Using method of Nadaraya-Watson kernel regression to detection outliers in multivariate data fusion

Omar A. abd Alwahab[*1]

[1] Statistic Department, College of Administration and Economics, University of Diyala, Iraq
Email: omaradil.d87@gmail.com

Abstract

In this paper, the researcher discussed a developed approach to the detection of outliers that is suited to multivariate data fusion. The challenge in outlier detection when dealing with multivariate data it is the detection of the outlier with more than two dimensions. To address this issue, the researcher developed a method to detect anomalies using methods based on local density including comparing a specific observations density with the densities of its neighboring observations. To make such comparisons, the researcher often employs an outlier score. In this study, various density estimation functions and distance metrics were utilized. Nadaraya-Watson kernel regression for multivariate data considered the KNN with multivariate data. Finally, the estimate of the Volcano kernel method is an essential method for outliers detection. In the simulation experiments of multivariate data with (4,6,8) variables and (60,120,180) observations, the results of simulation experiments by using the criterion of the precision evaluation showed that the N-W method is better than the VOL method in outlier detection in multivariate data.

Keywords: K-nearest neighbor; density of kernel function; outlier score; N-W regression; Volcano kernel method; data fusion.

1. Introduction

Detecting outliers in data by researchers can provide valuable information for making more informed decisions [1]. Failing to identify all outliers can result in inaccurate results, biased parameter estimation, and false assumptions [4]. Thus, it's crucial identifying the outliers before modeling and analysis [14]. At times, researchers may specifically focus on detecting outliers, such as credit card fraud detection, cybersecurity intrusion detection, and medical diagnosis [10]. In those cases, outliers may be integral data points, or the researcher may seek to remove them to clean the data [10]. Outlier definitions vary among several researchers [13]. Typically, an outlier is a data point that substantially deviates from other data points and appears suspicious [17]. Such observation might have been generated using a different mechanism than the rest of the data [15].

Numerous methods and approaches exist for detecting outliers, and they can be classified into different categories [14]. One way to categorize them is by differentiating between univariate and multivariate methods [1]. Historically, outlier detection research has predominantly focused on univariate methods [5]. However, this study concentrates on multivariate methods [12]. Another categorization of outlier detection methods is by learning methods, which can be classified into three scenarios: supervised, semi-supervised, and unsupervised learning methods [4]. Supervised learning means learning by example; this kind of learning examines training data and makes new functions based on function applications from training data [7]. Unsupervised learning seeks to discover hidden patterns in unlabeled data [17]. It cannot be used directly for a classification issue, as the output values are unknown [11]. Semi-supervised learning lies between the labeled data and unlabeled data [2]. Semi-supervised learning aims to determine how combining unlabeled, and some labeled input affects learning behavior [14]. The third categorization is parametric and non-parametric [8]. The parametric approach or the statistical method presupposes that the underlying distribution for the observation is known, often unsuitable for large datasets with several dimensions [2].

Doi: https://doi.org/10.54216/FPA.100207
Received: November 27, 2022 Accepted: March 11, 2023

2. Methods

2.1 Volcano kernel method

This function is presented to avoid the drawbacks in the Gaussian kernel for anomaly estimation [6]. We mean that in some methods that use a Gaussian kernel, we cannot guarantee that the normal data point is approximately equal to one for the outlier score, so we must use a threshold value (τ).

The volcano kernel is determined as follows [8]:

$$K(c) = \begin{cases} \beta & \text{if } \|c\| \leq 1 \\ \beta g(\|c\|) & \text{otherwise} \end{cases} \qquad (1)$$

Where β ensures that the kernel function is a probability density function, the condition of K(c) integration is equal to 1. g(c) is a function that decreases monotonically, with the close interval [0,1], and at infinity, equal to zero. The $g(c) = e^{-|c|+1}$ is a standard function in this method.

When we deal with univariate space, figure (1) explained the Volcano kernel function curve. Where β is a constant value in the kernel value when $\|c\| \leq 1$. For that, the outlier scores of samples within a group are close to 1. in case of ‖c‖ > 1, the Kernel value is less than one and decreases monotonically as ‖c‖ grows.

As a result, the outlier score for anomalies is substantially greater than one.

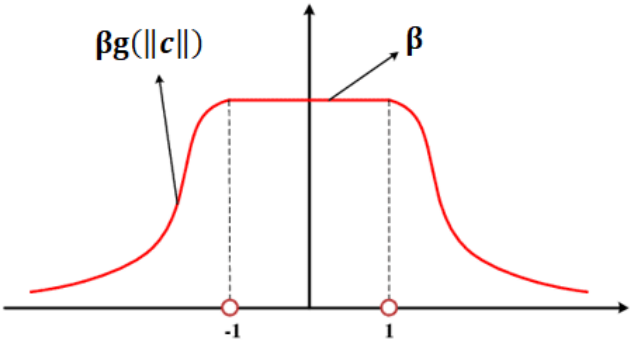

Figure 1: Volcano kernel curve in univariate data [8].

The Volcano kernel was created to identify anomalies. Its goal is to develop outlier scores for a normal sample close to 1, and for the anomaly, the outlier scores are more than 1. The number of k neighborhoods required in the VOL kernel is smaller than that needed for the Gaussian kernel. Regular samples take the wide plurality of the data set. Where ‖c‖ is a random variable that takes values between {-1,1} and figure (1) shows the densities of the VOL kernel. In the VOL kernel, the estimate of the density of a sample need lower neighboring samples than the Gaussian kernel [8].

2.2 Volcano kernel algorithm

1- Compute D (distance matrix) from C_{nxp} by Euclidean distance.

2- Squared proximities matrix $P^{(2)}$ where P is a matrix of the distance between each point.

3- Finding B where

Doi: https://doi.org/10.54216/FPA.100207
Received: November 27, 2022 Accepted: March 11, 2023

$$B = -0.5 * J * P^{(2)} * J \quad (2)$$

4- Find the r largest positive eigenvalues $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_r$ of matrix β and the congruous r eigen vectors $e_1, e_2, e_3, \ldots, e_r$

5- derive from the coordinate matrix r multivariate spatial of the n objects disposition

$$Z = e_r * \Lambda^{1/2} \quad (3)$$

Where

e_r is the matrix of r eigenvectors.

Λ_r is the diagonal matrix of r eigenvalues of β.

6- Let Z_1 and Z_2 are first and second dimensions from Z where Z_1 and Z_2 has the largest positive eigenvalues.

7- Calculate all the distances between each c_1 and c_2 data point.

8- Compute KNN.

9- Compute $kde(c_1) = \frac{\sum_{c_2 \epsilon N_k(c_1)} \frac{1}{(C.k-distance(c_2)^{\alpha})^2} K\left(\frac{c_1 - c_2}{C.k-distance(c_2)^{\alpha}}\right)}{|N_k(p)|} \mathcal{V}$ (4)

10- Compute $w_{c_2} = exp\left\{ - \frac{\left(\frac{k-distance(c_2)}{min_k} - 1\right)^2}{2\sigma^2} \right\}$ (5)

11- Compute $wde(c_1) = \frac{\sum_{c_2 \epsilon N_k(c_1)} w_{c_2} \cdot kde(c_2)}{\sum_{c_2 \epsilon N_k(c_1)} w_{c_2}} \; w_{c_2}$ (6)

3. The kernel regression of Nadaraya-Watson

Data collection d for this method will be $\{(x_1, c_1), \ldots, (x_N, c_N)\}$ Where $F(x, c)$ is the joint pdf, the function of the regression of X on C [5]

$$m(c) = E(X|C = c)$$

$$m(c) = \frac{\sum_{i=1}^{N} \frac{1}{\lambda_i^{\gamma}} K\left(\frac{c - c_i}{\lambda_i}\right) x_i}{\sum_{i=1}^{N} \frac{1}{\lambda_i^{\gamma}} K\left(\frac{c - c_i}{\lambda_i}\right)} \quad (7)$$

Where

λ_i: the adaptive kernel width (k-distance between points).

γ: The sensitivity parameter is set to two by default since a higher gamma value in the local density estimate results in a more sensitive KDE (c) for Multivariate data, which is not a desirable property for outlier detection. Where K(∗) is the multivariate kernel.

The estimator of Nadaraya-Watson calculates the regression coefficient for each data point (x,c) based on a weighted average to $\{c_1, \ldots, c_N\}$. the multivariate kernel function is used to calculate the weight.

The estimator of local kernel regression is based on the kernel regression of Nadaraya-Watson which is calculated within a k-distance radius of a particular data point.

The estimator local kernel regression for a data point p is given by :

$$m(c_1) = \frac{\sum_{q \epsilon N_k(p)} \frac{1}{(k-distance(q))^{\gamma}} K\left(\frac{c_p - c_q}{k-distance(q)}\right) x_p}{\sum_{q \epsilon N_k(p)} \frac{1}{(k-distance(q))^{\gamma}} K\left(\frac{c_p - c_q}{k-distance(q)}\right)} \quad (8)$$

Doi: https://doi.org/10.54216/FPA.100207
Received: November 27, 2022 Accepted: March 11, 2023

Where

q: is equal to the data point the k-distance neighbor to the point p.

x_p: outlier vector of point q.

In general, there are differences between the estimator of local kernel regression and the estimator of kernel regression of Nadaraya-Watson in the below:

1- The regression estimator for the point p calculated locally in the k-distance radius for p compared to the Nadaraya-Watson kernel regression global calculation, local computation reduces calculation is greatly complicated.

2- Parameter γ equal to two, but in Nadaraya-Watson kernel regression, γ is set to the number of dimensions is d. in high dimensional data k-distance very small that, $(k-distance)^d$ is approximately equal to zero when k-distance is big, this makes them, $(k-distance)^d$ very big.

3- The bandwidth control by λ_i for point i, Nadarya-Watson regression turns into the k-distance for the estimator of local kernel regression. The use of the k neighborhood distance of each data point to control the bandwidth size to ensure the adaptive bandwidth is adjusted while the choosing of λ is avoided in the estimator of the Nadaraya-Watson kernel.

The determination of local kernel regression is found by the kernel K(∗) as equation (8). That follows the prerequisite for the kernel function in local regression for local outlier detection, It should efficiently show higher outlier factors for isolated outliers that are easily discovered by local outlier detection methods.

Additionally, the function of the local kernel should be able to take into account the relationship between a datapoint and its k neighbor's distance when determining the weight of each neighbor.

The kernel function in this method the function of multivariate local kernel regression is the Gaussian kernel

$$K(\frac{c_i-c_j}{h_j})_{Gaussian} = \frac{1}{(2\pi)^d} exp\left(-\frac{\|c_i-c_j\|^2}{2*h_j^2}\right) \quad (9)$$

Where

$K(*)$ is the Gaussian kernel function, and K(x) integral is equal to 1.

The kernel smooth function is requisite to find the smoothness in the estimation of density. The smooth kernel function can be shown:

$$\int K(x)dx = 1$$

$$\int x\,K(x)dx = 0 \quad (10)$$

$$\int x^2K(x)dx > 0$$

Doi: https://doi.org/10.54216/FPA.100207
Received: November 27, 2022 Accepted: March 11, 2023

Assuming that the amount of $k - distance(c_2)$ is fixed. According to multivariate local kernel regression. Where the data point c_1 weight of for k neighbor distance c_2 is computed as :

1) If c_1 is also a k-distance neighborhood for c_2 then $(\|c_1 - c_2\| \leq d_k(c_2)$.
2) If c_1 is not a k-distance neighborhood for c_2 then $(\|c_1 - c_2\| > d_k(c_2)$.

This method's general framework:

Step one:

According to the equation:

$$\{c_1, \dots, c_N\} \overset{\tilde{F}(c)}{\Longrightarrow} \{(x_1, c_1), \dots, (x_N, c_N)\} \quad (11)$$

, process the data. In this step, the parameter k does not affect the detection performance.

Step two:

Calculate the parameter k range for estimating the local kernel regression, this range's lower and higher bounds respectively are denoted by k_{min} and k_{max}. The method performs well in a specific range of k points for a given data set. The interval $[k_{min}, k_{max}]$ must intersect with a specific range of the density based on equation (11).

Step three:

The estimator of regression that was computed in this interval $[k_{min}, k_{max}]$

Let $T = \frac{k_{min} - k_{max}}{\alpha}$ Where α: is the transition factor. The factors of outlier calculated in the preprocessing stage are denoted by $\{x_1^0, \dots, x_N^0\}$, for t = 1,2,3,…, T

3) Compute the k neighborhoods distance for all of the data points, where
$$k = k_{min} + \alpha.(t-1)$$
4) Determined $\{x_1^t, \dots, x_N^t\}$ using the estimator of local kernel regression together with $\{x_1^{t-1}, \dots, x_N^{t-1}\}$ whether
$\left\{\left\|\frac{c_p - c_q}{k-distance(q)}\right\| > 1, \forall\, q \epsilon\, N_k(p)\right\}$, the outlier factor $x_p^t = x_p^{t-1}$.
5) Let £ $= \sum_{i=1}^{N} |c_i^t - c_i^{t-1}|$, if £ $< \varepsilon$ where ε is the specified threshold or number of iterations, the procedure should be stopped and the output should equal $\{x_1^t, \dots, x_N^t\}$.

Step four:

the outlier factors output.

$$\{OF(c_i) = x_i^t | i = 1,2, \dots, N\}$$

3.1 N-W kernel regression algorithm

1- Compute D (distance matrix) from C_{nxp} by Euclidean distance.

2- Squared proximities matrix $P^{(2)}$ where P is a matrix of the distance between each point.

3- Finding B where

Doi: https://doi.org/10.54216/FPA.100207
Received: November 27, 2022 Accepted: March 11, 2023

$$B = -0.5 * J * P^{(2)} * J \quad (12)$$

4- Find the r largest positive eigenvalues $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_r$ of matrix β and the congruous r eigen vectors $e_1, e_2, e_3, \ldots, e_r$

5- derive from the coordinate matrix r multivariate spatial of the n objects disposition

$$Z = e_r * \Lambda^{1/2} \quad (13)$$

Where

e_r is the matrix of r eigenvectors.

Λ_r is the diagonal matrix of r eigenvalues of β.

6- Let Z_1 and Z_2 are the first and second dimensions from Z, where Z_1 and Z_2 have the largest positive eigenvalues.

7- Calculate all the distances between each c_1 and c_2 data point.

8- Compute KNN.

9- Compute Nadaraya-Watson kernel regression: $m(c_1) = \frac{\sum_{q \epsilon N_k(p)} \frac{1}{(k-distance(q))^{\gamma}} K\left(\frac{c_p - c_q}{k-distance(q)}\right) x_p}{\sum_{q \epsilon N_k(p)} \frac{1}{(k-distance(q))^{\gamma}} K\left(\frac{c_p - c_q}{k-distance(q)}\right)}$

(14)

4. Criterion

The criteria measure that deals with the efficiency performance of the outlier approach in this study are precision criteria, and the criteria used are [15]:

4.1 precision criteria:

Precision criteria can define as the percentage of dividing the correct outliers number by all points that are filtered to be outliers:

$$Pre = (v/N) * 100 \quad (15)$$

Where

v= correct outliers in the data set.

N= all points be outliers.

5. Concept of Simulation

Simulation can be defined as the process of a set of equations to represent a real phenomenon. Simulation is also expressed as a mathematical method that works to find similar data [4]. In this research, generating different sample sizes with different numbers of variables and the K-nearest neighborhood as a sample of the theoretical is designed to represent the community in the state of the real community [15].

6. Experimental design of generated data

In this study, data generated random numbers naturally subject to the normal distribution of $\sigma^2 = 0.5$ and mean = 0. The data was divided into different K-nearest neighborhoods with (3,4,..,,11) a number of variables (4, 6, and 8), and sample sizes (60, 120, and 180). The number of iterations for each is (itr =1000) according to Tabel (1). Many experiments were conducted.

Table 1**:** Variables and Samples sizes of generated data according to the K-nearest neighborhood

Doi: https://doi.org/10.54216/FPA.100207
Received: November 27, 2022 Accepted: March 11, 2023

Variables	**Samples sizes**			**K-nearest neighborhood**
4	60	120	180	3 4 5 6 7 8 9 10 11
6				
8				

7. Result

Table 2: shows 60 observations with 1000 replicates of the Average number of outliers and precision

Variables	**4**		**6**		**8**	
K	**NW**	**VOL**	**NW**	**VOL**	**NW**	**VOL**
3	16	15	15	33	17	36
4	15	15	17	31	16	31
5	15	14	15	29	17	32
6	17	15	15	29	15	30
7	13	13	12	27	13	28
8	14	13	10	29	12	30
9	13	14	9	28	12	28
10	12	14	10	27	12	29
11	14	13	15	26	14	27
Precision	26.40	28.13	27.69	28.58	22.18	32.89

Table 2 shows the results of (4,6 and 8) variables with 60 observations, of the VOL and NW methods, which were NW superior because the average number of outliers decreased as the number of nearest neighbors increased to 11. The Precision of NW has lower than VOL.

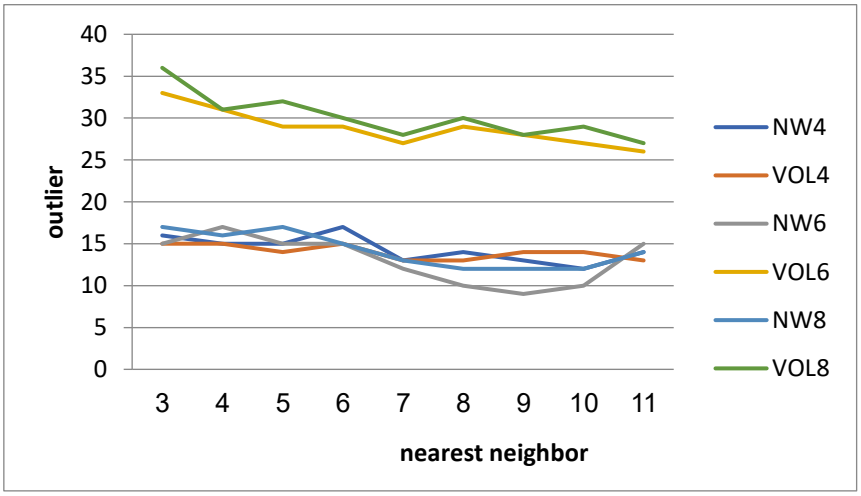

Figure 2: shows 4, 6 and 8 variables with 60 observations for NW and VOL

Table 3: shows 120 observations with 1000 replicates of the Average number of outliers and precision

Doi: https://doi.org/10.54216/FPA.100207
Received: November 27, 2022 Accepted: March 11, 2023

variables	4		6		8	
K	**NW**	**VOL**	**NW**	**VOL**	**NW**	**VOL**
3	33	70	32	70	34	74
4	34	62	35	64	31	64
5	31	60	33	64	33	65
6	28	58	36	60	30	69
7	25	60	29	58	28	68
8	23	60	24	59	23	68
9	19	59	23	57	19	65
10	20	59	22	55	20	63
11	21	58	21	55	18	58
Precision	23.07	27.04	23.53	24.49	17.91	29.62

Table 3 shows the results of (4,6 and 8) variables with 120 observations, of the VOL and NW methods, which were NW superior because the average number of outliers decreased as the number of nearest neighbors increased to 11. The Precision of NW has lower than VOL.

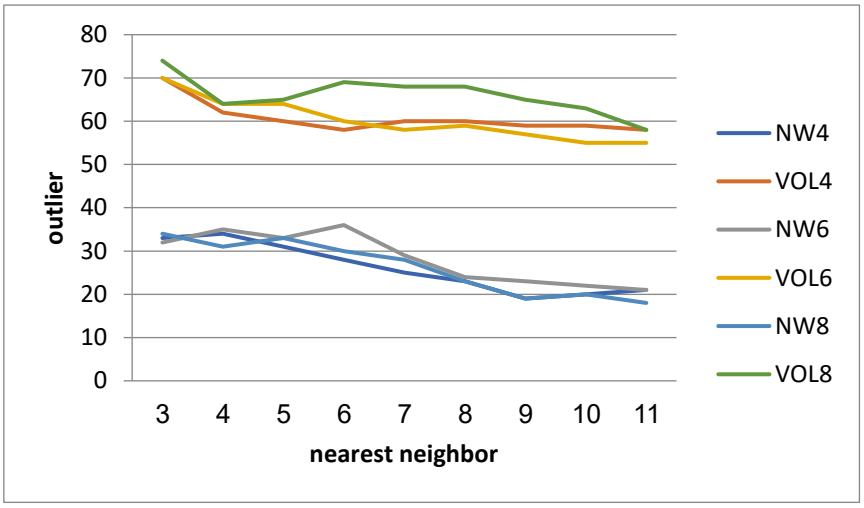

Figure 3: Shows 4, 6 and 8 variables with 120 observations for NW and VOL

Table 4: Shows 150 observations with 10.0 replicas of the Average number of outliers and precision

variables	4		6		8	
K	**NW**	**VOL**	**NW**	**VOL**	**NW**	VOL
3	51	102	51	112	53	100
4	53	100	57	96	57	94
5	59	96	57	93	50	91
6	57	91	52	88	46	89
7	54	91	45	86	39	86
8	52	90	39	83	34	83
9	49	84	42	82	31	81
10	46	83	38	80	29	79
11	42	84	36	79	26	78
Precision	26.68	26.61	27.39	22.56	27.32	21.39

Table 4 shows the results of (4, 6 and 8) variables with 180 observations, of the VOL and NW methods, which were NW superior because the average number of outliers decreased as the number of nearest neighbors increased to 11. The Precision of NW has lower than VOL

Doi: https://doi.org/10.54216/FPA.100207
Received: November 27, 2022 Accepted: March 11, 2023

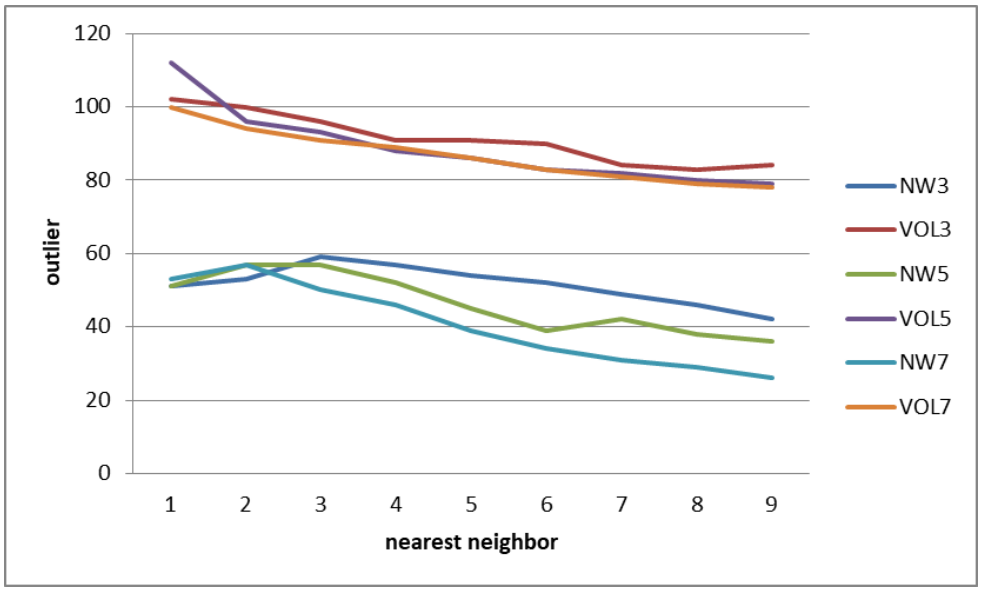

Figure 4: Shows 3, 5 and 7 variables with 150 observations for NW and VOL

Table 5: Shows results of 1000 replicate precision for all variables and observation

Variables	Methods	Observations		
		60	120	180
4	NW	26.4	23.07	26.68
	VOL	28.13	27.04	26.61
6	NW	27.69	23.53	27.39
	VOL	28.58	24.49	22.56
8	NW	22.18	17.91	27.32
	VOL	32.89	29.62	21.39

Doi: https://doi.org/10.54216/FPA.100207
Received: November 27, 2022 Accepted: March 11, 2023

Figure 5: Shows Precision for 4, 6 and 8 variables with 60,120 and 180 observations for NW and VOL

6. Conclusion

In this paper, the researcher discussed a new approach to the detection of outliers, which is suited to multivariate data fusion. From the experiment result, the researcher notice that increasing neighbor's points have a significant impact on outlier's detection, especially in the method of (NW), but the method of VOL was less impact on outlier's detection. From that, it was clear the kernel function that used for multivariate data fusion is better when comparing the results. The Euclidean distance with variables for the (NW) and (VOL) was used. In addition, the outlier's average number in (VOL) is great than the outlier's average in the Euclidian distance. The increase in neighbors' number leads to a decrease in outliers' number.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] B. Tang, and H. He, "A local density-based approach for outlier detection," Neurocomputing, Vol. 241, pp. 171–180, Jun. 2017.
[2] Hu, W., Gao, J., Li, B., Wu, O., Du, J., & Maybank, S. (2018). "Anomaly detection using local kernel density estimation and context-based regression." IEEE Transactions on Knowledge and Data Engineering, 32(2), 218-233.
[3] F. E.Grubbs, "Procedures for detecting outlying observations in samples." Technometrics Vol. 11, no.1, pp 1-21.1969
[4] Fan, H., Zaïane, Foss, O. R., A., and J. Wu, "A nonparametric outlier detection for effectively discovering top-n outliers from engineering data,". In Pacific-Asia conference on knowledge discovery and data mining Springer, Berlin, Heidelberg, pp. 557-566, 2006.
[5] Gao, J., Hu, W., Li, W., Zhang, Z., and O. Wu, "Local outlier detection based on kernel regression," In 2010 20th International Conference on Pattern Recognition, pp. 585-588, IEEE.2010 .
[6] Gao, J., Hu, W., Zhang, Z. M., Zhang, X., and O. Wu, "RKOF: robust kernel-based local outlier detection," In Pacific-Asia conference on knowledge discovery and data mining. Springer, Berlin, Heidelberg, pp.270-283, 2011.
[7] Hawkins D M, "Identification of Outliers", Chapman and Hall., London, Vol 11, 1980.
[8] Hu, W., Gao, J., Li, B., Wu, O., Du, J., & Maybank, S.. Anomaly detection using local kernel density estimation and context-based regression. IEEE Transactions on Knowledge and Data Engineering, 32(2), 218-233, 2018.
[9] Latecki, L.J. Lazarevic, A. and D. Pokrajac, "Outlier detection with kernel density functions," in Proc. of International Conference on Machine Learning and Data Mining in Pattern Recognition, pp. 61-75, 2007.
[10] L. Zhang, J. Lin, and R. Karim, "Adaptive kernel density-based anomaly detection for nonlinear systems," Knowledge-Based Systems, Vol. 139, pp. 50–63, 2018.
[11] Manly, Bryan FJ, and Jorge A. Navarro Alberto. Multivariate statistical methods: a primer. Chapman and Hall/CRC, 2016.
[12] Zhang, L., Lin, J., & Karim, R. (2018). "Adaptive kernel density-based anomaly detection for nonlinear systems." Knowledge-Based Systems, 139, 50-63.
[13] Papadimitriou, S., Kitagawa, H., Gibbons, P. B., & C. Faloutsos, "Loci: Fast outlier detection using the local correlation integral," Proceedings 19th international conference on data engineering., Cat. No. 03CH37405, pp. 315-326, IEEE, 2003.
[14] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," ACM Sigmod Record, Vol. 29, no. 2, pp. 427–438, May 2000.
[15] V. A. Epanechnikov, "Non-parametric estimation of a multivariate probability density," Theory of Probability & Its Applications, Vol. 14, no. 1, pp. 153–158, 1969.
[16] W. Jin, A. K. H. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in Pacific-Asia Conference on Knowledge Discovery and Data Mining, Berlin, pp. 577–593, 2006.

Doi: https://doi.org/10.54216/FPA.100207
Received: November 27, 2022 Accepted: March 11, 2023

[17] X. Xu, H. Liu, L. Li, and M. Yao, "A comparison of outlier detection techniques for high-dimensional data," International Journal of Computational Intelligence Systems, Vol.11, No. 1, pp. 652-662, 2018.

Doi: https://doi.org/10.54216/FPA.100207
Received: November 27, 2022 Accepted: March 11, 2023