Int. J. Nonlinear Anal. Appl. 13 (2022) No. 1, 1683–1699
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.5784

A local density-based outlier detection method for high dimension data

Shahad Adel Abdulghafoor[b,a,*], Lekaa Ali Mohamed[a]

[a] University of Baghdad, College of Management and Economics, Department Of Statistics, Iraq
[b] Ministry of planning,Central Statistical Organization, Iraq

(Communicated by Madjid Eshaghi Gordji)

Abstract

The researchers faced challenges in the outlier detection process, mainly when deals with the high dimensional dataset; to handle this problem, we use The principal component analysis. Outlier detection or anomaly detection, with local density-based methods, compares the density of observation with the surrounding local density neighbors. We apply the outlier score as a measure of comparison. In this research, we choose different density estimation functions and calculated different distances. Weighted kernel density estimation with adaptive bandwidth for multivariate kernel density estimation(Gaussian) considered the KNN and RNN. KNN is considered too for the Epanenchnikov kernel density estimation. Lastly, we estimate the LOF as a base method in detecting outliers. Extensive experiments on a synthetic dataset have shown that RKDOS and EPA are more efficient than LOF using the precision evaluation criterion.

Keywords: local density; K-nearest neighbor; R-nearest neighbor; outlier score; WKDE.

1. Introduction

the most recent research on outlier detection methods which based on a probability density estimate (pdf) that follow up the structure of local outlier factor(LOF).
There are many definitions for outliers or anomalies. Grubbs defines an outlier as an outlying observation that appears to deviate from other members of the sample in which occurs [8]. Hawkins defines an outlier as an object that differs so much from other observations that it raises the possibility that it was caused by a different mechanism [9].

*Corresponding author
Email addresses: shahed.adel1001a@coadec.uobaghdad.edu.iq (Shahad Adel Abdulghafoor), lekaa.a@coadec.uobaghdad.edu.iq (Lekaa Ali Mohamed)

Received: May 2021 *Accepted:* October 2021

For the past decades, the outlier detection methods varied from: univariate to multivariate, unsupervised to supervised, parametric to nonparametric, global to local approach.
The classification of outlier detection methods into several types [16]: the first methods model-based or distribution-based, the second methods proximity-based (the approach based on distance and that based on density), the third methods Clustering-based.
Breunig et al. presented a density estimation method based on the local outlier factor (LOF). This method explains that it is more significant to give each object a grade of being an outlier. This grade named (LOF) means how much the observation is isolated locally from other surrounding neighborhoods [1].
Papadimitriou, S. et al. propose a method for detecting outlier and group of outliers called Local Correlation Integral for finding outliers in multiple dimensional data set [13]
Shekhar, S. et al. introduced a method for detecting spatial outliers in traffic data for multidimensional datasets. Defined the test and analyze the statistical model of this approach and provide a good algorithm and the cost model to detect spatial outlier [15].
Fan, H. et al. give a modern solution and data mining algorithm for nonparametric outlier detection. The outlier algorithm results take into account the local and global objects of the dataset. Both synthetic and real-life on large building contractor datasets are applied on the algorithm, and compared with another previous mining algorithm, this method proved effective and superior [4].
Latecki, L. J. et al. proposed a nonparametric algorithm using the variable kernel density estimation function to detect outliers locally. The local density of each object is compared with the local density of its surrounding neighbours. The algorithm was compared with the local outlier factor (LOF) and local correlation integral (LOCI) and proved efficiency and effectiveness [11]
Gao, J. et al. give a Multi-Scale Local Kernel Regression in a classic nonparametric regression by using the primary local density method in the local regression estimator of a kernel in neighborhoods that multi-scaled in detection outliers [6]
The LOF is not accurate enough when the data set is big. Gao, J. et al. present the density estimate of the variable kernel and the density estimation of the weighted neighborhood. They use various k parameters to improve robustness and the LOF framework. Besides, they propose another method of detection based on kernel density function called Volcano kernel, real and synthetic data set explains that these methods are suitable for a good detection performance, and also work in large data sets [7].
Fink, O. et al. detected outliers using approach based on multivariate kernel density estimation, The other approach is an unsupervised algorithm based on artificial neural gas named the growing neural gas (GNG). These two methods are applied in the railway field of turnout systems. Both approaches proved their appropriateness in detecting novel patterns. Moreover, the GNG was most suitable for input data dimensionality and online learning [5].
Tang, B. and He, H. introduce a density-based measure for local outliers detection, named relative density outlier detection. This method by which the object is estimated locally with local KDE by using the extended nearest neighbours (k nearest neighbours, reverse nearest neighbours, and shared nearest neighbours) to estimate the density distribution of an object[16].
Zhang, L. et al. presented a nonlinear system to measure outlier detection. They used the Gaussian kernel and adoptive kernel bandwidth for better smoothness and improved the distinguish power. The method put an outlier degree defined as a proportional measure for each sample to show the deviation for each sample from its neighbors in the local density[19].
Wahid, A. and Rao, A. C. S. propose an approach to detect outliers based on density estimation at the location of an object. In this method, the researchers use weighted kernel density estimation with an adoptive kernel width by the extended nearest neighbours using both KNN and RNN to

estimate the density of an object[17].
This paper applies LOF, WKDE for The Gaussian kernel and Epanchinikov kernel once for the Euclidian distance and once for the Chebyshev distance.

2. Methods

2.1. LOF

LOF [1] is an algorithm in data mining and belongs to density-based approaches. This method is presented by Breunig et al., which is supposed the density around regular observation is the same as the density around its neighbours. The density surrounding an outlier observation is exceptionally different from the density around its neighbours. It's a way in multidimensional datasets of finding outliers. Local in outlier factor means each point taking into account and restricted the neighborhood of that point only. The presented LOF algorithm attempts to find the outlier data points by measuring the local deviation of the given object while taking into account all other neighbors. In this algorithm, the outlier-score will tell us whether the observation is an outlier or not.
Let's have the data set D=$\{c_1, c_2, \ldots, c_n\}$, where c_1,c_2 and c_3 are observations in the dataset. This method uses the math abbreviation d(c_1, c_2) which implying the distance between objects c_1 and c_2.

- The k-distance of object c_1

  Let's k be a positive integer. it represents the k-distance between two observation$s$ c_1, c_2 ϵ D known as k-distance (c_1), and it is denoted as the distance d(c_1, c_2) such that :

  For at least k observations c_2' ϵ $D\diagdown\{c_1\}$ under the condition d(c_1, c_2')$\leq d(c_1, c_2)$.

  For at most k-1 observations c_2' ϵ $D\diagdown\{c_1\}$ under the condition d(c_1, c_2')$< d\left(c_1, c_2\right)$.

  The nearest distance between observation c_1 and its k-neighbors, if we have KNN to object c_1 then the k-distance of c_1 will equal to the maximum distance among all pairs of observation c_1.

- k-distance neighborhood of an observation c_1

  set k-distance of an observation c_1, the k-distance neighborhood of an observation c_1 consist of each observation whose distance from c_1 is not larger than k-distance:

$$N_{k-distance(c_1)}\left(c_1\right) = N_k(c_1) = \{c_3 \; \epsilon \; D\diagdown\{c_1\}| \; d\left(c_1, c_3\right) \leq k - distance(c_1)\} \tag{2.1}$$

  These observations c_3 is the k-nearest neighbor of c_1.

- reachability distance of an object $c_1 with\ respect\ to\ object$ c_2

  For any natural number k, The reachability distance of an observation $c_1$with respect to observation c_2 is determined as:

$$reach - dist_k\left(c_1, c_2\right) = max\left\{k - distance\left(c_2\right), d\left(c_1, c_2\right)\right\} \tag{2.2}$$

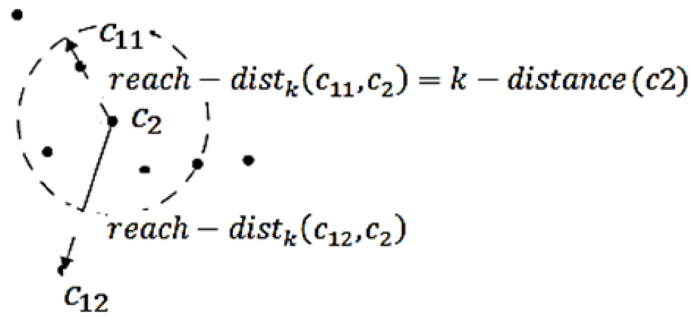

Figure 1: to clarify the concept of reachability distance with $k = 5$

When c_1 faraway from observation c_2 (c_{12} in figure1 the reach- dist. between them is their actual distance, but if the two observations are close (c_{11} in figure 1), the actual distance is equal to the k-distance of c_2 .So we do that to significantly reduce the statistical fluctuation for the distance d(c_1, c_2) to all the c_1's that near to c_2. Parameter k is the locality parameter it takes control of the smoothing effect. The greater the value of k, The closer the reachability distances between observations in the same neighborhood.

- local reachability density of an object c_1

  For the observation $c_1$we can calculate the local reach- dist. as:

$$Lrd_k\,(c_1) = \frac{1}{\left(\frac{\sum_{c_2 \epsilon N_k(c_1)} (c_1, c_2)}{|N_k(c_1)|}\right)} \tag{2.3}$$

  The average reachability distance of observation $c_1$depend on KNN for observation c_1, the local reachability density of an observation c_1, is the inverse of the average reachability distance.

  If we have an observation whose neighbors are entirely distant from it. Then the distance of the observation from their neighbors would become larger, which means the average distance would be higher so that when we divide one by the amount of the average distance, we gain a small density.

  That reasonable cause all the observation neighbors are entirely distant from it, and then we get this observation has low density.

  Furthermore, the close the neighbors are to the observation (the small the distance between the neighbors and the observation), the object's density will increase.

- local outlier factor of an observation c_1

  LOF of an observation c_1 is determined as:

$$Lof_k\,(\boldsymbol{c_1}) = \frac{\sum_{c_2 \epsilon N_k(c_1)} \frac{lrd_k(c_2)}{lrd_k(c_1)}}{|N_k(c_1)|} \tag{2.4}$$

of the local reachability density of $c_1$is represented by Lof which it is an average ratio and its nearest neighbors, Lof is good way to measure the degree to which $c_1$is an outlier. The smaller the local reachability density of the c_1 is the larger local reachability density of the c_1 nearest neighbors are, and the larger Lof value is.

2.2. Estimation Of Weighted Kernel Density With Adaptive Band Width

This local outlier detection method estimates weighted kernel density (WKD) with adaptive kernel bandwidth [17]. This method of estimation takes into account the observation neighborhoods instead of taking all the observations in the dataset. This approach takes into account two kinds of neighbors (K-nearest neighbor (KNN), reverse nearest neighbor (RNN)). This method calculates the Average Density Fluctuation (ADF) to measure the fluctuation of a data point with the rest of the objects in the influence set then evaluate the density for each observation by using the Relative kernel density-based outlier score (RDOS).

Let D be the given data set $D = \{c_1, c_2, c_3, c_4 \ldots, c_n\}$ where n is the sample size from a given data set or the data space from the euclidian distance

The presented approach calculates the degree of deviation or outlierness score locality for the data points. In order to calculate the local outline measurement, the method first performs a density estimate.

In estimating the density of the data points, we depend on the given data set, which uses an approach with nonparametric weight for KDE with adaptive bandwidth. In this approach, the density estimate is given by adapting the kernel estimator, where Each observation is given a sample weight.

The adaptive bandwidth for KDE depends on the random sample $c_1, c_2, c_3, c_4 \ldots, c_n$

Where $c_i \epsilon R^d \;\; for \;\; i = 1, 2, 3, 4, \ldots, n$ with weight $w_1, w_2, \ldots, w_n$, which have been normalized to equal to $1(\sum_{j=1}^{n} w_j = 1)$,and the weigted KDE is:

$$p(c_i) = \sum_{j=1}^{n} \frac{w_j}{h_j^d} K\left(\frac{c_i - c_j}{h_j}\right) \tag{2.5}$$

Where K($*$) is denoted as the kernel function, The smoothness of the estimator is controlled by the bandwidth h_j the smoothing parameter, w_j is performing the observation's weight, which the formula can represent as :

$$w_j = \frac{a - \sum_{j=1}^{n} Euclidean(x_i, x_j)}{a} \tag{2.6}$$

Where $[a]$ is known as the highest Euclidean dist. between the points and the point that applied normalization $[\sum_{j=1}^{n} Euclidean(x_i, x_j)]$ is denoted as the sum of Euclidean dist. for the point x_j of the jth Gaussian to the point x_i, including the outliers.

The kernel function that commonly used is many, but we will use Gaussian and Epanenchnikov[3]. The smoothing kernel is necessary to obtain smoothness in density estimation. The characteristic of smoothing kernel function is:

$$\int K(u)du = 1, \quad \int u \; K(u)du = 0, \quad \int u^2 K(u)du \; > 0 \tag{2.7}$$

The Gaussian kernel and Epanchinikov kernel the most widely used function in outlier detection

$$K(\frac{c_i - c_j}{h_j})_{Gaussian} = \frac{1}{(2\pi)^d} exp\left(-\frac{\|c_i - c_j\|^2}{2 * h_j^2}\right) \tag{2.8}$$

$$K(\frac{c_i - c_j}{h_j})_{Epanechnikov} = \left(\frac{3}{4}\right)^d \left(1 - \frac{\|c_i - c_j\|^2}{h_j^2}\right) \tag{2.9}$$

Where $\|c_i - c_j\|$ is the Euclidean distance between the points c_i and c_j ,d: is dimensions, h: is kernel bandwidth.
Chebyshev distance is used to calculate the maximum distance[2] between any two points c_1 and c_2 or any two coordinates c_i and c_j. It is also referred to as the chessboard distance since chess is a game of strategy, where a king's minimal number of moves from one square on the chessboard to another is the Chebyshev distance between squares centres.

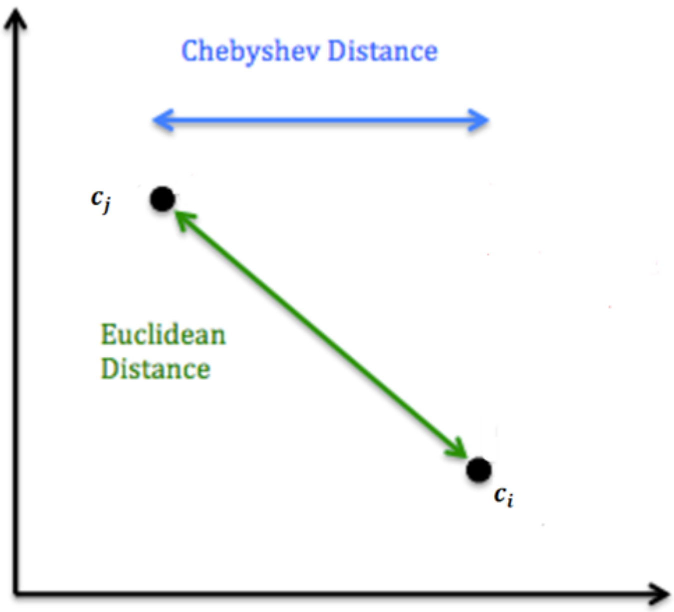

Figure 2: the difference between the Euclidean and Chebyshev distances.

The calculation of Chebyshev distance, if we have two vectors with N-dimensional points $c_1 = \{c_{11}, c_{12}, \cdots, c_{1N}\}$ and $c_2 = \{c_{21}, c_{22}, \cdots, c_{2N}\}$ the Chebyshev distance can be determined as :

$$d = max(|c_{11} - c_{21}|, |c_{12} - c_{22}|, \cdots, |c_{1N} - c_{2N}|) \tag{2.10}$$

If we have two-point $c_1 = \{4,\ 5\}$ and $c_2 = \{3, 8\}$ so the distance calculation would be equal to:

$$d = max(|4 - 3|,\quad |5 - 8|) = max(1, 3) = 3$$

2.3. Estimation of density at the location

Estimation of density at the location c_i[17] It is determined by considering its neighbors as kernels instead of the data point in the data set. Because if we first estimate the density for all the data points, we may lose the local density differences, and we will be unsuccessful in identifying the local outliers. Second, The approach for detecting outliers determines the outlying degree for all data points in the data sets will lead to high computation complexity, especially in O(n^2) where n is the data set's total number of samples.
For accurate density estimation in the neighborhood, this approach applies the influence set (I_Set), the set of KNN, and the RNN to a data point. Influential observations that influence c are included in the influence set for an observation c. more precisely estimation of density in c's neighborhood

with regards to these objects. In recent research, the RNN has been present for best local distribution data information and applied to detect outliers and classification [10][3].
If $NN_j(c_i)$ is the j^{th} nearest neighbors of the observation c_i , the set of K nearest neighbor can be written as:

$$KNN\left(c_i\right) = \left\{NN_1\left(c_i\right), NN_2\left(c_i\right), ..., NN_k\left(c_i\right)\right\} \tag{2.11}$$

Let we have two data points, c and x_i taken from the collection of a data set, we can define KNN by using the Euclidean distance and calculating the distance between c and x_i as $d(c, xi)$ from all the points $i = 1, 2, 3, \ldots, n$:

$$KNN = \{xi : \ d(c, xi) \leq d_k \tag{2.12}$$

Where d_k is the minimum distance from c to x_i
The RNNs of the observation c_i, are these points that have c_i is considered one of their K-nearest neighbors, or we say that the observation c is consider one of the RNNs of c_i if

$$NN_j\left(c\right) = c_i \qquad j \leq k \tag{2.13}$$

So that the RNN could be equal to zero or one or more data points.
Then the kernel function in (2.5) would be equivalent to kernel estimation at the location c_i :

$$p\left(c_i\right) = \sum_{c\epsilon I\_set\left(c_i\right)} \frac{w_p}{h_c^d} K\left(\frac{c_i - c}{h_c}\right) \tag{2.14}$$

from equation (2.14), the final formula of density estimation at the location of c_i is:

$$p\left(c_i\right) = \sum_{c\epsilon I\_set\left(c_i\right)} \frac{w_p}{h_c^d * \ (2\pi)^{\frac{d}{2}}} exp(-\frac{\left\|c_i - c\right\|^2}{2 * h_c^2}) \tag{2.15}$$

where the $|I - set\left(c_i\right)|$ Defined as the influence set, which is a number of data points.

2.4. The computation of the influence set

For each observation $c\epsilon D$[17], We get at least k results from the NN_K search, while reverse nearest neighbors can be equal to zero, one or more observations. By merging $NN_k(c)$ and $RNN_k\left(c\right)$ as k-influence space for c and determined as I_$Set_k(c)$,We will generate a space of local neighborhood around c in estimating the density distribution.
Figure 3 explain the RNN and how it obtained the data points $[c, c_1, c_2, c_3, \ c_4 and, c_5]$, Where $k =$ 4. , $NN_K\left(c\right) = \{\ c_1, c_2, c_3, c_4\}$, $NN_K\left(c_1\right) = \{c, c_3, c_4, c_5\}$, $NN_K\left(c_2\right) = \{c, c_1, c_3, c_4\}$, $NN_K\left(c_3\right) = \{c, c_1, c_4, c_5\}$, $NN_K\left(c_4\right) = \{c, c_1, c_2, c_5\}$ and $NN_K\left(p_5\right) = \{c, c_1, c_3, c_4\}$.
While KNN searches for the points c, c_1, c_2, c_3, c_4, and c_5, the $RNN_K\left(c\right) = \{\ c_1, c_2, c_3, c_4, \ c_5\}$ is gradually structured. And in the same way $RNN_K\left(c_1\right) = \{c, c_2, c_3, c_4, c_5\}$, $RNN_K\left(c_2\right) = \{c, c_4\}$,$RNN_K\left(c_3\right) = \{c, c_1, c_2, c_5\}$ $RNN_K\left(c_4\right) = \{c, c_1, c_2, c_3, c_5\}$ *and* $RNN_K\left(c_5\right) = \{c_1, c_3, c_4\}$.
We notice that:
$NN_K\left(c\right) = \{c_1, c_2, c_3, c_4\}$ and $RNN_K\left(c\right) = \{\ c_1, c_2, c_3, c_4, \ c_5\}$ will be equal to $I\_Set_4(c) = \{\ c_1, c_2, c_3, c_4, \ c_5\}$ and similarly, we calculate $I\_Set_4(c_1)$, $I\_Set_4(c_2)$, $I\_Set_4(c_3)$, $I\_Set_4(c_4)$, $I\_Set_4(c_5)$.

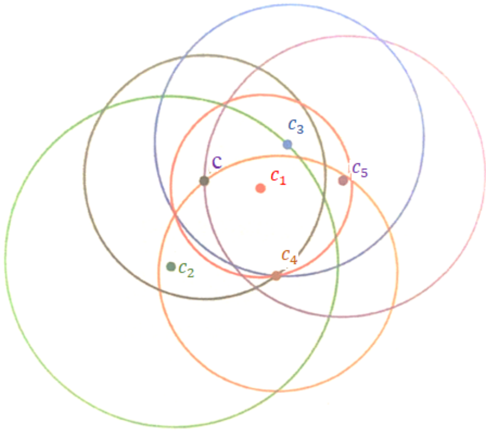

Figure 3: the influence set KNN and RNN of c

2.5. Calculation of Adaptive Width

In the density estimation at c_i Location, We use the KDE to determine the density for c_i at the location by the criterion for outlier detection using the function of kernel smooth to perform the smoothness. We use the adaptive kernel bandwidth to improve the power of differencing between anomalies and regular data points.

In calculation KDE, the bandwidth parameter h of all data points is fixed. Therefore h optimum is determined by the specific locations in the data space. Still, when using large width parameter h in the low-density region could result in over-smoothing, while small h parameter can result in noise estimates.

Adaptive kernel width will have two outcomes: the data points that are distant from others are the more isolated, and there is a hazy difference between the regular data points. We shall employ the notion of adaptive kernel width shown in [19].

The estimated kernel adaptive bandwidth is equal to:

$$h_i = \alpha \left[d_{k-max} + d_{k-min} + \delta - d_k\left(c_i\right)\right] \tag{2.16}$$

Where

$d_k\left(c_i\right)$: indicates the average distance between one point and the KNN. Whereas

$$d_k\left(c_i\right) = \frac{1}{k} * \sum_{j \epsilon kNN(c_i)} d(c_i, c_j)$$

d_{k-max} : the biggest amount in the dataset $\{d_k(c_i) \mid i = 1, 2, \ldots, n\}$

d_{k-min}: the smallest amount in the dataset $\{d_k(c_i) \mid i = 1, 2, \ldots, n\}$

And a ($\alpha > 0$): The smoothing effect is controlled by the scale factor.

δ: a minimal positive numeric number that ensures that the kernel width does not equal zero.

2.6. RKDOS Method

following the evaluation of each object's density, RKDOS is used to determine how different observation c's density is from the surroundings, as it's known [17]:

$$RKDOS_k\left(c_i\right) = \frac{\sum_{p \epsilon I\_Set(c_i)} ADF\left(c\right)}{ADF\left(c_i\right) * \left|I_{Set(c_i)}\right|} \tag{2.17}$$

$RKDOS_k$: is the average density fluctuation (ADF) of influence set observations divided by the ADF of a test point c_i, If $RKDOS_k(c_i)$ is considerably more than 1, the observation c_i Is outside the cluster density, indicating that it is an outlier. If $RKDOS_k(c_i)$ equals or is less than 1, the observation c_i is surrounded by a dense population of neighbors. This proves that c_i Observation is not an anomaly.
We should mention that the average density fluctuation ADF can be determined as:

$$ADF\left(p_i\right) = \frac{\sum_{p\epsilon N}\left(\rho\left(p_i\right) - \rho(p)\right)^2}{|N|} \tag{2.18}$$

The RKDOS method uses KNN as an input graph to give a detailed description of RKDOS. KNN is a directed graph in which every observation is a vertex that is linked to its nearest neighbors in an outbound manner. In the KNN-G, the outward edge of a data point is k, and it has zero, one, or more incoming edges. From the KNN-G, the KNN and RNN may be simply calculated. We create an influence set (I_set) and describe a method for combining KNN with RNN. Then, based on the densities of neighbors in I_set, compute the RKDOS for each observation.
In this research, we will use the principal component analysis PCA[12]. Its linear transformation reduces the information and keeps it as a component, and the variance represents this information. It's also a mathematical style that transforms a set of correlated explanatory variables into a new orthogonal set of uncorrelated variables named the principal component. The PCA employs the variance-covariance matrix or the correlation matrix for the explanatory variables in this analysis. The PCA is a linear combination of explanatory variables $c_1, c_2, \ldots, c_n$ determined by eigen vectors a_i which is related to eigen values λ_i, the eigen value came from var-cov matrix or correlation matrix. We should mention that the number of principal components is equal to the number of explanatory variables. The mathematical expression is:

$$PC_i = a_{i1}C_i + a_{i2}C_i + \cdots + a_{in}C_i,\ i = 1, 2, \ldots, n \tag{2.19}$$

$$\begin{bmatrix} PC_1 \\ PC_2 \\ \vdots \\ PC_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_n \end{bmatrix} = A'C \tag{2.20}$$

Each column in matrix A represented eigen vector that corresponding to the eigen value and correlated to it.
The first PC is a linear combination of explanatory variables with the coefficient of eigen vector corresponding to the first eigen value $\lambda_1$represent the biggest eigen value. The eigen values will be represented as:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq\ \lambda_n \tag{2.21}$$

Altho the number of PC is equal to the explanatory variables, but we will use the dimension reduction we will eliminate the weak PC specially the variance of eigen value that less than 1.
In the calculation steps; we will use a variance-covariance matrix if we have the same measuring units and a correlation matrix if we have different measuring units, finding the first chara vector $\underline{a}_1$ which its equal to the first eigen value λ_1, the element of $\underline{a}_1$ should satisfy :

$$\underline{a}'_1\underline{a}_1 = 1 \tag{2.22}$$

The formula of the first principle component is :

$$PC_1 = a_{11}C_1 + a_{12}C_2 + \cdots + a_{1n}C_n \tag{2.23}$$

The second:

$$PC_2 = a_{21}C_1 + a_{22}C_2 + \cdots + a_{2n}C_n \tag{2.24}$$

The points of the characteristic vector are chosen under the two constrain:

$$\underline{a}_2'\underline{a}_2 = 1, \quad \underline{a}'_1\ \underline{a}_2 = 0 \tag{2.25}$$

Where PC_1 & PC_2 under this condition are orthogonal.
the criterion for evaluating the performance of outlier approaches in this paper is Precision(P)[18], it is defined as the ratio that divided the number of correct outliers by the total number of points that filtered to be outliers:

$$Precision\frac{m}{t} \tag{2.26}$$

m=number of correct outliers that found in the set, t=total number of points that filtered to be outliers

3. Experiment Analysis and Result

The dataset is about three groups of random numbers naturally generated according to the normal distribution of mean= 0 and variance 0.5. These groups are represented in three different sizes (N=50, N=100, N= 150). Three explanatory variables were generated (P= 3, P=5, P=7). The nearest neighbours are in the range (2 to 10). And the number of iterations for each dataset and explanatory variables is (itr =100) according to Tabel 1. Several experiments were conducted.

Table 1: The order of initial variables generated according to the size of the variables and the sizes of the samples

Explanatory variables	**Sizes of samples**			**nearest neighborhood**
3	50	100	150	2,3,4,5,6,7,8,9,10
5	50	100	150	2,3,4,5,6,7,8,9,10
7	50	100	150	2,3,4,5,6,7,8,9,10

LOF, RKDOS and EPA are the methods applied to the simulated data set of size (N=50) of three explanatory variables (3, 5, 7) with Euclidian distance and Chebyshev distance, as shown in Table 2 and Tabel 2 as shown below where the average number of outliers in the Chebyshev distance is larger than in Euclidian. The figures 4, 6, 8, 5, 7 and 9 explain the difference between these two distances in the three methods.
The same methods applied for sample size (N=100) and the three variables (3, 5, 7). Table 4, 5 show that (RKDOS, EPA) is significantly larger in the average number of outliers than the same methods in the Echiedian distance while the LOF is slightly decreasing in the Chebyshev distance.

The figures 10, 12, 14, 11, 13 and 15) illustrate the increase and the decrease in (LOF, RKDOS and EPA).

For sample size (150), the RKDOS method significantly increased when calculating the average number of outliers in the Chebyshev distance, especially when the number of variables was (3, 5). Slightly increases in the average number of outliers when we use the Chebyshev distance for The EPA method but minimal fluctuation for the LOF method. The performance of Euclidian and Chebyshev distance is in the Figures (16, 18, 20, 17, 19 and 21).

Table 2: The results of 50 observations for the methods with 100 replicate for Euclidean distance

	average number of outliers								
k	3			5			7		
	LOF	RKDOS	EPA	LOF	RKDOS	EPA	LOF	RKDOS	EPA
2	15	21	25	15	18	27	15	17	28
3	16	14	21	16	14	24	16	11	25
4	17	13	17	16	12	22	16	9	22
5	17	11	17	16	12	21	16	11	19
6	16	10	16	16	11	20	16	9	20
7	16	11	16	16	10	19	16	11	19
8	16	11	16	15	12	18	15	10	19
9	15	11	15	15	10	17	15	9	18
10	14	11	15	15	11	17	15	9	17

Table 3: The results of 50 observations for the methods with 100 replicate for Chebyshev distance

	average number of outliers								
K	3			5			7		
	LOF	RKDOS	EPA	LOF	RKDOS	EPA	LOF	RKDOS	EPA
2	15	26	24	16	20	27	16	19	26
3	16	15	22	16	11	24	17	9	25
4	17	12	20	17	11	23	17	11	23
5	16	9	19	16	11	23	17	10	21
6	16	13	18	16	11	22	16	9	20
7	16	11	18	15	10	21	16	10	19
8	16	12	17	15	8	20	16	10	19
9	15	10	16	14	10	19	15	11	19
10	15	11	16	14	9	18	15	10	18

Table 4: The results of 100 observations for the methods with 100 replicate for Euclidean distance

k	average number of outliers								
	3			5			7		
	LOF	RKDOS	EPA	LOF	RKDOS	EPA	LOF	RKDOS	EPA
2	35	40	43	32	34	46	31	37	48
3	37	35	28	35	31	36	35	26	34
4	37	23	24	35	23	31	35	22	30
5	37	23	23	35	21	29	34	18	29
6	36	22	23	35	16	27	34	19	29
7	35	18	24	34	17	26	33	19	27
8	34	21	24	34	22	26	32	14	27
9	33	21	24	33	22	25	31	17	26
10	32	20	23	32	22	25	30	19	26

Table 5: The results of 100 observations for the methods with 100 replicate for Chebyshev distance

K	average number of outliers								
	3			5			7		
	LOF	RKDOS	EPA	LOF	RKDOS	EPA	LOF	RKDOS	EPA
2	33	52	40	32	41	45	31	39	45
3	35	28	30	34	33	34	34	26	34
4	36	25	26	36	22	31	35	19	32
5	38	22	24	36	20	29	35	18	32
6	36	23	24	35	20	29	35	18	31
7	35	20	24	33	19	29	35	20	30
8	33	21	24	32	19	28	34	19	29
9	33	21	24	31	17	28	34	16	28
10	32	19	24	32	17	27	32	20	27

Table 6: The results of 150 observations for the methods with 100 replicate for Euclidean distance

K	average number of outliers								
	3			5			7		
	LOF	RKDOS	EPA	LOF	RKDOS	EPA	LOF	RKDOS	EPA
2	49	52	46	49	58	56	50	59	58
3	54	51	29	54	47	39	55	41	43
4	55	45	25	55	36	34	57	35	36
5	54	39	24	55	29	30	56	35	34
6	54	33	23	55	34	30	54	26	34
7	54	29	22	54	28	30	53	29	35
8	52	32	23	52	30	30	51	23	34
9	52	28	23	51	30	30	50	24	33
10	49	27	23	51	27	30	49	29	33

Table 7: The results of 150 observations for the methods with 100 replicate for Chebyshev distance

	average number of outliers								
K	3			5			7		
	LOF	RKDOS	EPA	LOF	RKDOS	EPA	LOF	RKDOS	EPA
2	53	76	49	51	85	55	49	59	62
3	56	45	33	55	44	35	54	44	40
4	57	40	28	56	40	34	55	35	36
5	57	33	26	55	35	32	55	34	36
6	55	40	25	55	32	32	55	24	35
7	54	28	25	56	32	33	54	29	36
8	54	31	25	53	27	34	52	22	36
9	52	31	26	54	32	34	51	22	36
10	50	35	27	52	27	35	50	26	35

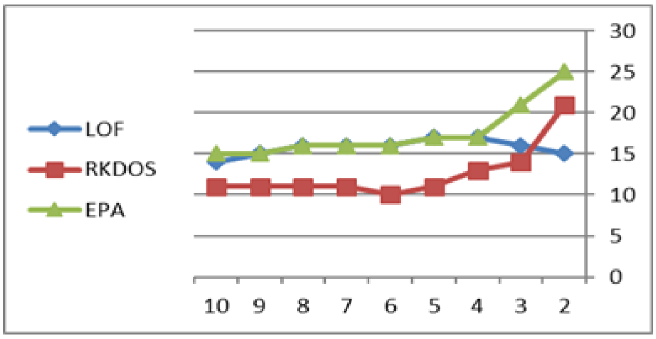

Figure 4: sample size 50 three variables with Euclidian distance for LOF, RKDOS, EPA

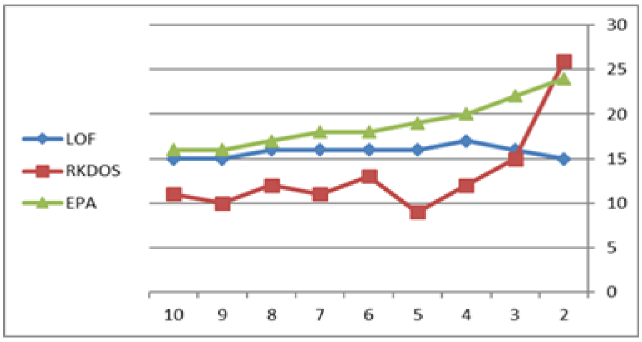

Figure 5: sample size 50 three variables with Chebyshev distance for LOF, RKDOS, EPA

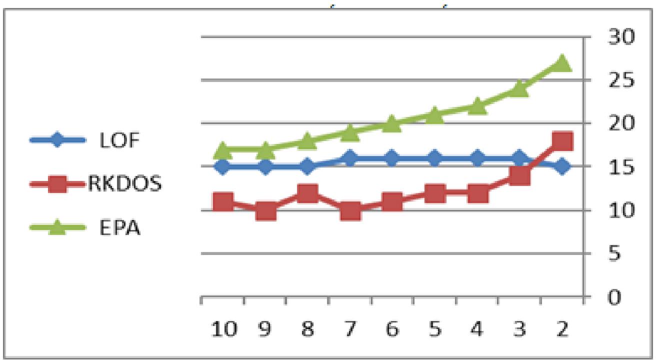

Figure 6: sample size 50 five variables with Euclidian distance for LOF, RKDOS, EPA

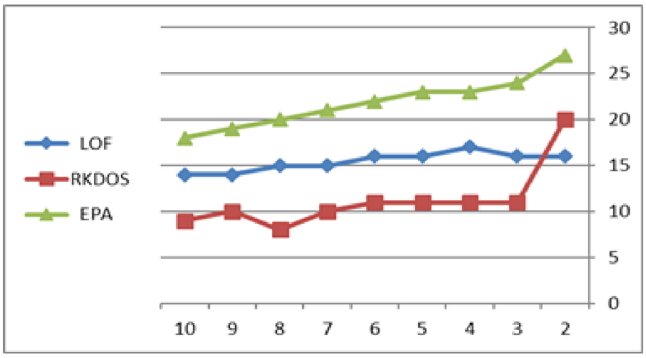

Figure 7: sample size 50 five variables with Chebyshev distance for LOF, RKDOS, EPA

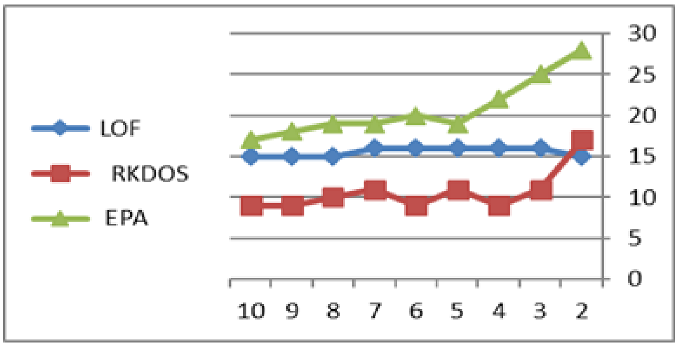

Figure 8: sample size 50 seven variables with Euclidian distance for LOF, RKDOS, EPA

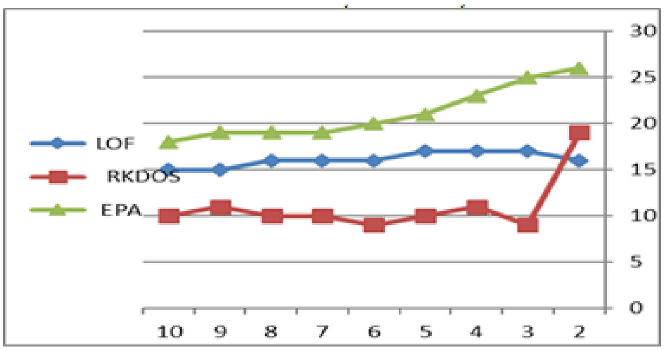

Figure 9: sample size 50 seven variables with Chebyshev distance for LOF, RKDOS, EPA

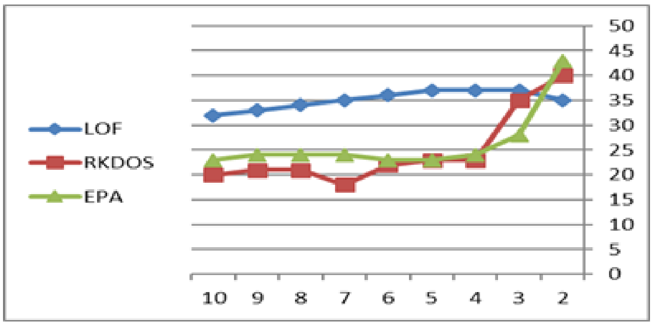

Figure 10: sample size 100 three variables with Euclidian distance for LOF, RKDOS, EPA

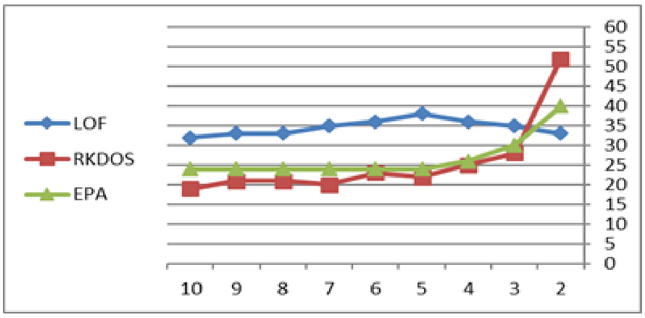

Figure 11: sample size 100 three variables with Chebyshev distance for LOF, RKDOS, EPA

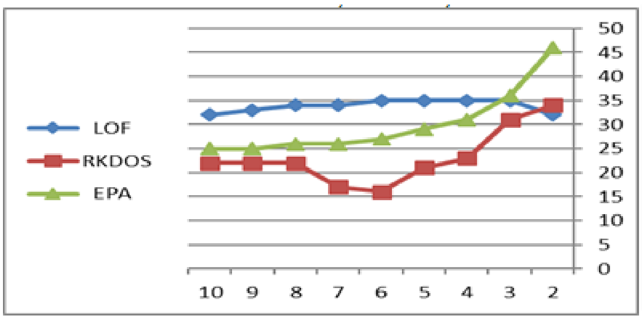

Figure 12: sample size 100 five variables with Euclidian distance for LOF, RKDOS, EPA

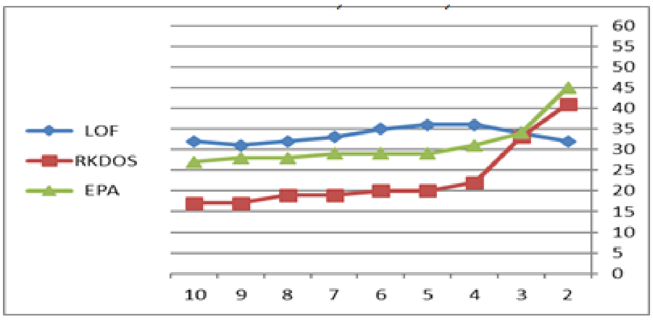

Figure 13: sample size 100 five variables with Chebyshev distance for LOF, RKDOS, EPA

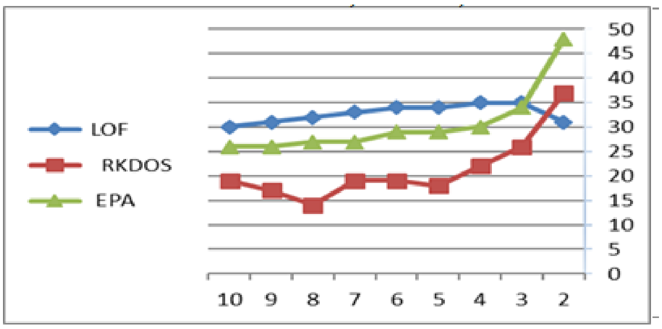

Figure 14: sample size 100 seven variables with Euclidian distance for LOF, RKDOS, EPA

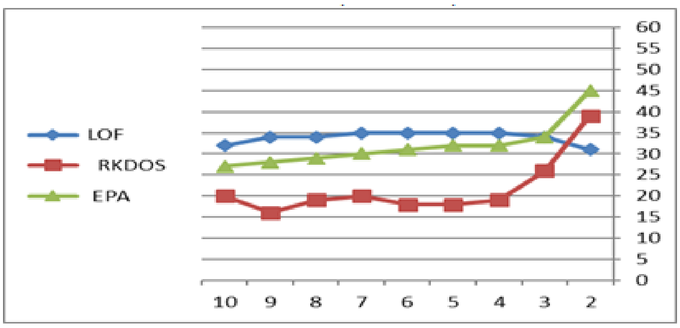

Figure 15: sample size 100 seven variables with Chebyshev distance for LOF, RKDOS, EPA

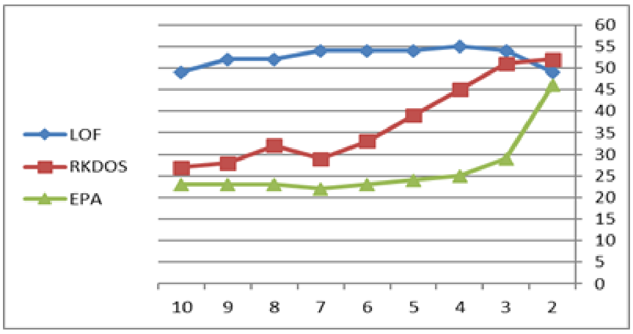

Figure 16: sample size 150 three variables with Euclidian distance for LOF, RKDOS, EPA

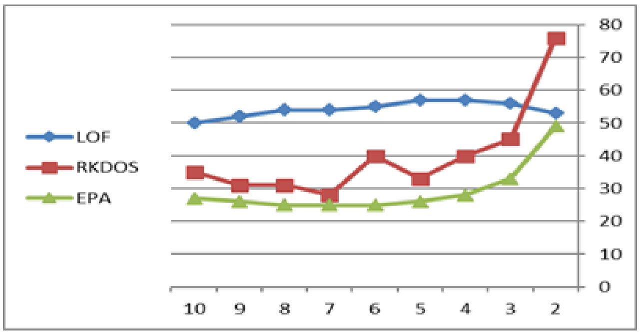

Figure 17: sample size 150 three variables with Chebyshev distance for LOF, RKDOS, EPA

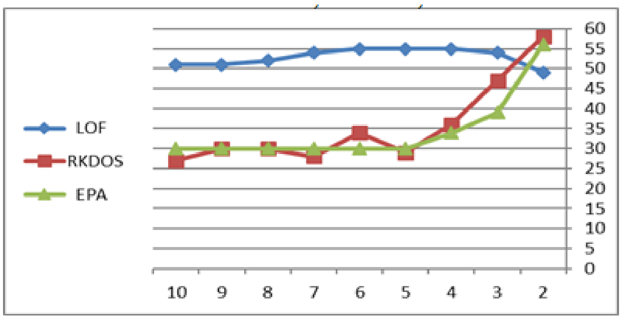

Figure 18: sample size 150 five variables with Euclidian distance for LOF, RKDOS, EPA

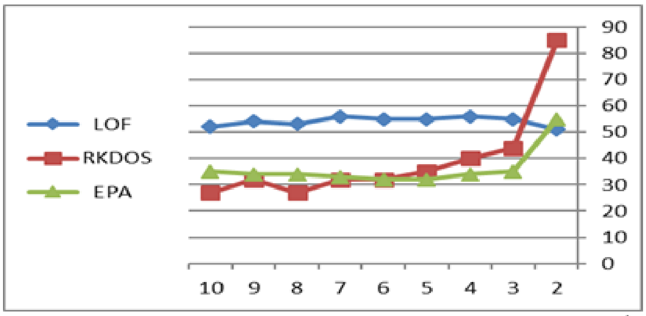

Figure 19: sample size 150 five variables with Chebyshev distance for LOF, RKDOS, EPA

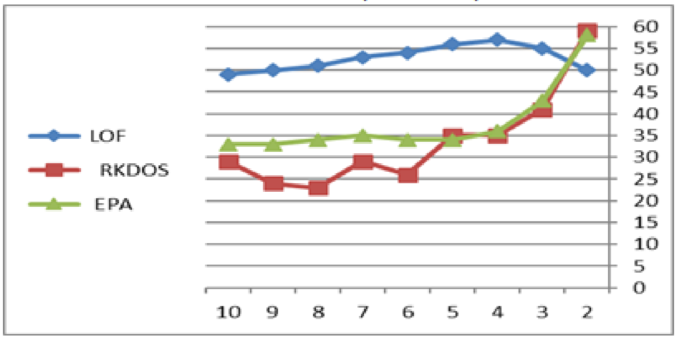

Figure 20: sample size 150 seven variables with Euclidian distance for LOF, RKDOS, EPA

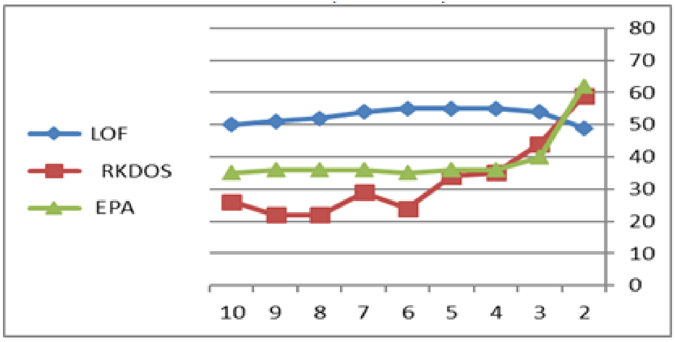

Figure 21: sample size 150 seven variables with Chebyshev distance for LOF, RKDOS, EPA

Table 8: The precision ratio of three sample sizes (50, 100 and 150) and with k from (2 to 10).

	testing the performans of outliers by the precision ratio on sample size 50 for the Euclidean distance								
k	3			5			7		
	LOF	RKDOS	EPA	LOF	RKDOS	EPA	LOF	RKDOS	EPA
2	0.54	0.75	0.89	0.54	0.64	0.96	0.54	0.61	1.00
3	0.57	0.50	0.75	0.57	0.50	0.86	0.57	0.39	0.89
4	0.61	0.46	0.61	0.57	0.43	0.79	0.57	0.32	0.79
5	0.61	0.39	0.61	0.57	0.43	0.75	0.57	0.39	0.68
6	0.57	0.36	0.57	0.57	0.39	0.71	0.57	0.32	0.71
7	0.57	0.39	0.57	0.57	0.36	0.68	0.57	0.39	0.68
8	0.57	0.39	0.57	0.54	0.43	0.64	0.54	0.36	0.68
9	0.54	0.39	0.54	0.54	0.36	0.61	0.54	0.32	0.64
10	0.50	0.39	0.54	0.54	0.39	0.61	0.54	0.32	0.61

	testing the performans of outliers by the precision ratio on sample size 50 for the Chebyshev distance								
K	3			5			7		
	LOF	RKDOS	EPA	LOF	RKDOS	EPA	LOF	RKDOS	EPA
2	0.54	0.93	0.86	0.57	0.71	0.96	0.57	0.68	0.93
3	0.57	0.54	0.79	0.57	0.39	0.86	0.61	0.32	0.89
4	0.61	0.43	0.71	0.61	0.39	0.82	0.61	0.39	0.82
5	0.57	0.32	0.68	0.57	0.39	0.82	0.61	0.36	0.75
6	0.57	0.46	0.64	0.57	0.39	0.79	0.57	0.32	0.71
7	0.57	0.39	0.64	0.54	0.36	0.75	0.57	0.36	0.68
8	0.57	0.43	0.61	0.54	0.29	0.71	0.57	0.36	0.68
9	0.54	0.36	0.57	0.50	0.36	0.68	0.54	0.39	0.68
10	0.54	0.39	0.57	0.50	0.32	0.64	0.54	0.36	0.64

	testing the performans of outliers by the precision ratio on sample size 100 for the Euclidean distance								
k	3			5			7		
	LOF	RKDOS	EPA	LOF	RKDOS	EPA	LOF	RKDOS	EPA
2	0.67	0.77	0.83	0.62	0.65	0.88	0.60	0.71	0.92
3	0.71	0.67	0.54	0.67	0.60	0.69	0.67	0.50	0.65
4	0.71	0.44	0.46	0.67	0.44	0.60	0.67	0.42	0.58
5	0.71	0.44	0.44	0.67	0.40	0.56	0.65	0.35	0.56
6	0.69	0.42	0.44	0.67	0.31	0.52	0.65	0.37	0.56
7	0.67	0.35	0.46	0.65	0.33	0.50	0.63	0.37	0.52
8	0.65	0.40	0.46	0.65	0.42	0.50	0.62	0.27	0.52
9	0.63	0.40	0.46	0.63	0.42	0.48	0.60	0.33	0.50
10	0.62	0.38	0.44	0.62	0.42	0.48	0.58	0.37	0.50

	testing the performans of outliers by the precision ratio on sample size 100 for the Chebyshev distance								
K	3			5			7		
	LOF	RKDOS	EPA	LOF	RKDOS	EPA	LOF	RKDOS	EPA
2	0.63	1.00	0.77	0.62	0.79	0.87	0.60	0.75	0.87
3	0.67	0.54	0.58	0.65	0.63	0.65	0.65	0.50	0.65
4	0.69	0.48	0.50	0.69	0.42	0.60	0.67	0.37	0.62
5	0.73	0.42	0.46	0.69	0.38	0.56	0.67	0.35	0.62
6	0.69	0.44	0.46	0.67	0.38	0.56	0.67	0.35	0.60
7	0.67	0.38	0.46	0.63	0.37	0.56	0.67	0.38	0.58
8	0.63	0.40	0.46	0.62	0.37	0.54	0.65	0.37	0.56
9	0.63	0.40	0.46	0.60	0.33	0.54	0.65	0.31	0.54
10	0.62	0.37	0.46	0.62	0.33	0.52	0.62	0.38	0.52

	testing the performans of outliers by the precision ratio on sample size 150 for the Euclidean distance								
K	3			5			7		
	LOF	RKDOS	EPA	LOF	RKDOS	EPA	LOF	RKDOS	EPA
2	0.58	0.61	0.54	0.58	0.68	0.66	0.59	0.69	0.68
3	0.64	0.60	0.34	0.64	0.55	0.46	0.65	0.48	0.51
4	0.65	0.53	0.29	0.65	0.42	0.40	0.67	0.41	0.42
5	0.64	0.46	0.28	0.65	0.34	0.35	0.66	0.41	0.40
6	0.64	0.39	0.27	0.65	0.40	0.35	0.64	0.31	0.40
7	0.64	0.34	0.26	0.64	0.33	0.35	0.62	0.34	0.41
8	0.61	0.38	0.27	0.61	0.35	0.35	0.60	0.27	0.40
9	0.61	0.33	0.27	0.60	0.35	0.35	0.59	0.28	0.39
10	0.58	0.32	0.27	0.60	0.32	0.35	0.58	0.34	0.39

	testing the performans of outliers by the precision ratio on sample size 150 for the Chebyshev distance								
K	3			5			7		
	LOF	RKDOS	EPA	LOF	RKDOS	EPA	LOF	RKDOS	EPA
2	0.62	0.89	0.58	0.60	1.00	0.65	0.58	0.69	0.73
3	0.66	0.53	0.39	0.65	0.52	0.41	0.64	0.52	0.47
4	0.67	0.47	0.33	0.66	0.47	0.40	0.65	0.41	0.42
5	0.67	0.39	0.31	0.65	0.41	0.38	0.65	0.40	0.42
6	0.65	0.47	0.29	0.65	0.38	0.38	0.65	0.28	0.41
7	0.64	0.33	0.29	0.66	0.38	0.39	0.64	0.34	0.42
8	0.64	0.36	0.29	0.62	0.32	0.40	0.61	0.26	0.42
9	0.61	0.36	0.31	0.64	0.38	0.40	0.60	0.26	0.42
10	0.59	0.41	0.32	0.61	0.32	0.41	0.59	0.31	0.41

Table 8 illustrates the precision ratio of the RKDOS experiments decreases when the number of neighbors approaches 10. As well, the precision ratio of the EPA gradually decrease when the number of neighbors increases. While the LOF experiments show fluctuations from increase to decrease, it's the least affected method in increasing the number of neighbors.

4. Discussion

From the results above, We notice that when we increase the number of neighbors, it significantly influences the number of outliers specially in the two methods of (RKDOS, EPA), but the method of LOF was the least affected.
When we compare the result from the Euclidean distance and Chebyshev distance specially for the (RKDOS, EPA) the average number of outliers in Chebyshev is larger than the average of outliers in the Euclidian distance, but when the number of neighbors increases, the number of outliers will decrease. Noticing that Local Outlier Factor(LOF) is less affected by the distance changes than RKDOS and EPA.

References

[1] M. M. Breunig, H. P. Kriegel, R. T. Ng and J. Sander, *LOF: identifying density-based local outliers*, ACM SIGMOD Record, 29 (2)(2000) 93-104.
[2] S. Dahal, *Effect of different distance measures in result of cluster analysis*, MS thesis, 2015.
[3] V. A. Epanechnikov, *Non-parametric estimation of a multivariate probability density*, Theory Probab. Appl., 14 (1)(1969) 153-158.
[4] H. Fan , O. R. Zaïane, A. Foss and J. Wu, *A nonparametric outlier detection for effectively discovering top-n outliers from engineering data*, In: Pacific-Asia conf. Knowl. Discovery Data Min. Springer, Berlin, Heidelberg, 2006, pp. 557-566 .
[5] O. Fink, E. Zio and U. Weidmann, *Novelty detection by multivariate kernel density estimation and growing neural gas algorithm*, Mech. Syst. Signal Proc., 50(2015) 427-436.
[6] J. Gao, W. Hu, W. Li, Z. Zhang and O. Wu, *Local outlier detection based on kernel regression*, In: 2010 20th Int. Conf. Pattern Recognit., 2010 pp. 585-588, IEEE.
[7] J. Gao, W. Hu, Z. M. Zhang, X. Zhang and O. Wu, *RKOF: robust kernel-based local outlier detection*, In: Pacific-Asia Conf. Knowl. Discovery Data Min. Springer, Berlin, Heidelberg, 2011, pp.270-283.
[8] F. E. Grubbs, *Procedures for detecting outlying observations in samples*, Technometrics, 11 (1) (1969) 1-21.
[9] D. M. Hawkins , *Identification of Outliers*, Chapman and Hall., London, Vol 11, 1980.
[10] W. Jin, A. K. H. Tung, J. Han and W. Wang, *Ranking outliers using symmetric neighborhood relationship*, In: Pacific-Asia Conf. Knowl. Discovery Data Min., Berlin,2006, pp. 577-593.
[11] L. J. Latecki, A. Lazarevic and D. Pokrajac, *Outlier detection with kernel density functions*, In: Proc. Int. Conf. Mach. Learn. Data Min. Pattern Recognit. , 2007 pp. 61-75 .
[12] B. F. J. Manly and J. A. N. Alberto, *Multivariate statistical methods: a primer*, Chapman and Hall/CRC, 2016.
[13] S. Papadimitriou, H. Kitagawa, P. B. Gibbons and C. Faloutsos, *Loci: Fast outlier detection using the local correlation integral*, In: Proc. 19th int. conf. data eng., Cat. No. 03CH37405, 2003, pp. 315-326 .
[14] S. Ramaswamy, R. Rastogi and K. Shim, *Efficient algorithms for mining outliers from large data sets*, ACM Sigmod Record, 29 (2)(2000) 427-438.
[15] S. Shekhar, C. T. Lu and P. Zhang, *Detecting graph-based spatial outliers*, Intell. Data Anal., 6(5)(2002) 451-468.
[16] B. Tang and H. He, *A local density-based approach for outlier detection*, Neurocomputing, 241(2017) 171-180.
[17] A. Wahid and A. C. S. Rao, *Rkdos: A relative kernel density-based outlier score*, IETE Technical Rev., 37 (5)(2020) 441-452.
[18] X. Xu, H. Liu, L. Li and M. Yao, *A comparison of outlier detection techniques for high-dimensional data*, Int. J. Comput. Intell. Syst., 11 (1)(2018) 652-662.
[19] L. Zhang, J. Lin and R. Karim, *Adaptive kernel density-based anomaly detection for nonlinear systems*, Knowledge-Based Syst., 139(2018) 50-63.