Meteorological outliers detection based on artificial intelligence

F Xue¹, W Zheng¹, M Zhang¹*, Q Wu¹, Z Yang² and X Ai¹

¹International Joint Research Laboratory of Magnetic Confinement Fusion and Plasma Physics, State Key Laboratory of Advanced Electromagnetic Engineering and Technology, School of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China

²China-EU Institute for Clean and Renewable Energy, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China

*Corresponding author’s e-mail: zhangming@hust.edu.cn

Abstract. To precisely evaluate the effect of artificial precipitation of Project Tianshui, anomaly data within the large dataset collected is supposed to be detected and dealt with reasonably, to enhance the analysis and prediction of the data. Using the data accumulated from 2008 to 2017 from Wushaoling Meteorological Station, and the data from May 2019 to October 2019 from the dedicated supervising network built for the project, taking temperature(0.1℃) and other meteorological data as example, anomaly detection is conducted by the arithmetics Local Outlier Factor and Isolation Forest respectively. Consistency check is also conducted via the two arithmetics. The main outliers of the accumulated data for ten years are isolated outliers. LOF works pretty well, while Isolation Forest performs unsatisfying. The main outliers of the dedicated supervising network data are of a continuous outlier sequence. LOF can hardly detect the anomaly, while Isolation Forest precisely find out the outliers. Both arithmetics perform well on consistency check in space. The differences of LOF and Isolation Forest are compared. New ideas of using artificial intelligence arithmetics to detect meteorological outliers and pre-process the data are provided.

1. Introduction
With the rapid development of China's meteorological industry and the continuous improvement of meteorological information level, meteorological data collection has become more and more information-based and intelligent. From the former manual data collection to the automatic meteorological station data collection, the collection frequency and accuracy of meteorological data are constantly improved, marking the development of meteorological observation in China to a new level [1]. The meteorological data collected by the automatic meteorological station has the characteristics of high frequency, high precision, large amount of data, obvious time sequence and spatial characteristics. If we can make good use of these meteorological data, we can make more accurate meteorological forecast and analysis of abnormal meteorological conditions, so as to bring convenience to the national production and life.

However, it is bringing a great convenience while also resulting in an urgent problem of how to deal with such a huge amount of data. Due to many uncertain factors such as local atmospheric disturbance, stability difference of automatic observation station, disturbance of observation surrounding environment and data communication, errors and anomalies often exist in the accumulated meteorological data. For example: data missing caused by abnormal collection of meteorological
equipment in a certain period of time, data error caused by sensor sensitivity reduction or failure caused by disturbance and poor working environment, etc., will affect the effect of data utilization later.

The traditional outlier’s anomaly detection of meteorological data mainly depends on the climate limit value check [2], internal consistency check, time consistency check [3] and space consistency check [4]. This method can screen out some data outliers, but it basically depends on the meteorological data boundary value in the historical data, and because of the huge amount of data, it needs a lot of manpower and material resources to clean the data, so it lacks sensitivity and efficiency to the abnormal changes.

With the rapid development of computer technology and the explosive growth of data, artificial intelligence technology has been widely used in various big data processing occasions [5]. For example, artificial intelligence is applied to aerospace anomaly detection events [6], power system planning [7], network security management [8], air quality analysis [9], etc.

This paper introduces two filtering algorithms of meteorological data outliers based on artificial intelligence, and makes some exploratory analysis on the application methods of artificial intelligence in meteorological scientific research, hoping to provide new ideas for filtering and eliminating meteorological data outliers.

2. Local Outlier Factor Detection Algorithm

LOF (Local Outlier Factor) algorithm gives each object in the data set a factor LOF to represent its degree of outlier, and identifies whether the object is an outlier by judging the size of LOF.

The density based local outlier factor LOF is based on the following definitions [10-15]:

Definition 1.1: k-distance of object p

p is a point in data set D. For any positive integer k, k-distance of object p is recorded as $k\text{-distance}(p)$. It is defined as the distance $d(p,o)$ between p and object o $\in D$, and meets the following requirements:

- There are at least k objects $o \in D(p)$, with $d(p,o) \leq d(p,o)$
- There are at most $k-1$ objects $o \in D(p)$, with $d(p,o) > d(p,o)$

Definition 1.2: k-distance neighborhood of object p

The k-distance neighborhood of object p includes all data objects whose distance from p is not greater than k-distance. It is defined as:

$$N_{k\text{-distance}}(p) = \{q \in D(p)|d(p,q) \leq k\text{-distance}(p)\}$$

Definition 1.3: Reachable distance between object p and object o

The reachable distance between object p and object o is defined as:

$$\text{reach-dist}_k(p,o) = \max \{k\text{-distance}(o), d(p,o)\}$$

If the object p is in the k-distance neighborhood of the object o (as $p_1$ in Figure 1), the reachable distance between the two can be replaced by the k-distance of o; if the object p is not in the k-distance neighborhood of the object o (as $p_2$ in Figure 1), the reachable distance between the two is the actual distance between p and o.

![Figure 1. The reachable distances of objects o and p1 and o and p2 when k = 4.](image)

Definition 1.4: Local reachable density of object p

The local reachable density of object p can be defined as:
In essence, the local reachable density of an object is an estimate of the density of the data object within its distance neighborhood, which is based on the reciprocal of the average reachable density of the nearest neighbor. If the sum of all reachable distances is 0, the local density is $\infty$.

Definition 1.5: Local outlier factor $LOF$ of object $p$

The local outlier factor $LOF$ of object $p$ can be defined as:

$$LOF_{\text{MinPts}}(p) = \frac{\sum_{o \in \text{MinPts}(p)} \frac{lr\text{d}_{\text{MinPts}}(o)}{\sum_{o \in \text{MinPts}(p)} \text{dist}_{\text{MinPts}}(p,o)}}{|\text{MinPts}(p)|}$$

The local outlier factor $LOF$ of object $p$ is the average of the local reachable density ratio of its nearest neighbor. The local outlier factor represents the outlier degree of an object. The smaller the local reachable density of $p$ is, the greater the local reachable density of the neighborhood object is, and the greater the $LOF$ value is, the greater the outlier degree is.

In general, the abnormal data only accounts for a small part of the total data, and the abnormal data is isolated from other data. Therefore, by calculating the local density of each group of data, and then calculating the $LOF$ value of each group of data, we can separate the abnormal data from the normal data.

The main steps of $LOF$ algorithm anomaly detection in this paper are as follows:

Inputs:
- n-dimensional data set $D$;
- Parameter MinPts (indicates the minimum number of objects in the neighborhood of $p$);
- Parameter Threshold (indicates the threshold value of $LOF$ used to filter abnormal data)

Output: Outliers in the dataset
- Determine the $k$-distance neighborhood of each object according to its attributes and spatial relations;
- Calculate the Euclidean distance between the object and all objects in its neighborhood, so as to calculate the local reachable density of each object;
- Calculate the local outlier factor $LOF$ of each object;
- Compare the local outlier factor $LOF$ and Threshold of each object;
- Output all objects with $LOF(p) > \text{Threshold}$, that is, abnormal data.

The flow of $LOF$ algorithm implementation is shown in Figure 2.

3. Isolation Forest Algorithm

Isolation Forest (iForest) algorithm is an anomaly detection method based on Ensemble, which has linear time complexity, high accuracy and high speed in processing large data. It is widely used in various occasions at present. Common scenarios include insider threat detection [16], covert communication detection [17], noise data filtering (data cleaning) [18], etc.

The main idea of the Isolation Forest algorithm is to segment the data space into two subspaces by randomly selecting random hyperplanes. Then, a hyperplane is randomly selected to continue to segment the data space until each subspace contains only one data point. The more times each data point is segmented, the higher the density of its cluster, the lower the probability that it is abnormal data [19].

iForest is composed of several isolation trees (iTree). The main steps to create iTree are as follows:

Input:
- n-dimensional data set $X$;
- Number of iTree $t$;
- Number of subsamples $\psi$

Output: Sets of $t$ iTrees
- Select $\psi$ points randomly from the training data set as sub samples and put them into the root node of an iTree;
Randomly specify a dimension. Within the data range of the current node, a segment point \( p \) is randomly generated, and \( p \) is between the maximum value and the minimum value of the currently specified dimension;

- The hyperplane generated by the segment points divides the data space into two subspaces, and classifies the points smaller than \( p \) in the current dimension into the left branch of the node, and the points larger than \( p \) into the right branch of the node;
- Recurse the second and third steps on the left and right branches of the node, and continuously generate new sub nodes until each sub node contains only one data, or the tree has grown to a limited height.
- Since the segment process is completely random, the result is converged by the method of ensemble, and the segment is repeated from the beginning, then the average value of each segment result is calculated. Repeat the above steps until \( t \) iTrees are obtained.

The reason why limiting the height of the tree is that we are more concerned about the branches with short paths, and the probability that the data contained in the end points of the branches with short paths are outliers is higher, while the branches with long paths are often normal data, which we do not care about.
After obtaining iTrees, we can use iTree to test and evaluate the data, that is, to calculate the abnormal score $s$. The calculation formula of abnormal score $s$ is as follows:

$$s(x, \psi) = 2^{\frac{E(h(x))}{c(\psi)}}$$

In the formula, $x$ is the sample, $h(x)$ is the height of $x$ in each tree, $c(\psi)$ is the average value of path length when the number of samples is given, which is used to standardize the path length $h(x)$ of sample $x$. According to this formula, the closer the abnormal score is to 1, the more likely it is to be an abnormal point; the closer it is to 0.5, the less likely it is to be an abnormal point.

4. Experimental Results and Analysis

The data used in this paper is the accumulated meteorological data of Wushaoling meteorological station from 2008 to 2017, in hours (hereinafter referred to as the ten-year cumulative data), and the accumulated meteorological data of dedicated supervising network from May to October 2019, in minutes (hereinafter referred to as the dedicated supervising network data). Take the ten-year cumulative data, the timed temperature (0.1°C) data of dedicated supervising network data and the precipitation data of dedicated supervising network data as examples.

4.1. Anomaly detection for ten-year cumulative data and dedicated supervising network data using LOF algorithm

Firstly, LOF algorithm is used to detect the anomaly of ten-year cumulative data. There are 87672 pieces of ten-year cumulative data. There are two adjustable parameters in the algorithm, which are the number $k$ of neighborhood points, that is, MinPts (the same below) in the algorithm, and the LOF threshold to determine whether the data is abnormal. Adjusting these two parameters will affect the detected outliers. In general, the threshold is constant, and the effect of anomaly detection can be changed by adjusting the number $k$ of neighborhood points. After many experiments, when the fixed threshold is 1.75, the detection effect of $k=50$ is the best. In order to briefly explain the effect of taking different $k$ values, take $k$ as 10, 50 and 100 respectively for anomaly detection of ten-year cumulative data, as shown in Figure 3.
In this case, if the $k$ value is relatively small, the neighborhood detected by each data is small, the local density calculation value is scattered, and the LOF value is too large, which will cause more normal points to be detected as abnormal points by mistake; if the $k$ value is relatively large, the neighborhood detected by each data is large, the local density calculation value is relatively centralized, and the LOF value is too small, which will cause the abnormal points not to be detected. In application, the specific value of $k$ should be judged according to the actual situation and the size of data set. It can also be seen from the comparison of Figure 3 (a) (b) (c), when $k$ is taken as 10, there are more abnormal data detected, but a small part of normal data is mistakenly detected as abnormal points; when $k$ is taken as 50, it can basically ensure that the actual abnormal points are detected, and at the same time, a small part of normal data is mistakenly considered as abnormal points; when $k$ is taken as 100, there is basically no normal data mistakenly considered as abnormal data, but some of the actual abnormal points have not been detected.

It can be seen from Figure 3(b) that when $k = 50$, most abnormal data can be detected by LOF algorithm. Compared with the calibrated data, the detection rate is over 97%. However, some normal points are mistaken as abnormal points, as shown in Figure 4. It can be seen in the figure that although the points shown in the figure are continuous in space, the local density of these two points in the data...
space is smaller than that of other points around them, so their LOF value is larger, and they are detected as abnormal points by the algorithm.

![Figure 4. Partial map of temperature data (normal points) on March 3, 2011.](image)

After that, LOF algorithm is used to detect the anomaly of the dedicated supervising network data, as shown in Figure 5 and Figure 6. There are 70237 pieces of dedicated supervising network data. It can be seen from Figure 5 that the obvious anomaly (-40.0°C) from August 24, 2019 to August 25, 2019 cannot be detected by LOF algorithm, and only the isolated outliers such as July 30, 2019 can be detected as shown in Figure 6. This is because LOF is an algorithm based on local density. As long as the local density reaches the threshold value, it will not be detected as an abnormal value. The anomaly of dedicated supervising network data is sequence anomaly. Because every point in the sequence has similar distribution points, the local density of the abnormal part is very large, and the LOF value is very small, so it can not be detected.

![Figure 5. Global map of anomaly detection on dedicated supervising network data using LOF.](image)
In addition, by comparing the overall LOF value of the ten-year cumulative data with the overall LOF value of the dedicated supervising network data, it is found that the overall LOF value of the dedicated supervising network data is far less than the overall LOF value of the ten-year cumulative data. This is because the dedicated supervising network data is counted in minutes, which is close to each other in time scale, with small deviation and large density between adjacent points, so the LOF value is small. However, the ten-year cumulative data is counted in hours, which is far away from each other in time scale, with large deviation and small density between adjacent points, so the LOF value is large.

4.2. Anomaly detection for ten-year cumulative data and dedicated supervising network data using Isolation Forest algorithm

Firstly, Isolation Forest is used to detect the anomaly of ten-year cumulative data, as shown in Figure 7. As can be seen from the Figure, the effect of Isolation Forest on anomaly detection of ten-year cumulative data is not good. On the one hand, the accuracy rate of abnormal data detection is low, only about 70%; on the other hand, a considerable part of normal data is wrongly identified as abnormal.
The reason for this phenomenon is that the Isolation Forest algorithm is not based on density or distance, but randomly segments the data space repeatedly, so as to use the number of times the subspace is segment or the length of the branch to determine whether it is abnormal. It can be seen from the figure that the abnormal data detected by the Isolation Forest is basically concentrated in the maximum and minimum values of the data, because their distribution in the data space is relatively sparse compared with the global. But for local and isolated abnormal data, it can not be well reflected.

Then, Isolation Forest is used to detect the anomaly of dedicated supervising network data, as shown in Figure 8. It can be seen from the Figure that after setting the appropriate threshold value, Isolation Forest can accurately detect the anomaly from August 24, 2019 to August 25, 2019, and there is no misidentification.

![Figure 8. Anomaly detection on dedicated supervising network data using Isolation Forest.](image)

4.3. Spatial consistency check for precipitation data under different stations using LOF and Isolation Forest algorithm

In the dedicated supervising network, there are 8 meteorological stations R0001 ~ R0008 specially used to detect precipitation data. The collection frequency of precipitation data is once per minute. Because most of the time there is no rain or snow, that is, the precipitation is zero, and in the case of precipitation, the precipitation per minute is also very small, so it is not significant to check the consistency of minute level precipitation. Therefore, the precipitation is summarized firstly, and then calculated and compared in the unit of daily precipitation.

However, due to the influence of equipment and experimental conditions, the precipitation data of meteorological stations are partially missing, and the missing dates of each meteorological station are different from each other. In order to eliminate the impact of missing measurement, the intersection of all meteorological stations is taken, so as to ensure that all meteorological stations monitor precipitation data within the intersection date. From May 1, 2019 to October 20, 2019, a total of 78 days of daily precipitation data are taken for consistency check.

The method of calculating standard deviation is used to check the consistency of precipitation data. However, since the standard deviation is one-dimensional data, and the inputs of the two algorithms used are required to be at least two-dimensional, so (standard deviation, standard deviation) is used as the input two-dimensional data for inspection, so that the one-dimensional data can be detected abnormally without adding other data features. First, two algorithms are used to screen out the date with poor consistency. Then (daily precipitation, daily precipitation) is used as the input of the algorithm to detect the abnormal stations in the eight stations of the day.
In the traditional method based on statistics, the measured value whose deviation from the average value is more than twice or three times of the standard deviation is removed as the abnormal value. However, in this case, because the spatial distance between the eight stations is very close, the actual situation is that the daily precipitation of most stations is exactly the same or very close, and the daily precipitation of one station has a relatively large deviation, but its deviation from the average value is not enough to reach two to three times the standard deviation. If the multiple of the standard deviation is reduced, some small errors with less daily precipitation will be detected. For example, if the daily precipitation of eight stations on September 19, 2019 is \([0.0, 0.1, 0.2, 0.0, 0.0, 0.0, 0.0, 0.0]\), the deviation between the measured value of R0002 and the average value is 2.15 times of the standard deviation, while the daily precipitation of each station on May 5, 2019 is \([30.0, 22.2, 10.5, 30.7, 30.7, 30.7, 30.7, 30.7]\), in which the deviation between the measured value and the average value of R0003 station with obvious abnormality is 1.97 times of the standard deviation. It can be seen that the effect of only using standard deviation for anomaly detection is not good. Considering that the measured values in this example are often close to each other and the standard deviation is often very small, the accurate results can be obtained by detecting the date firstly when the standard deviation is abnormal, and then detecting the site exception for each date.

The abnormal data detected by LOF algorithm and IsolationForest algorithm are shown in Table 1 and Table 2.

| Date          | Standard Deviation | LOF           | Problemed Sites |
|---------------|--------------------|---------------|-----------------|
| August 26, 2019 | 1.92280            | 3.90454       | R0003           |
| September 9, 2019 | 2.12901            | 4.23833       | R0002, R0003    |
| October 16, 2019 | 2.22450            | 4.40155       | R0002, R0003    |
| July 28, 2019  | 2.32652            | 4.57799       | R0002, R0003    |
| July 29, 2019  | 2.52756            | 4.92572       | R0002           |
| August 25, 2019 | 2.96009            | 5.69077       | R0003           |
| May 5, 2019    | 3.25925            | 6.22472       | R0003           |
| July 7, 2019   | 7.29711            | 13.95820      | R0003           |
| July 7, 2019   | 9.90700            | 19.19186      | R0003           |

| Date          | Standard Deviation | Decision Value | Problemed Sites |
|---------------|--------------------|----------------|-----------------|
| July 7, 2019  | 9.90700            | -0.26842       | R0003           |
| May 5, 2019   | 7.29711            | -0.23014       | R0003           |
| May 6, 2019   | 3.25925            | -0.12582       | R0003           |
| August 25, 2019 | 2.96009           | -0.08867       | R0003           |
| July 29, 2019 | 2.52756            | -0.05436       | R0002           |
| September 18, 2019 | 1.28397        | -0.04481       | R0002           |
| September 9, 2019 | 2.12901           | -0.03050       | R0002, R0003    |
| August 26, 2019 | 1.92280           | -0.03024       | R0003           |
| October 16, 2019 | 2.22450           | -0.02860       | R0002, R0003    |
| July 28, 2019  | 2.32652            | -0.02830       | R0002, R0003    |

In Table 2, Decision Value has the same effect as the abnormal score mentioned in the paper. The smaller Decision Value is, the greater the probability of data abnormality is. It can be seen from the results that normal and abnormal data can be distinguished by LOF and IsolationForest, and the results detected by the two algorithms are basically the same, with only one day difference. After further anomaly detection of the rainfall data of eight meteorological stations on the date of the problem standard deviation, it is found that the problems of R0002 and R0003 meteorological stations are particularly prominent. Among them, the anomaly of R0003 mainly shows that the precipitation on the abnormal day is far less than that measured at other stations. After the confirmation of the test site, it is
found that there is a long-term disconnection problem in R0003 meteorological station, which leads to the fact that the precipitation data measured in real time can not be transmitted to the server, resulting in a large number of missing data. This is also consistent with the results of data analysis. The anomaly of R0002 is that the precipitation is more than that of other stations. Since there is no problem with the collection equipment itself, it is inferred that the tipper rain sensor is not placed horizontally, and further confirmation is needed.

5. Conclusion

In this paper, LOF algorithm and Isolation Forest algorithm are used to detect the temperature data of Wushaoling meteorological station from 2008 to 2017 and the temperature data of dedicated supervising network from May 2019 to October 2019. Among them, the abnormal data in the temperature data of 2008 ~2017 is basically a single outlier, while the abnormal data in the temperature data of May 2019 ~2019 October is the whole series. Through the detection of two algorithms, we can see that LOF algorithm is more suitable for the detection of local single outlier, and the detection success rate is high. While the IsolationForest algorithm is more suitable for the detection of global outliers, and the detection success rate is high.

When checking the spatial consistency of the precipitation data of eight stations at the same time, two kinds of algorithms are adopted respectively, which makes reasonable use of the characteristics of the collected data, first calculate the standard deviation and determine the abnormal date, and then check the precipitation data of eight stations according to the specific date. By using these two algorithms, the abnormal date and abnormal meteorological station can be detected accurately, and the detection results are compared with the field equipment to confirm the accuracy of detection. At the same time, the disadvantages and shortcomings of anomaly detection based on traditional statistical methods are illustrated by examples.

As the two algorithms are suitable for local and global anomaly detection respectively, but can not reflect both local and global features, the next step can be considered to improve the LOF algorithm. From the initial detection of the neighborhood reachable density of each point, it is transformed into clustering data simply, and then detecting the neighborhood reachable density of each data, so as to weaken part of its function of detecting local exceptions and increase part of its function of detecting global exceptions. At the same time, due to the high time complexity of LOF algorithm, clustering the data in advance is also helpful to shorten the running time of the program.

Because the data used in this paper includes the data of dedicated supervising network from May to October 2019, there is not enough data for more in-depth analysis, and the conclusion also needs more data test. Wushaoling dedicated supervising network still needs to solve the faults detected in the paper and keep stable operation on the current basis to have more continuous and complete data, so as to better prepare for the effect evaluation of artificial rainfall and snow.

Acknowledgments

This work was supported in part by the National Key Research and Development Plan of China under Grant 2016YFC0401006 and in part by the National Natural Science Foundation of China (51821005).

References

[1] Hu, Y. (2004) Difference Between Data of Automatic and Manual Observation. Journal of Applied Meteorological Science, 06: 719-726.
[2] Yu, Y., Zhu, C., Zhao, Q. (2016) Integration and Preliminary Quality Control of Global Surface Hourly Temperature and Precipitation Data. Journal of Meteorology and Environment, 32: 94-101.
[3] Xu, Z., Chen, X., Wang, Y. (2013) Quality Control Scheme for New-built Automatic Surface Weather Observation Station’s Data. Journal of the Meteorological Sciences, 33: 26-36.
[4] Hua, Y., Lin, H. (2012) The Technology of Large Capacity Meteorological Data Visualization and Real-Time Spatial Analysis. Atmospheric Science Research and Application, 01: 76-84.
[5] Peng, Y., Pang, J., Liu, D. (2015) Big Data: Connotation, Technical Framework and Its Development. Journal of Electronic Measurement and Instrument, 29: 469-482.

[6] Zhang, K., Cai, Y., Ren, Y. (2019) Space Anomaly Events Detection Approach Based on Generative Adversarial Nets. Journal of Beijing University of Aeronautics and Astronautics, 45: 1329-1336.

[7] Yang, T., Zhao, L., Wang, C. (2019) Review on Application of Artificial Intelligence in Power System and Integrated Energy System. Automation of Electric Power Systems, 43: 2-14.

[8] Chu, M. (2012) Research on Key Technologies of Network Security Management Based on Artificial Intelligence Theory. Computer CD Software and Applications, 15: 95-108.

[9] Xiao, W. (2019) Anomaly Detection and Analysis of Air Quality Data Based on Isolation Forest Algorithm [J]. China Computer & Communication, 31: 38-40.

[10] Sander, J.(2000) LOF: Identifying Density-Based Local Outliers. Acm Sigmod Record, 29:93-104.

[11] Wang, F. (2015) iLOF*: An Optimized Local Outlier Detection Algorithm. Computer Systems & Applications, 24: 233-238.

[12] Liu, D., Li, X., Li, X. (2019) K-means Clustering Method Based on LOF and Its Application in Micro Seismic Monitoring. Journal of Safety Science and Technology, 15: 81-87.

[13] Chen, M. (2016) Research on Credit Card Fraud Detection Based on Fuzzy Local Outlier Factor (LOF) [J]. Financial Theory and Practice, 10: 54-57.

[14] Zhuang, C., Zhang, B., Hu, J. (2016) Anomaly Detection for Power Consumption Patterns Based on Unsupervised Learning. Proceedings of the CSEE, 36: 379-387.

[15] Zhou, P., Cheng, Y. (2017) An Improved LOF Outlier Detection Algorithm. Computer Technology and Development, 27: 115-118.

[16] Liu, Y., Luo, S., Qu, L., Pan, L., Zhang, J. (2019) Full-featured Information Equation Modeling for Insider Threat Detection. Journal of Zhejiang University (Engineering Science) 53: 777-784.

[17] Ni, Y., Yan, L., Cui, Y., Li, S. (2018) Convert Communication Detection Mechanism for Software Defined Work. Computer Systems & Applications 27: 143-150.

[18] Li, X., Shi, Z., Kang, Z. (2018) Data Cleaning Method for Power Operation and Maintenance Based on Isolated Forest Algorithm and BP Neural Network Algorithm. Electrotechnical Application, 37: 72-79.

[19] Liu, F., Ting, K., Zhou, Z. (2008) Isolation Forest. In: 2008 Eighth IEEE International Conference on Data Mining. Pisa. pp. 413-422.