the 14th SIAM International Conference on Data Mining (SDM), Philadelphia; 2014, p. 542–50.
40. Liu Y, et al. Generative adversarial active learning for unsupervised outlier detection. IEEE Trans Knowl Data Eng. 2020;32(8):1517–28.
41. Abe N, Zadrozny B, Langford J. Outlier detection by active learning. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, vol. 2006; 2006, p. 504–9
42. Yang X, Latecki LJ, Pokrajac D. Outlier detection with globally optimal exemplar-based GMM. In: Proceedings of the applied mathematics, society for industrial and applied mathematics—9th SIAM international conference on data minning 2009, vol. 1; 2009, p. 144–53.
43. Cohen G, Sax H, Geissbuhler A. Novelty detection using one-class parzen density estimator. An application to surveillance of nosocomial infections. Stud Health Technol Inform. 2008;136:21–6.
44. Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC, Scholkopf B. Estimating the support of a high-dimensional distribution. Neural Comput. 2001;13(7):1443–71.
45. Kriegel H, Schubert M, Zimek A. Angle-based outlier detection in high-dimensional data. In: Proceedings of the 14th ACM international conference on knowledge discovery and data mining (SIGKDD), Las Vegas; 2008, p. 444–52.

Publisher's Note