Research on data outlier detection method based on sample parameter selection LOF

Yanwei Huo[1], *Lei* Yin[1,2,*], and *Ran* Wang[2]

[1]School of Computer Science, Xi'an University of Posts and Telecommunications, Xi'an 710121, China
[2]School of Electro-Mechanical Engineering, Xidian University, Xi'an 710126, China

Abstract. The LOF data anomaly detection method has some defects, such as the value of *k* has great influence on the accuracy of detection results, and the selection of *k* value usually adopts trial method, which consumes a lot of calculation time. Therefore, this paper proposes an anomaly detection method for LOF data based on sample parameter selection, Tagged according to the sample data set point of normal and abnormal point, the adaptive selection of *k* value and outlier detection, so as to improve the accuracy of data outlier detection and calculation speed, and through the example of meteorological data outliers detection showed that LOF abnormal data points based on sample parameter selection method in the detection accuracy and reliability are improved significantly.

1 Introduction

In general prediction problems, the model is usually a form of expression of the data structure of the whole sample, which usually captures the general properties of the whole sample, and the points that are completely inconsistent with the whole sample in these properties are called outliers [1]. These anomalies often lead to inaccurate prediction results and serious adverse consequences. At present, the detection methods of outliers mainly include distance-based, clustering-based, distribution statistics-based, support vector machine-based and density-based methods [2-4]. Density-based and distance-based detection methods are widely used in various scenes. The set of outliers generated by these anomaly detection methods and their scores may be highly dependent on the number of clusters used and the existence of the total outliers in the data, as well as sensitive to the choice of parameters. The LOF outlier detection algorithm [5-7] combines the data point *q* with the surrounding *k* points for analysis, which makes the final outlier factor value more reasonable, reduces the impact of density maximum and density minimum on the whole data, and uses the numerical form to represent the outlier degree of data points, which is easier to understand. Only one parameter *k* needs to be set, which is easy to operate and implement [8]. However, in the LOF outlier detection method, there is a problem of selecting the parameter *k*. whether the value of *k* is properly selected will directly affect the effect of the outlier detection algorithm.

* Corresponding author: Yinlei_w@163.com

2 Research on LOF anomaly detection method based on sample parameter selection

Considering the advantages and disadvantages of various methods and the characteristics of meteorological data, among the conventional density-based outlier detection algorithms, LOF outlier factor (local outlier factor) detection algorithm is relatively mature and defined based on distance. The method defines the *k* neighborhood distance of the object and calculates the local anomaly factor LOF, which reflects the degree of anomaly, according to the local density of the object. All objects whose LOF value is greater than a specified threshold value will be judged as outliers [8]. However, in the LOF outlier detection method, there is a problem of selecting the parameter *k*. whether the value of *k* is properly selected will directly affect the effect of the outlier detection algorithm. In order to eliminate the problem caused by *k* selection, this paper adopts the selection method of sample parameters to adaptively select the appropriate *k*.

2.1 LOF anomaly detection method

LOF algorithm is a density-based unsupervised anomaly detection algorithm. This algorithm determines whether the point *p* is an isolated point by comparing the density of points in each point *p* and its adjacent area [9]. Suppose D is the noise data set, and each point in the region is the object in the data set, as shown in Figure 1. The calculation steps of LOF are as follows:

(1) For any natural number *k*, calculate the distance of the point *k* far away from point *p*, which is recorded as $k-dist$, that is $o\{s,t,v,q,h\}$, there are at least *k* point O objects in data set D, which satisfies $d(p,o) \le k-dist$; There are at most *k*-1 point O objects, which are included in the D dataset except points *p*.

(2) Calculate the k-th distance neighborhood of *p*, expressed as $N_k(p)$, That is, the set of all points within the k-th distance of the *p* object. Namely

$$N_k(p) = \{o \in D \setminus p \mid d(p,o) \le k-dist\} \tag{1}$$

(3) Calculate the reachable distance of *p*, expressed as $reach-dist(p,o)$, if the reachable distance is less than or equal to k - $dist$, It indicates that the O object is in the array of the k-th neighborhood, and the reachable distance at this time is $k-dist$; If point O is not in the k-th range, the reachable distance is the real distance between the two objects. It can be expressed as the following formula:

$$reach\_dist(p,o)_k = \max\{k-dist(o), d(p,o)\} \tag{2}$$

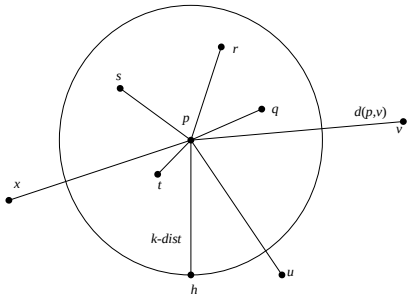

Fig. 1. k-nearest neighbor distance of point *p*.

(4) And the set of mutual distances of objects in a data set D can be expressed as $dist$, then the sum of k nearest neighbor distances of p with respect to D can be expressed $d(p,D)_k$, and its calculation formula is:

$$d(p,D)_k = \sum_{o \in N_k(p)} reach\_dist_k(p,o) \tag{3}$$

(5) Calculate the local reachable density of p and record it as $lrd_k(p)$, represents the reciprocal of the near sum of the data points adjacent to point p and k. Expressed as:

$$lrd_k(p) = \frac{|N_k(p)|}{d(p,D)_k} \tag{4}$$

The local reachable density represents the degree of dispersion of objects around the point. Objects with small local reachable density are more likely to be judged as isolated points, while objects with large local density are more likely to belong to normal points [10].

(6) Calculate the local isolation coefficient of p, the value of k is given arbitrarily, $lof(p)_k$. The local isolation coefficient depends on the value of k. for different values of k, the same data point may have different local isolation coefficients. LOF algorithm measures the isolation of a data point by focusing on its relative density with adjacent data points, rather than its absolute local density. Therefore, it is expressed by the average of the ratio of the local reachable density of the neighborhood of point p, $N_k(p)$ to the local reachable density of point p. Namely

$$lof(p)_k = \frac{1}{|N_k(p)|} \frac{\sum_{o \in N_k(p)} lrd_k(o)}{lrd_k(p)} \tag{5}$$

where: $lrd_k(o)$ Represents the neighborhood point $N_k(p)$ of point p local reachable density.

If the local isolation coefficient of object p is large, it indicates that there are few objects in the local range of the object, that is, the local density is small, which indicates that the object has a high probability of being an isolated point, and vice versa.

2.2 A LOF method based on sample parameter selection

When detecting outliers based on LOF anomaly detection algorithm, the performance of the algorithm is directly related to the selection of parameter k, and the value of k directly affects the detection results of the algorithm. If k value is too small, outliers will be lost and recall rate will decrease. If k value is too large, the number of abnormal points will be increased to a certain extent, which will increase the misjudgment rate. Therefore, the reasonable selection of parameter k is very important for LOF algorithm anomaly detection. In the current research, the selection of k usually adopts the heuristic method, which starts from 1 and proceeds to the most effective k. Although this method is simple, it is too blind. It will calculate more useless k values and consume more time. In order to eliminate the problems caused by k selection, this paper adopts the method based on sample parameter selection, calculates and analyzes the marked normal points and abnormal points in the sample data set, and adaptively selects the appropriate k.

In order to prevent misjudgment of outliers due to a small value of k or omission of outliers due to a large value of k, this paper selects k reasonably by calculating the dispersion to maximize the difference between outliers and normal points.

According to Formula (3), the smaller the $d(p,D)_k$ value of an object *p*, it means that the neighborhood objects of *p* are dense; On the contrary, the larger $d(p,D)_k$, the more sparse the neighborhood range of *p*, and the more likely the *p* object is an outlier. That is $d(p,D)_k$, outliers and outliers can be distinguished. Since $d(p,D)_k$ is greatly affected by *k*, the ratio of k-nearest neighbor distance and the number of edges $k(k-1)/2$ between k-nearest neighbors is recorded as dispersion $D_k(p)$ to maximize the difference between normal points and abnormal points. The calculation formula is:

$$D_k(p) = \frac{2d(p,D)_k}{k(k-1)} \tag{6}$$

Therefore, use $D_k(p)-k$ to observe that $D_k(p)$ is affected by the value of *k*. Fig. 2 shows the k-dispersion diagram of an abnormal point and a normal point.

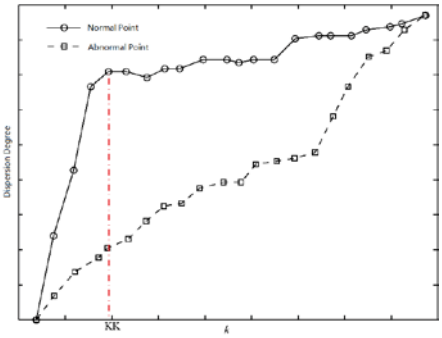

Fig. 2. Schematic diagram of k-dispersion of an abnormal point and a normal point.

It can be seen from Figure 2 that there are significant difference between abnormal points and normal points. When the value of *k* is small, $D_k(p)$ rises rapidly because *k* nearest neighbors are far away from each other. When *k* rises to $D_k(p)$ certain value, a tends to slow down under the influence of normal data, while the normal point is relatively flat with the rise of *k*. In order to ensure the accuracy of subsequent detection of outliers, the difference between outliers and normal points should be maximized, so *k* value corresponding to the point with the largest longitudinal interval between the two lines should be selected. FIG. 2 Take *kk* from *k*.

To sum up, the LOF anomaly detection algorithm flow based on sample parameter selection is presented, as shown in Figure 3.

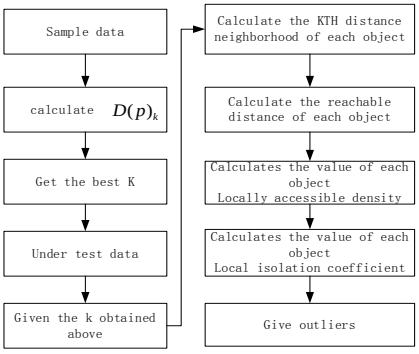

Fig. 3. LOF anomaly detection algorithm flow based on sample parameter selection.

Step 1: given the sample data of marked normal points and isolated points, respectively calculate the change of normal points and isolated points with *k*, the ratio of k-nearest neighbor distance and the edge number $k(k-1)/2$ between k-nearest neighbors, and obtain the *k* value corresponding to the maximum difference of calculation results, that is, the *k* value selected by the sample method.

Step 2: take the value of *k* as the parameter *k* of anomaly detection based on LOF algorithm.

Step 3: calculate the k-th distance neighborhood of each point according to Formula (1), and calculate the reachable distance of each point according to Formula (2).

Step 4: The local accessible density is calculated according to Formula (4), and the local isolation coefficient is calculated according to Formula (2-5). According to the local isolation coefficient, outliers are outputted.

2.3 experimental verification

The accuracy of wind speed measurement has an important influence on the numerical prediction and calculation results in wind sand environment. Therefore, this paper takes the wind speed data measured by a meteorological station in Northwest China as the verification object, and takes the wind speed as the data object for outlier detection. The test set of this experiment comes from a meteorological station in the northwest. Its value range is limited to a certain extent, and some data points not within this range are generated according to a certain probability, which are regarded as isolated points. When the LOF anomaly detection algorithm is running, the sample test set composed of normal data points and outliers is extracted from the experimental test set in a certain proportion. Table 1 lists some experimental test sets.

Table 1. Experimental data of anomaly detection.

attribute / amount	temperature（℃）	wind speed（m/s）	Relative humidity（%）
1	8.6	5.9	0.41
2	10.9	5.5	0.42
3	17.1	5.2	0.39
4	30.7	5.0	0.35
……	……	……	……
20	31.5	7.8	0.32
21	30.8	7.7	0.34
22	29.8	7.5	0.35
23	28.5	7.3	0.36
24	27.6	7.1	0.33
……	……	……	……

Based on LOF anomaly detection algorithm, the isolation degree can be quantified by calculating the local reachable density and isolation coefficient. The isolation degree of each data object can be seen very intuitively. The tester can determine the real anomaly point according to the actual situation, which is more practical than directly determining whether it is an anomaly point.

Generally, the performance of outlier detection algorithm is evaluated from recall rate and misjudgment rate [17]:

Recall is defined as follows:

$$\eta=\frac{m_d}{m_t}\times 100\% \tag{7}$$

where: η represents the recall rate, m_f and m_n represent the number of outliers actually contained and detected respectively.

Recall ratio indicates the detection capability of the algorithm. The higher the recall ratio is, the more outliers the algorithm detects, that is, the stronger the ability of detecting outliers.

The false positive rate is defined as follows:

$$\xi=\frac{m_f}{m_n}\times 100\% \tag{8}$$

where, ξ is the false positive rate, m_f and m_n are the number of misjudged normal points and the number of actual normal points respectively.

The false positive rate indicates the detection accuracy of the method. The smaller the false positive rate, the fewer the number of normal points misjudged, that is, the higher the detection accuracy of the algorithm.

Figure 4 shows an isolated point and a normal point $k-D_k(p)$.

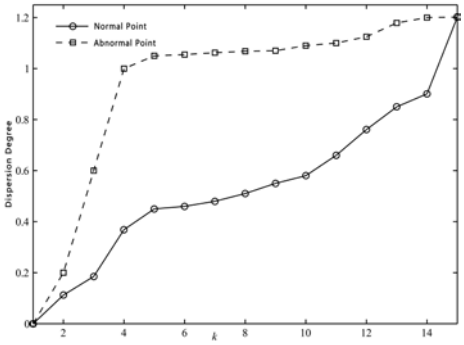

Fig. 4. Example diagram of k-dispersion of an abnormal point and a normal point.

As shown in Figure 4, when k = 4, the difference between isolated points and normal points is the largest. For the normal point, the LOF value changes steadily and is small. For outliers, the LOF value is much larger than that of normal points. The accuracy of selection based on sample data k is proved. Therefore, in this paper, the parameter k =4 is taken for the detection of outlier points based on LOF in test data.

Figure 5, Figure 6 and Figure 7 show the curves of wind speed original data and LOF value when k =4, 6 and 8 respectively.

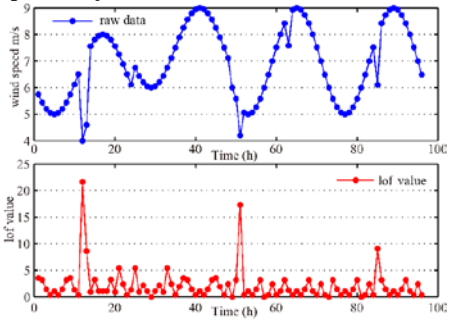

Fig. 5. Original data of wind speed and LOF value when k = 4.

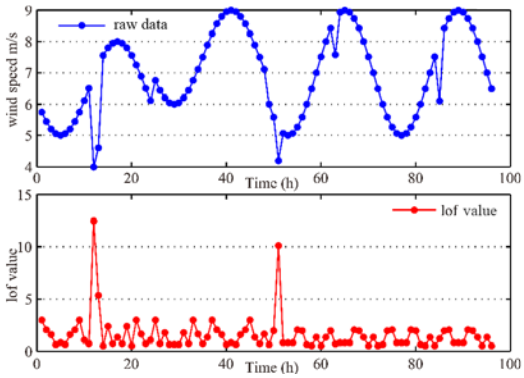

Fig. 6. Original data of wind speed and LOF value when k = 6.

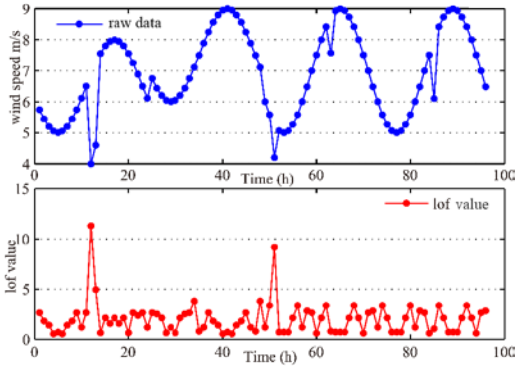

Fig. 7. Original data of wind speed and LOF value when k =8.

According to Figure 5, Figure 6 and Figure 7, certain abnormal points can be detected no matter when *k* is 4, 6 and 8. From the original wind speed data, t there are 5 outliers, and the LOF calculated by LOF algorithm is significantly higher than other points in 4. According to formula (7) and (8), when *k* is 4, the recall rate is 100%, when *k* is 6 and 8, the recall rate is 80%, and the misjudgment rate is 0. It can be seen that LOF anomaly detection algorithm has high accuracy and stability in detecting outliers.

Through the above experiments and result analysis, the accuracy and reliability of sample data selection *k* and LOF anomaly detection algorithm are verified. Thus, it can effectively solve the problem of abnormal value detection of climate data and improve the quality of data used in environmental prediction calculation.

3 Conclusion

Based on the analysis of LOF outlier detection algorithm, a LOF data outlier detection method based on sample parameter selection is proposed in this paper. This method marks normal points and abnormal points in the sample data set, calculates the reachable distance and local reachable density of each data point, realizes the adaptive selection of *k* value and data abnormal point detection, and greatly improves the accuracy and reliability of LOF data abnormal point detection.

This work was supported by the National Key R&D Program of China [grant number 2019YFB1705404].

References

1. Huang Jingtao, Ren Zhiwei, Luo Wei. Research on Outlier Detection Algorithm of Power Station Boiler Monitoring Data [J]. Computers and Applied Chemistry **30(10)**, 1153-1156(2013)
2. MIAO Runhua. Research on data preprocessing method based on clustering and outlier detection [D]. Beijing: Beijing Jiaotong University,(2012)
3. BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: identifying density-based local outliers[C]// ACM SIGMOD International Conference on Management of Data, Dallas, Texas, USA May 15-18, 2000, Oettingenstr, Munich, Germany: MDDT, **29(2)**, 93-104(2000)
4. Tian Jiang. Research on Outlier Detection Method Based on Support Vector Machine [D]. Liaoning: Dalian University of Technology, (2009)
5. MA Y, SHI H, MA H, et al. Dynamic process monitoring using adaptive local outlier factor[J]. Chemometrics & Intelligent Laboratory Systems, **127(18)**, 89-101(2013)
6. GAO Z. Application of Cluster-Based Local Outlier Factor Algorithm in Anti-Money Laundering[C] International Conference on Management and Service Science, 1-4 (September 20-22, 2009, Wuhan, China: MSS, 2009)
7. VINTROVA V, VINTR T, REZANKOVA H. Comparison of Different Calculations of the Density-Based Local Outlier Factor[J], 60-67(2012)
8. Wang Fei. ILOF *: An Improved Local Anomaly Detection Algorithm [J]. Applications of Computer Systems, **24(12)**,233-238(2015)
9. Xue Anrong, Yao Lin, Ju Nimui, et al. A Review of Outlier Mining Methods [J]. Computer Science, **(11),** 13-18, 27(2008)
10. Wang Qian, LIU Shuzhi. Improvement of Local outlier data Mining Method based on Density [J]. Application Research of Computers, **31(6),** 1693-1696, 1701(2014)