Outlier detection algorithm based on k-nearest neighbors-local outlier factor

He Xu¹,², Lin Zhang¹, Peng Li¹,² and Feng Zhu¹,²

Abstract
The main task of outlier detection is to detect data objects which have a different mechanism from the conventional data set. The existing outlier detection methods are mainly divided into two directions: local outliers and global outliers. Aiming at the limitations of the existing outlier detection methods, we propose a novel outlier detection algorithm which is named as kNN-LOF. First, the k-nearest neighbors algorithm is applied to divide different areas for outlier attributes, which is more suitable for outlier detection in different density distributions. Secondly, a hierarchical adjacency order is proposed to hierarchize the neighborhood range according to the link distance. The average sequence distance is calculated from the data objects in the hierarchy, and the reachable distance of an object is redefined to introduce a new local outlier factor. Experimental results show that the proposed algorithm has good performance in improving the accuracy of outlier detection.

Keywords
Outlier detection, k-nearest neighbor, local outlier factor

Received 21 May 2021; Revised received 13 October 2021; accepted 22 November 2021

Introduction
Data mining is a kind of data analysis technology to find general data patterns with potential value in data. Outliers are relatively sparse isolated points. As one of the tasks of data mining, outlier detection is to find "outliers" from a large amount of data which has different characteristics among other data. It is widely used in network intrusion detection, fault diagnosis, health monitoring, satellite image analysis and so on.¹

Outlier detection was used for the data preprocessing in database knowledge discovery. The data was cleaned to find and remove outliers from the original data in order to improve the accuracy of data mining. With the era of big data, the actual value and significance of outlier detection become important. Analyzing the existence significance and outlier attributes is the main research hotspot. According to the difference from the overall data set, outliers are divided into global outliers and local outliers. According to the size of outliers, the outliers are divided into cluster points and outlier clusters.² Current outlier detection techniques are mainly divided into statistical-based methods, cluster-based methods, distance-based methods, and density-based methods.

The outlier detection method started from the statistical learning-based outlier detection method,³ which assumes that the data obey a certain regular distribution. It creates a probability distribution function model for the data set by estimating the parameters of the probability distribution to find out the data which deviates from the statistical distribution curve is outlier data, when the occurrence probability is lower than the threshold are outliers. The existing statistical learning is divided into research methods based on distribution and depth. The distribution-based outlier detection method is based on statistical theory. After selecting the distribution model according to the characteristics of the data, the outlier is found by fitting the probability model. The most commonly used is the Gaussian distribution model. However, this method requires a high level of user’s prior knowledge. It is necessary to know the characteristics of the data in advance in order to select a suitable distribution model. However, in practical applications, the data is

¹School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China
²Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, China

Corresponding author:
He Xu, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, 210023, China.
Email: xuhe@njupt.edu.cn

Creative Commons Non Commercial CC BY-NC. This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).
unknown and complex. The data is mainly multi-dimensional. Thus, step-based detection methods show huge limitations and shortcomings. The depth-based outlier detection method is based on the improvement of the step-based detection methods. Based on the theory of computer geometry, the k-d convex hull of each layer is calculated. A depth is given for each data point, and the depth of the data object is assigned according to the degree of deviation from the center value. The data objects are hierarchically organized in the data space, and the outlier detection is performed on the level with a smaller depth value. However, the time complexity of this method is very high, so it is not suitable for high-dimensional data.

The cluster-based outlier detection method examines the relationship between objects and clusters. The purpose of the clustering algorithm is to divide objects into subsets so that objects with the same characteristics are grouped together. Outliers are “isolated object” which is a distant object in the feature space. The cluster-based detection method is to detect objects that do not belong to any cluster or a small cluster. This method focuses on the overall distribution of the data and performs outlier detection after clustering the data set. Zhu et al. proposed a minimum spanning tree outlier detection method based on fast k-nearest neighbors (kNNs), which is an algorithm that combines density and clustering. The algorithm proposes a new data structure to divide the dimensions of the data set to find the kNN of the data point. The number of clusters is determined by calculating the edge threshold and the cluster splitting threshold. Finally, several clusters are formed and the clusters whose data points are significantly smaller than other clusters are output as outliers. The algorithm is good at performing on circular data sets, but the algorithm requires more parameter settings when determining the number of clusters. Such methods often focus on detecting small clusters, which will affect the accuracy in the detection of outliers and have a strong dependence on parameter settings.

The concept of distance-based outlier definition was first proposed by Knorr and Ng: An object O in a dataset T is a DB(p, D)-outlier if at least fraction p of the objects in T lies greater than the distance D from O, which indicates that the object is larger than p in the database are more than D which is away from the object O, then the object O is an outlier. This algorithm proposes an index-based algorithm and a nested loop algorithm, which are suitable for those without a standard distribution model dataset. However, such algorithms are not suitable for high-dimensional data because the time complexity is high. Ramaswamy et al. further modified the definition of outliers. The method is based on the distance between the object and its kNNs: given k and n, a point is an outlier if no more than n−1 other points in the data set have a higher value for D than p. By calculating the distance from an object to its neighbors and sorting, the object with the largest value in the order is marked as an outlier. This method does not require the user to specify a distance parameter for the definition of outliers, so the data has strong independence. The experimental results also verify the method has independent of parameters. In addition, in order to find outliers, a partition-based outlier mining algorithm is proposed. The clustering algorithm is used to block the data and calculate the upper and lower bounds of each block area. Outliers are calculated in the non-run partition. This algorithm has good scalability in both the size and the dimension of the data set, which saves the amount of calculation. Angiulli and Pizzuti proposed a new definition of distance-based outliers and the proposed algorithm called HiOut, which aims to effectively detect outliers in large and high-dimensional data sets. The weight of a data point is defined as the sum of its distance from the kNNs. The point with the largest weight is the outlier. The calculation of weights tends to reduce the efficiency of the entire algorithm. They define a set of approximate outliers so that the weight of the data points is greater than the weight of the true outliers in a small range. The set of candidate points is true outliers. Single outlier detection has gradually failed to meet the complexity of the data. Jiang et al. proposed a boundary and distance-based outlier detection, which combines the concept of rough concentrated boundaries with the distance-based outlier detection method proposed by Knorr et al. In distance-based outlier detection, whether the object is an outlier is determined by calculating the distance between the object in the data set and each object in the abnormal boundary, B-primary boundary, and B-lower approximation.

Density-based outlier detection is proposed to overcome the shortcomings of distance-based detection of global outliers. Breunig first proposed the concept of the Local Outlier Factor (LOF). A local outlier coefficient is assigned to the data object to indicate the degree of an outlier of the object relative to its neighborhood. The calculation method of LOF is supported by a series of definitions. It examines the ratio of the local density around an object to the average reachable density of its neighborhood objects, where the neighborhood of the object is determined according to the minimum neighbor parameter k given by the user and the nearest neighbor distance. A LOF value which is close to 1 indicates that the object is in a region with a relatively uniform density. A larger LOF value indicates a higher probability that the object is an outlier. This method differs in that the determined outlier is a degree of isolation relative to the local neighborhood density of the object, so it is a local outlier. The restricted neighborhood of each object is considered and the degree of an outlier of the object is quantified. Since the outlier factor was proposed, many measurement methods have emerged. Tang et al. proposed a connectivity-based
Outlier Factor (COF), which calculates the ratio of the chain distance of a data point to the average chain distance of all nearest neighbors at that point to define the outlier factor. The weighted subspace outlier detection algorithm SPOD of local information entropy proposed by Ni et al.\textsuperscript{13} through the neighborhood information entropy analysis of the data object in each dimension, generates the corresponding outlier subspace and attribute the weight vector of the data object. The attributes in the outlier subspace are given higher weights, and the concepts of subspace weighted distance are further proposed. Tang and He\textsuperscript{14} proposed a simple and effective density-based outlier detection method based on local kernel density estimation. The author introduces the relative density-based outlier (RDOS) to measure the local outlier of the object and uses the local KDE method to estimate the density distribution (RDOS) to measure the local outlier of the object and uses the local KDE method to estimate the density distribution of the object position. Hu and Qin\textsuperscript{15} proposed a density-based outlier detection algorithm DLOF, which uses information entropy to determine the outlier attributes of each object. When calculating the distance between objects, the weighted distance is used to improve the accuracy of outlier detection.

The difference between density-based detection methods lies in the difference between the way to determine the local neighborhood and to calculate outliers. Tang et al.\textsuperscript{12} redefine outliers. Ni et al.,\textsuperscript{13} Tang and He,\textsuperscript{14} and Hu and Qin\textsuperscript{15} make further improvements. The calculation of the outlier factor is to find the local density of its neighborhood objects. The controllability of the neighborhood range is determined by the $k$ parameter, which is full of variable factors and uncertainties. The effect of different parameter settings on the results cannot be ignored. The selection of neighborhood objects is related to the distance-based outlier detection algorithm. For this reason, the joint improvement of the determination of the local neighborhood and the calculation of the outlier coefficient is also a fusion of distance and density methods. An outlier detection algorithm consisting of outlier factors with multiple conditional mixing parameters has become an important research direction in this field. Liu\textsuperscript{16} proposed a two-parameter-based outlier detection algorithm-DDPOS algorithm, which uses the adaptive Gaussian kernel density estimation function to calculate the kernel density of the object, thereby obtaining the local density estimated by the exponential kernel density. Each data object in descending order is sorted according to the local density to get the set of local density of all data objects in the data set. Calculate the average distance between the object and the first $k$ denser than its larger object is the global distance. It will investigate the degree of deviation between each object point to get the global distance set for each object in the data set. The newly defined outlier factor is the ratio of the global distance to the local density of each object in the data set to determine whether the object is an outlier. Wang et al.\textsuperscript{17} proposed a weighted outlier factor (DRWROF) algorithm based on distance ratio. It is defined as the product of the ratio of the distance from the object to its $k$NN to the average distance of the $k$NN and the weight level in the distance from the object to its $k$NN. It is also a new algorithm for calculating outlier factors.

The contribution of this paper is as follows: This paper proposes an outlier detection algorithm based on KNN-LOF. First, the data set is divided into different regions by $k$NN to accurately query the scope of the neighborhood. Secondly, the average sequence distance is introduced to assign different weights to the data of different neighborhood ranges. By considering the degree of proximity of points in different neighborhoods to data objects, the influence of different adjacent ranges can be distinguished, so as to further accurately calculate the outlier coefficient.

**Materials and methods**

**k-Nearest neighbors**

Suppose $D$ is a data set and $p$ is any data point in $D$, $p \in D$. The distance from $p$ to any point $o$ in $D$ is called $d(p, o)$. The local neighborhood and the $k$-Nearest Neighbor are standardized by Breunig.\textsuperscript{11}

**Definition 1 (local neighborhood):** The $r$ is a real number and $r \geq 0$, the $r$-neighborhood range of the data point $p$ is denoted as $N(p,r)$ and defined as

$$N(p,r) = \{o \in D | d(o,p) \leq r\} \quad (1)$$

**Definition 2 (k-nearest neighbor):** For any positive integer $k$, the $k$NNs of the data point $p$ in $D$ are composed of all objects in $D$ whose distance from the $p$ point does not exceed the threshold $k$, which is recorded as $N_k(p)$, defined as

$$N_k(p) = \{ o \in D | d(o,p) \leq k, o \neq p \} \quad (2)$$

Local neighborhood and $k$NN were the two kinds of measurement benchmarks for distance-based outlier detection, both of which depend on the setting of parameter $k$.

**Local outlier factor**

The basic idea of the LOF is to define the degree of an outlier of the point by assigning an object’s deviation factor in the data set. It is not a clear definition of which data points are outliers. In essence, the point is judged from the region whether the overall layout is in a more concentrated area. The calculation method of LOF is supported by a series of definitions.

**Definition 3 (k-distance of object $p$):** For any positive integer $k$, the $k$th distance of the data object $p$ is marked
as \( k \)-distance\( (p) \). In dataset \( D \), the distance between \( p \) and \( o \in D \) is written as \( d(p, o) \).

Let \( d(p, o) \) be \( k \)-distance\( (p) \):

For at least \( k \) objects \( o' \in D \setminus \{p\} \), satisfy \( d(p, o') \leq d(p, o) \);
For up to \( k-1 \) objects \( o' \in D \setminus \{p\} \), satisfy \( d(p, o') < d(p, o) \).

From Definition 3, the \( k \)-distance of the object \( p \) is the distance between \( p \) and the farthest point among its \( k \)NNs.

**Definition 4** (\( k \)-distance neighborhood of object \( p \)): Knowing the \( k \)-distance of the data object \( p \), the \( k \)-distance neighborhood of the data contains all data objects whose distance to the data object \( p \) is not greater than the \( k \)-distance of \( p \), denoted as \( N_{k\text{-distance}}(p) \), abbreviated as \( N_k(p) \), let \( k \)-distance\( (p) \) be simply \( k \)-distance\( (p) \):

\[
N_k(p) = \{q \in D \setminus \{p\} | d(p, q) \leq \text{k-distance} (p) \} \quad (4)
\]

These objects \( q \) are called \( k \)-distance neighborhoods of \( p \). The neighborhood range includes all data objects in the area with the data object as the center and the \( k \)-distance of the object as the radius. In other words, the set must contain at least \( k \) data objects. In general, the number of data objects contained will not be much more than \( k \).

**Definition 5** (reachable distance of object \( p \) with respect to object \( o \)): Let \( k \) be a natural number, then the reachable distance of the data object \( p \) relative to the object \( o \) is written as

\[
\text{reach-distance}(p, o) = \max \{k\text{-distance}(o), d(p, o)\} \quad (4)
\]

That is, when the object \( p \) exceeds the \( k \)-distance neighborhood of the object \( o \), the reachable distance between \( p \) and \( o \) is the actual distance \( d(p, o) \). The \( k \)-distance of the reachable distance between \( p \) and \( o \) is denoted by \( k \)-distance\( (o) \), when \( p \) is sufficiently close to \( o \) and in the \( k \)-distance neighborhood of \( o \).

**Definition 6** (local reachable density of object \( p \)): Set \( k \) as a positive integer parameter of the data object. In this paper, we focus on specific instances of \( k \) to define outliers. Specify the minimum number of neighbors of the object under investigation with the parameter \( \text{MinPts} \), and set \( \text{MinPts} \) to a positive integer. The local reachable density of \( p \) can be expressed as

\[
\text{lrd}_{\text{MinPts}}(p) = \frac{1}{\sum_{o \in N_{\text{MinPts}}(p)} \text{reach-distance}(p, o) / |N_{\text{MinPts}}(p)|} \quad (5)
\]

By this definition, the denominator of formula (5) is the average reachable distance of the smallest neighborhood of \( p \). The calculation of the local reachable density first obtains the sum of the reachable distances of all data objects in the \( k \)th neighborhood of the data object \( p \) to the point \( p \). Reach-distance\( (p, o) \) will take \( d(p, o) \) if the deviation of \( p \) is larger. Fewer data objects in its \( k \) neighborhood will result in \( |N_{\text{MinPts}}(p)| \) and lower reachable densities. Reach-distance\( (p, o) \) will take \( k \)-distance\( (o) \) if \( p \) is in a more clustered group, it will cause the local reachable density values of all data objects in the area to be close and large. Therefore \( \text{lrd}_{\text{MinPts}}(p) \) can effectively characterize the density of the local area of the data object.

**Definition 7** (local outlier for object \( p \)):

\[
\text{LOF}_{\text{MinPts}}(p) = \frac{\sum_{o \in N_{\text{MinPts}}(p)} \frac{\text{lrd}_{\text{MinPts}}(o)}{\text{lrd}_{\text{MinPts}}(p)}}{|N_{\text{MinPts}}(p)|} = \frac{\sum_{o \in N_{\text{MinPts}}(p)} \frac{\text{lrd}_{\text{MinPts}}(o)}{|N_{\text{MinPts}}(p)|}}{|N_{\text{MinPts}}(p)|} \quad (6)
\]

From this definition, \[ \sum_{o \in N_{\text{MinPts}}(p)} \frac{\text{lrd}_{\text{MinPts}}(o)}{|N_{\text{MinPts}}(p)|} \] is the average density of the data objects in the \( \text{MinPts} \) neighborhood of object \( p \), and \( \text{LOF} \) is the ratio of it to the local reachable density, which reflects the degree of an outlier of \( p \). That is, a high \( \text{LOF} \) value indicates that the corresponding data object is more likely to be an outlier.

**Outlier detection algorithm based on kNN-LOF**

The continuous improvement of the density-based outlier detection method is mainly reflected in two aspects. One is the determination of local neighborhoods. The other is the calculation of the outlier factor. In order to improve the accuracy of outlier detection, this paper makes improvements in two aspects. Firstly, the kNN algorithm is introduced to determine the \( k \)-distance neighborhood of the data object \( p \) more accurately; Secondly, when calculating the distance between the data object and the objects in its neighborhood, the sequence distance weighting calculation method is used to determine the weighted distance among the data objects.

**Related concepts of algorithms**

In the traditional density-based outlier detection algorithm, the \( \text{LOF} \) calculation may be inaccurate for some outliers in the case of special density distribution. Compared with a sparse area, a data point close to a denser area may have a higher \( \text{LOF} \) value calculation result. Because the average density of data objects in the latter neighborhood leads to a high \( \text{LOF} \) ratio, it is easy to cause misjudgment of outliers according to the traditional relevant definition standards. In addition, the search for outliers needs to involve all data points in the calculation of the outlier factor, which leads to the high time complexity of the algorithm.

It can be seen from Figure 1 that the outlier degree of point \( p \) is relatively small, and the point \( q \) has more...
outlier properties than point $p$. If the LOF algorithm is used for calculation, the neighborhood density of $p$ point is higher than the $K$ nearest neighbor density of $p$, which causes the LOF value of $p$ to be higher than that of $q$ point, which is more likely to be misjudged as an outlier. It can be seen that the existing algorithms are not effective in detecting data sets with irregular or uneven distribution. Density-based outlier detection can only exclude outliers with a lower density that are close to non-distance clusters, and the simple use of distance to measure the degree of outliers is slightly insufficient. In the traditional method, the kNN algorithm is used to determine whether the object is far away from most of the divisions. Euclidean distance or Manhattan distance can be used to calculate the distance between each object and other remaining objects, and the first $k$ smallest values are selected to calculate score. kNN is a classification algorithm. It finds the nearest kNNs in the training set, and the category determines the classification set of the sample to be recognized, which brings great inspiration to our proposed algorithm.

Therefore, this paper uses kNN to divide the effective range of the data set. We divide different areas for different outlier attributes and perform effective outlier detection in this area. The purpose of this method is to make the algorithm more suitable for outlier detection in different density distributions. The more rigorous the consideration of the neighborhood of the data point is, the more relativity and practical significance of the selected outliers are.

For point $p$ in Figure 2, we set $k=9$. Generally, there will not be a situation where the number of two categories is equal when $k$ is set to an odd number. We can see that there are more samples belonging to class A than B in the 8 sample points closest to the classification sample $p$. Therefore, we attribute the $p$ point to the class B sample set and calculate the local density of the point in the class B range. In the same way, the $q$ point is classified into the C-type sample set and the local density of the point is calculated in the C-type range.

We introduce the concepts of low density and isolation in this paper. The former means that the number of objects is the smallest in the neighborhood close to the object. The latter means the degree of connection between an object and other objects, and the relationship between isolation and low density is not absolute. The degree of isolation of an object can be quantified by its distance to the nearest neighbor. The closest path based on the set has been studied in the past based on the model of a single linear set and individual outliers to reflect the “shift from the model” in an appropriate way. The set-based closest index is proposed to trace all the closest neighbors.
this paper, we propose a hierarchical adjacency order, which aims to be suitable for various nonlinear low-dimensional structures. All the points in the neighborhood range in the traditional detection algorithm are "equally" involved in the conceptual hierarchy of the calculation process. According to the different link distances, the neighborhood range is hierarchized and the average sequence distance of data objects is calculated in different levels. While redefining the reachable distance of object \( p \) with respect to object \( o \), different weights are assigned to the neighborhoods of different \( k \) ranges. Considering the closeness between points in different neighborhoods and data objects, distinguishing the influence of different proximity ranges is to further accurately calculate the outlier coefficient to improve accuracy.

**Definition 8 (sequence division):** The calculation of the chain distance is based on a single chain set. Setting the fixed distance between each object point is too restrictive in practice, and iteratively expands when only a collection of objects is initially included. Find the nearest neighbors in the remaining data and add them to each iteration in turn. The time taken by this method increases with the increase of the data set when the data set is larger.

The division process is shown in Figure 3. Determining the range of a circle is based on taking the object \( p \) as the center and assigning different radius \( R \) and removing the intersection. In the data set \( D \), let \( O_1 \) be the set with a circle with \( R_1 \) as the radius and \( O_1 \in D \), \( O_1 = \{p_1, p_2, \ldots, p_m\} \), let \( O_2 \) be the

Figure 2. Schematic diagram of nearest neighbor method.

Figure 3. Schematic diagram of sequence division.
set with a circle with \( R_2 \) as the radius and \( O_2 \in D \), \( O_1 \cap O_2 = \emptyset \), \( O_2 = O_1 - O_1 \), \( O_2 = \{p_{m+1}, p_{m+2}, \ldots, p_n\} \), and let \( O_3 \) be the set with a circle with \( R_3 \) as the radius and \( O_3 \in D \), \( O_3 \cap O_2 = \emptyset \), \( O_3 = D - O_1 \cup O_2 \), \( O_3 = \{p_{m+1}, p_{m+2}, \ldots, p_n\} \). Among them, let \( R_3 = 2R_2 = 4R_1 \). \( O_1 \), \( O_2 \), and \( O_3 \) are the three different levels of division. And the weight of sequence distance is unique in the same level.

**Definition 9 (average sequence distance):** From the definition of the local reachable density of the object \( p \), we know that the distance between a data object and its kNN is inversely proportional to its density distribution. In traditional detection algorithms, the inaccuracy is due to the fact that all points in the neighborhood range “equally” participate in the calculation process. Based on the concept of sequence division, this paper calculates Table 1. The pseudo code of the kNN-LOF based outlier detection algorithm.

| Input: \( D, k \), and \( \delta \) |
|-----------------------------------|
| Output: The first \( m \) outliers of dataset \( D \) |
| 1 Collect \( p \) as the first training point; |
| 2 if \( d(p, q) \leq k\text{-dis}(p) \), the object \( q \) is the \( k \)-distance neighborhood of \( p \); |
| 3 Divide the effective range of the data set according to the kNN algorithm; |
| 4 Hierarchize the neighborhood range according to the link distance; |
| 5 Assign different weights to data objects on different levels based on the concept of level division; |
| 6 Calculate the sequence distance between data points and other objects according to formula (7); |
| 7 Use equation (10) to calculate the average sequence distance of each object; |
| 8 Introduce the concept of average sequence distance to calculate the new local outlier coefficient \( NLOF \); |
| 9 if \( NLOF_k(p) > 0 \), continue to the next training point; |
| 10 Sort \( NLOF \) in descending order; |
| 11 Output the first \( m \) outliers. |

**Figure 4.** Different results under \( k = 20 \): (a) 45 outliers under the LOF algorithm; (b) 59 outliers under the RDOS algorithm; (c) 62 outliers under the kNN-LOF algorithm; (d) 81 outliers under the COF algorithm.

LOF: Local Outlier Factor; RDOS: relative density-based outlier; kNN-LOF: k-nearest neighbors-local outlier factor; COF: connectivity-based Outlier Factor.
the average sequence distance by assigning different weights to data objects on different sequences.

\[ O_1 = \{ p_1, p_2, \ldots, p_r \} \] is a collection within a radius of \( a \). The sequence distance from \( p_1 \) to \( O_1 - \{ p_1 \} \) is

\[ \text{dis}_{R_1}(p_1) = \frac{1}{r} \sum_{i=2}^{r} \frac{r - i}{r} d(p_1, p_i) \] (7)

Because the degree of outlier of data objects in different sequences is different, the sequence distance is related to \( R_k \). \( k \) is a positive integer, and the average sequence distance of all objects in the neighborhood of the object \( p \) and its effective \( k \)-distance is defined as

\[ \text{avg} - \text{dis}_{R_k}(p) = \frac{1}{R_k} \sum_{o \in N_k(p)} \text{dis}_{R_k}(o) \] (8)

**Definition 10 (new local outlier factor):** The local neighborhood density of object \( p \) is the reciprocal of the average distance of the \( k \)NNs of object \( p \) and is defined as

\[ \text{ln}_{dk}(p) = \frac{1}{\text{avg} - \text{dis}_{R_k}(p)} \] (9)

The local neighborhood density of the data object in the effective influence space of object \( p \) is

\[ \text{ln}_{dk}(N_k(p)) = \frac{\sum_{o \in N_k(p)} \text{ln}_{dk}(o)}{|N_k(p)|} \] (10)

The local outlier coefficient of object \( p \) is redefined as:

\[ \text{NLOF}_{k}(p) = \frac{\text{ln}_{dk}(N_k(p))}{\text{ln}_{dk}(p)} \] (11)

Table 1 gives the pseudo code of kNN-LOF algorithm.

**Results**

In this section, the accuracy and efficiency of the algorithm proposed in this paper are evaluated through experiments. We will compare the proposed method with other detection methods. The results show that our method has better performance. The experimental platform tested was a CPU 1.70 GHz, a PC with 4 GB of memory, and an operating system of Windows 7.

The data set used in this experiment is the activity recognition system based on a multisensor data fusion (ARMe)

![Figure 5. Different results under k = 45: (a) 51 outliers under the LOF algorithm; (b) 38 outliers under the RDOS algorithm; (c) 101 outliers under the kNN-LOF algorithm; (d) 65 outliers under the COF algorithm. LOF: Local Outlier Factor; RDOS: relative density-based outlier; kNN-LOF: k-nearest neighbors-local outlier factor; COF: connectivity-based Outlier Factor.](image-url)
data set from the University of California Irvine, which includes six categories of data: bending, cycling, lying down, sitting, standing, and walking. This paper selects two types of data in the data set: sitting and walking. And the data that is relatively marginal or affected by environmental factors is deleted. The outliers data occupy 5% of the whole data set.

The sensitivity of different algorithms to the value of k

Because LOF, COF, RDOS, and KNN-LOF algorithms depend on the parameter of k value, the choice of k value directly affects the detection effect of the algorithm. The determination of the value of k is a key factor for the effective use of the algorithm in practice. In the practical application of outlier detection, the four kinds of algorithms have different adaptability to the k value.

In order to prove that the algorithm has different performance in outlier detection due to different k values, this paper sets different k values for detection according to different data set sizes. The results are shown in Figures 4 to 7.

Here are the results for sitting data. The blue points are the data points, and the red points are the outliers.

Here are the results for walking data.

It can be seen from the experimental results that the four algorithms show different effects on the detection of outliers. In general, outliers detected by the LOF algorithm increase with the increase of k value. The COF algorithm performs well when k value is small. For large data, the RDOS algorithm detects fewer outliers and deteriorates performance with the increasing value of K. For the walk data set, RDOS detects more outliers with the increase of k value. However, the performance of the kNN-LOF algorithm proposed in this paper is not as good as that of the COF algorithm while k value is small. However, with the increase of k value, its performance is better than the other three algorithms. It also shows a quick search rate for finding outliers.

ROC versus k

Detection rate and false alarm rate are usually two indicators for judging the pros and cons of outlier detection algorithms, but due to the conflict between the two indicators, a higher detection rate is generally accompanied by a higher
false alarm rate. So we use ROC curve to show the relationship. The ROC curve is the receiver operating characteristic curve, which is used to graphically display the relationship between the true rate and the false positive rate in the model. In this paper, outliers are expressed as positive and non-outliers are expressed as negative. The state and decision obtained are shown in Table 2.

Table 2. The relationship between state and decision.

| Status      | Decision making | Positive                     | Negative                     |
|-------------|----------------|------------------------------|------------------------------|
| Positive    | True positive (TP) | False positive (FP)         |
| Negative    | False negative (FN) | True Negative (TN)         |

The ROC curve is a curve made by drawing the false positive rate (FPR) as the abscissa and the true rate (TPR) as the ordinate. Among them:

\[
FPR = \frac{FP}{FP + TN}
\]

\[
TPR = \frac{TN}{FP + TN}
\]

where AUC (area under the ROC curve) is used to determine the quality of the detection method. The NLOF algorithm proposed in this paper with the traditional LOF algorithm, COF algorithm, and RDOS algorithm for AUC is compared. The value of AUC is between 0 and 1. The comparison result is shown in Figure 8.

This paper runs the outlier detection algorithm on the AUC data set. After using these methods for outlier
detection, the comparison results of LOF, COF, RDOS, and kNN-LOF on AUC scores are shown in Figure 8. This paper sets up multiple possibilities of $k$ value to detect the change of AUC score under the diverse $k$ value. It can be drawn that when $k = 45$, the highest AUC score obtained by the kNN-LOF algorithm is 0.965. Although the COF algorithm has the highest AUC score of 0.935 when $k = 20$, the kNN-LOF algorithm has the best hit rate for outliers in the case of a relatively large data set.

**Algorithm complexity analysis**

The complexity of density-based outlier detection algorithms mainly depends on the determination of neighborhoods. In the traditional method, the time complexity of the LOF algorithm mainly depends on the neighborhood query operation. The LOF calculation of each data object involves the neighborhood query so that the time complexity of the LOF algorithm is $O(n^2)$. The outlier detection algorithm based on kNN-LOF proposed in this paper improves the traditional algorithm in time complexity.

The algorithm proposed in this paper initially finds the $k$-nearest neighborhood range of the data object. Using kNN to divide the effective range of the data set is accurate to a certain extent the neighborhood query range. Through the hierarchical adjacency order, the neighborhood range is hierarchized under different link distances. The average sequence distance is calculated based on the data objects in the hierarchy. Different weights are assigned to neighbors in different $k$ ranges to distinguish the influence of different neighboring ranges. Finally, the calculation of the outlier coefficient is accurate and the accuracy is improved.

The reduction of the neighborhood range for the $k$th distance neighborhood of the subsequent calculation object also greatly reduces the amount of calculation. The range is sequenced according to the different impact factors, and the concept of average sequence distance is introduced to calculate the new local outlier coefficient, which not only improves the accuracy to a certain extent, but also improves the time efficiency of the algorithm execution by times.

**Conclusions**

In this paper, we propose an average sequence distance, which assigns different weights to the data of different neighborhood ranges. By considering the proximity degree of the points in different neighborhoods to the data objects, the influence of different adjacent ranges can be distinguished to further accurately calculate the outlier coefficient. The partition data set is divided into different regions by introducing the kNN algorithm. The scope of the neighborhood query is more accurate. Finally, the effectiveness of the proposed algorithm is verified by experiments.

To sum up, this paper effectively combines the two detection methods based on distance and density, and proposes an outlier detection algorithm based on kNN-LOF, so as to improve the accuracy of outlier detection. In the future, how to reduce the time complexity and improve the efficiency of the algorithm are the main research directions.

**Declaration of conflicting interests**

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the National Key R&D Program of China under Grant 2019YFB2103003, in part by the Scientific and Technological Support Project of Jiangsu Province (nos. BE2019740, BK20200753, and 20KJB520001), and in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province (nos. SJKY19_0761, SJKY19_0759, KYCX20_0759, SJCX21_0283).

**ORCID iDs**

[He Xu](https://orcid.org/0000-0003-2809-2237)
[Lin Zhang](https://orcid.org/0000-0003-4873-3353)
[Peng Li](https://orcid.org/0000-0001-5026-5347)
[Feng Zhu](https://orcid.org/0000-0002-7129-1082)

**References**

1. Catani S, Colla V and Vannucci M. Outlier detection methods for industrial applications. Italy: I-Tech Education and Publishing KG, 2008.
2. Jiawei H. and Micheline K.: Data mining: concepts and techniques. 3rd ed. pp. 1–5. Morgan Kaufmann Publishers, Waltham, USA (2011).
3. Barnett V and Lewis T. Outliers in statistical data. Wiley series in probability and mathematical statistics. Wiley, USA (1984).
4. Johnson T., Kwok I. and Ng R.: Fast computation of 2-dimensional depth contours. In: 4th International Conference of Knowledge Discovery and Data Mining. pp. 224–228. AAAI Press, USA (1998).
5. Chintalapudi K.K. and Kam M.: The credibilistic fuzzy c means clustering algorithm. In: 1998 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2034–2039. IEEE, San Diego, USA (1998).
6. Zhu L, Qiu YY, Yu S, et al. Outlier detection method based on fast k-nearest neighbors for minimum spanning tree. Chin J Comput 2017; 40: 2856–2870.
7. Knorr E. and Ng R.: Algorithms for mining distance-based outliers in large datasets. In: 24rd International Conference on Very Large Data Bases, pp. 392–403. Morgan Kaufmann Publishers Inc, New York (1998).
8. Ramaswamy S., Rastogi R. and Shim K.: Efficient algorithms for mining outliers from large data sets. In: 2000 ACM SIGMOD Int Conf on Management of Data, pp. 427–438. Association for Computing Machinery, New York (2000).
9. Angiulli F and Pizzuti C. Outlier mining in large high-dimensional data sets. IEEE Trans Know Data Eng 2005; 17: 203–215.
10. Jiang F, Du JW, Sui Y, et al. Outlier detection based on boundary and distance. *J Electron* 2010; 38: 700–705.
11. Breunig M.: Lof: identifying density-based local outliers. In: ACM Sigmod International Conference on Management of Data, pp. 93–104. Association for Computing Machinery, New York (2000).
12. Tang J., Chen Z.X., Fu A., et al. Enhancing effectiveness of outlier detections for low-density patterns. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 535–548. Springer, Berlin (2002).
13. Ni WW, Chen G, Lu JP, et al. Detection algorithm of weighted subspace outliers based on local information entropy. *Comp Res Develop* 2008; 45: 1189–1194.
14. Tang B and He H. A local density-based approach for outlier detection. *Neurocomputing* 2017; 241: 171–180.
15. Hu CP and Qin XL. Density-based local outlier detection algorithm DLOF. *J Comput Res Dev* 2010; 47: 2110–2116.
16. Liu HJ. Research and application of outlier detection method based on density and distance. pp. 17–30. Xi’an University of Technology, Xi’an, China 2019.
17. Wang K. and Zhou Z.: Distance ratio-based weighted rank outlier detection on wearable health data. In: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp. 583–588. IEEE, Chengdu, China (2019).