IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON EMERGING TRENDS OF ENERGY AND SPECTRUM HARVESTING TECHNOLOGIES

Received February 24, 2021, accepted March 17, 2021, date of publication March 25, 2021, date of current version April 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068756

An Algorithm for Detection of Traffic Attribute Exceptions Based on Cluster Algorithm in Industrial Internet of Things

LIDONG FU[1], WENBO ZHANG[1], XIAOBO TAN[1], AND HONGBO ZHU[1,2], (Member, IEEE)
[1]School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China
[2]School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China

Corresponding author: Wenbo Zhang (zhangwenbo@yeah.net)

This work was supported in part by the Project of Security Monitoring Platform of Heterogeneous Terminal in Ubiquitous Power Internet of Things, in part by the China Academy of Military Sciences Fund, in 2019, in part by the Liaoning BaiQianWan Talents Program, in 2016, and in part by the Natural Science Foundation of Liaoning Province Project under Grant 20170540793.

ABSTRACT With the focus on network security and the goal of meeting the requirements of fast speed and high accuracy of abnormal traffic detection of industrial Internet of Things, this paper proposes a hierarchical abnormal traffic detection method for the industrial Internet of Things. This method includes two abnormal detection methods: the first (crude) one detects traffic frequency based on statistical analysis; the second (sophisticated) one detects traffic attributes based on a clustering algorithm. The hierarchical detection method first detects the abnormal frequency of the network traffic. Property exceptions are then detected for suspected traffic. The paper first calculates the difference in the value of traffic frequency and smoothes it. Then, the exponential weighted moving average model is used to make the data conform to the statistical law, and the deviation correction of the model is proposed to reduce the error caused by the initial value. Focusing on the fact that a fixed threshold is not suitable for a dynamic network environment, a two-layer threshold interval method is proposed to reduce the rate of false alarms. For traffic attribute detection, this paper designs a clustering optimization anomaly detection algorithm for complex attribute feature data. The algorithm classifies the weighted distance and safety coefficient of the data according to the priority of the traffic attribute features, selects the data with the high safety coefficient as the clustering center, clusters the multi-feature data around the center, and applies it to the attribute anomaly detection. The simulation results show that the traffic frequency detection algorithm based on statistical analysis proposed in this paper can quickly detect the traffic frequency anomalies in the network. Moreover, the clustering optimization anomaly detection algorithm based on complex attribute features proposed in this paper can effectively detect the malicious attributes contained in network traffic and achieve a high detection rate and a low false detection rate to ensure the safety and reliability of the industrial Internet of Things.

INDEX TERMS Clustering algorithms, industrial Internet of Things, traffic attribute detection, traffic frequency detection.

I. INTRODUCTION

With the advancement of the internet and artificial intelligence technology, traditional industrial production and management technologies have fallen behind in meeting the development needs of the time. Driven by new technologies, the traditional industrial manufacturing field is constantly transforming, ushering in new opportunities as well as challenges [1]. With the industrial reforms, diversified information technology and industrial production have become integrated, and new concepts, such as ''smart factory'' and ''intelligent manufacturing,'' have emerged [2]. In the industrial field, with ''intelligence'' being the core element, the concept of the Industrial Internet of Things (IoT), which includes production automation, intelligent logistics, advanced computing, and other features, has been put forward. This technology has been widely used in the field of industrial production, making the traditional manufacturing industry usher in a new look [3]. Through network interconnection, industrial IoT technology exchanges information between the device layers on the general control terminal and integrates management decisions with production operations.

The associate editor coordinating the review of this manuscript and approving it for publication was Deyu Zhang.

However, with the development of the internet, the scale of the network is growing, and the types of ports accessing the network are becoming more diversified. As a result, the general control system, data input, and networking equipment in the industrial IoT are vulnerable to attacks, which damages the normal operation of the network and even interrupts industrial production, causing economic losses [4]. Therefore, it is necessary to study detections of traffic security anomalies in the Industrial IoT.

In the Industrial IoT, abnormal network performance may be caused by network attacks, virus invasions, operation failures, unexpected access of illegal users, the sudden addition of new nodes, etc. These lead to network congestion and operation overload, causing a burst of traffic in the network, the router, and the communication links, and creating irreversible damage, which may eventually result in the network's non-responsiveness and collapse [5]. The abnormalities in the industrial IoT are mainly traffic anomalies that have two causes. First, network technologies always have defects, and vulnerabilities in protocols, management, and services are exploited by attackers [6]. Second, many new technologies that get integrated into the industrial IoT do not have reliable and verified security, bringing hidden security risks to the network [7]. The new technologies also introduce diverse data input interfaces, resulting in complex communication traffic, which makes the industrial IoT vulnerable to abnormal traffic [8]. In modern industries, the architecture of IoT depends on mature automation technologies, while also incorporating many new technologies. In recent years, abnormal network events have exposed the hidden security dangers that result from multi-technology integration and that hinder the development and popularization of the industrial IoT [9]. After a global announcement about the fourth Industrial Revolution, China put forward its 2025 Industrial Manufacturing Strategy. According to it, industrial IoT is expected to be an important technology in a variety of industries [10]. Therefore, studying the detection of the industrial IoT's abnormal traffic is critical to ensuring the safety of its assets.

II. RELATED WORKS

The main tasks of abnormality detection in machine learning are evaluating the network traffic components, identifying the traffic anomalies and their types, locating abnormal traffic, and analyzing the causes of the anomalies [11]. Many experts in China and abroad divide abnormal traffic detection based on machine learning into three categories: supervised learning, semi-supervised learning, and unsupervised learning [12]. Supervised learning requires an existing data set to establish the model with the features of known data. Unknown data is identified and classified based on the model [13], [14]. Semi-supervised learning also needs normal data to assist with modeling, and normal data is used as a standard for model diagnosis. Unsupervised learning, however, does not need prior data marking. Unknown data is analyzed and processed directly, and the discrete values between the unknown data and the normal data are calculated. If the deviation exceeds a certain degree, the data is diagnosed as abnormal [15].

The application of industrial IoT requires guaranteed network security because it has the features of an ordinary network. It is derived from the internet, but it is different from it in many aspects, so the security technologies that protect the internet cannot be directly applied to the industrial IoT. The following studies propose relevant security technologies based on the features of the industrial IoT.

Reference [16] proposed a risk assessment method for the industrial IoT system to protect data privacy and standardization. This method contains ten steps, with each step designed for a different vulnerability. It is designed to predict the impact of such vulnerabilities on industrial production. This method can effectively detect some of the dangers in the IoT. However, the method architecture is complex, the application conditions are strict, and it cannot be widely used in various industrial scenarios.

Reference [17] proposed an anomaly detection method for IoT by tracking traffic monitoring transactions of Transmission Control Protocol (TCP). This method is more sensitive to denial of service (DoS) intrusion attacks, and can quickly detect high traffic conditions. It can also handle some abnormal problems caused by the system's failure or human misoperation. However, this method has a high degree of dependence on the serial communication protocols, and the network abnormalities that can be handled are relatively simple.

Reference [18] proposed a security scheme that analyzes the architecture of each IoT layer and proposes security measures for the vulnerabilities existing in the other network layers (i.e., perception and transport).

Reference [19] proposed an improved K-Nearest Neighbor (KNN) algorithm in combination with the linear complexity measures which cluster unknown data by traversing the data set within the threshold range in the detection process. The advantage of this method is that it makes the calculation of the traditional KNN algorithm more convenient, improves the accuracy rate, and reduces the false alarm rate. Its disadvantage is that the initial value setting of the cluster center node has a great impact on the result of the algorithm.

Reference [20] proposed a detection method based on the k-means algorithm with information entropy. This method calculates data features based on the entropy value, compares normal data with unknown data, and makes judgments based on the comparative analysis. The advantage of this method is that the introduction of entropy makes the detection results more accurate and the calculation speed faster, but the k value of the algorithm cannot be accurately set, and the classification result requires knowing the number of categories.

Reference [21] proposes a Bayesian network detection method, which manages the database, establishes a Bayesian detection model, and filters abnormal behaviors of the network through the model. Reference [22] proposed

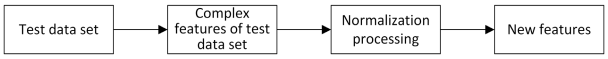


FIGURE 1. Data pre-processing.

optimization of a traffic detection model based on the Bayesian algorithm and introduced a time series to improve the accuracy of the detection model. Reference [23] established an implicit Markov chain model, analyzing and processing the network data as a semaphore, and then entering the Markov chain to detect abnormal network traffic.

III. DESIGN OF ANOMALY DETECTION ALGORITHM FOR TRAFFIC ATTRIBUTES BASED ON A CLUSTERING ALGORITHM

A. DEFINITION OF COMPLEX ATTRIBUTE FEATURES CLUSTERING

The clustering algorithm is an unsupervised machine learning algorithm. Because there is no pre-labeled data, it is necessary to make a quantitative comparison of multiple features contained in the network traffic. Traffic is divided into multiple categories and then classified according to the similarity between the new data and each existing category. The clustering algorithm for complex attributes needs to formulate a set of rules to divide the data into multiple clusters. Each cluster contains similar data and exhibits a common feature, while the data points between clusters are not similar.

Definition 1: let $X = \{X_1, X_2, \ldots, X_n\}$. The clustering algorithm divides X into K clusters: $U_1, U_2, \ldots, U_k$.

The following three conditions must be true:

- No cluster can be an empty set.
- The union of all clusters is X.
- intersection of any two different clusters is an empty set.

The clustering algorithm for complex attribute features includes three steps.

1) DATA PRE-PROCESSING

Due to a variety of network traffic attributes, the algorithm needs to face high-dimensional data processing, which is prone to large deviations in the iterations and produces an unnecessary time cost. This step can effectively extract the features and standardize the feature values to reduce the difference caused by different dimensions. In this paper, the data were normalized in the range of 0 to 1 to improve the accuracy and efficiency of the clustering algorithm when detecting abnormal network traffic, as shown in Fig.1. The test data set was represented by X. $Max(X)$ and $Min(X)$ were the maximum and minimum values of X, respectively. Any data in the data set X can be standardized into a new data value with the following equation:

$$x' = \frac{x - Min(X)}{Max(X) - Min(X)}. \tag{1}$$

2) CLUSTERING OF COMPLEX ATTRIBUTE FEATURES

In this step, data is assigned to different clusters according to specific rules. This process requires setting the number of clusters and the center of each cluster. In the clustering algorithm, the distance measurement function is a common method for measuring the similarity of complex attributes. The shorter the distance between two data points, the higher the similarity. Let $X_i = \{X_{i1}, X_{i2}, \ldots, X_{iM}\}$ and $X_j = \{X_{j1}, X_{j2}, \ldots, X_{jM}\}$ be data objects i and j in the data set X that contains M attribute features. In this paper, Mahalanobis distance is introduced, and the covariance matrix is used to calculate the distance based on the similarity between the data using (2). Mahalanobis distance is an effective method for calculating the similarity between two samples. It has the advantage of not being disturbed by the units of the original data and taking into account the relationship of various features.

$$d(X_i, X_j) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)}. \tag{2}$$

In the expression, S is the covariance matrix. When S is the identity matrix, the distance is approximately equal to the Mahalanobis distance. The distance measurement function is often used to evaluate the similarity of the multiple feature data clusters. Since each metric has a specific application scenario, the clustering algorithm produces different clustering results in different situations. Therefore, this paper uses similarity to evaluate the degree of similarity between data objects in the complex attribute feature data set X.

3) ANALYSIS OF CLUSTERING RESULTS

In this step, an evaluation function is used as an index to evaluate and analyze clustering results. Clustering algorithms with complex attributes have two conditions for ending iterations: the number of iterations of the algorithm has reached a preset maximum value or the best clustering effect has been achieved. The optimal criterion of clustering is calculated after each iteration using the evaluation function. If the algorithm meets the end condition, the iteration is terminated; otherwise, the algorithm continues until the optimal result is reached. Calculating the square error is a common method of evaluating the effect of clustering.

Definition 2: Let σ be the sum of squared errors, as shown below:

$$\sigma = \sum_{j=1}^{k} \sum_{X_i \in U_j} \|X_i - u_j\|^2. \tag{3}$$

where u is the center of the jth cluster. The smaller the σ, the smaller the distance between the data points in each cluster and cluster center, the higher the similarity, and the better the clustering effect. Therefore, lack of change in σ shows that the cluster center has the smallest distance to all data in the current cluster and that it has reached an optimal point. At this time, the algorithm stops, and the clustering is completed.

B. MATHEMATICAL MODEL

Let the data set $X = \{X_1, X_2, \ldots, X_n\}$, each data object $X_i = \{X_{i1}, X_{i2}, \ldots, X_{iM}\}(1 \leq i \leq n)$ be an m-dimensional vector with M attribute features. The $k-th$ attribute feature is

$F_k = \{x_{1k}, x_{2k}, \ldots, x_{nk}\}$, and $w_k (1 \le k \le M)$ is the weight of the $k-th$ feature attribute.

Definition 3: Let all the data form a graph G, and each node in the graph corresponds to a data object X_i in the data set X. The weighted value of the line edge e_{ij} between X_i and X_j is the product of the differences of the corresponding features and the weighted values, which can be calculated as follows:

$$W_{e_{ij}} = \sum_{k=1}^{M} \sqrt{w_k (x_{ik} - x_{jk})^2}. \tag{4}$$

Definition 4: Let $N(X_i) = \{X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(L)}\}$ be the set of L data points closest to X_i. In clustering, due to the high similarity between the adjacent and current data, the more adjacent is the data and the closer is the distance, the more the data can be characterized by the common features of the data set and the safer the data can be considered. Therefore, the safety factor $S(X_i, L)$ of the current data can be calculated according to the adjacent data in the following equation:

$$S(X_i, L) = \frac{1}{\sum_{j=1}^{L} d(X_i, X_i^{(j)}}, \quad 1 \le i \le n. \tag{5}$$

Let the weighted value of data point X_i be $W_{X_i} = s(X_i, L)$ and the centers of k clusters be set to $u_1, u_2, \ldots, u_k (u_i \in X, i = 1, 2, \ldots, k)$. Then calculate the distance y from each data point X_i in the data set X to the cluster center, match X_i to the cluster U_i, and calculate $\sum_{i=1}^{n_i} X_i / n_i$, where n_i is the number of data points in the $i-th$ cluster. This formula calculates the arithmetic mean of all data in the current cluster, namely the center of the mass which is used as the basis for adjusting the cluster center. Since the mirroring points appear in the multi-dimensional space by using the distance from the previous cluster center as the reference condition when selecting the cluster center, data of remote must be excluded. In the process of clustering iterations, the cluster center X can be adjusted based on the principle of clustering average value and optimization of the evaluation function. Finally, when the clustering center or evaluation function value does not change, $U_1, U_2, \ldots, U_k$ becomes a total of k clusters.

The anomaly detection rules of safe clustering are generated based on the cluster center, cluster size, and the average safety factor of each cluster. To improve the accuracy of anomaly detection and enhance the stability of clustering, in this paper, the weighted value w_k of attribute features, the number of adjacent points L of the clustering center, and the clustering center $U_1, U_2, \ldots, U_k$ are considered in-depth, as presented below:

1) SELECTION OF THE WEIGHTED VALUE OF TRAFFIC ATTRIBUTE FEATURES

In the selection process of w_k, priority should be given to the attribute features of the clustering process. Some redundant or irrelevant attributes of the test data might reduce the classification accuracy and increase the time of calculation. Because of this, this paper proposes an evaluation method that measures the importance of attribute features. This method converts the importance into a weighted value w_k and aggregates the key attributes into a set, simplifying the feature calculation of the attributes.

Definition 5: In this paper, the attributes included in the process of anomaly detection include source/destination IP, source/destination port, a service type field, a protocol type field, and a logical input port. Information entropy can be analyzed to ensure the flow remains a discrete information source and its attributes remain discrete events. Let the $k-th$ attribute feature of n flows be $F_k = \{x_{1k}, x_{2k}, \ldots, x_{nk}\}$, and let $p(x_{ik})$ be the frequency of occurrence of F_k in the flow concentration. The information entropy can be calculated with (6), where the larger the information entropy, the higher the value of the attribute feature.

$$H(F_k) = \sum_{x_{ik} \in F_k} p(x_{ik}) \log \frac{1}{p(x_{ik})}. \tag{6}$$

Definition 6: When the value y of the attribute feature F_k is given, the conditional information entropy of F_k can be expressed as follows:

$$H(F_{k'}|F_k) = -\sum_{x \in F_k} p(x) \sum_{u \in F_{k'}} p(y|x) \log p(y|x). \tag{7}$$

Definition 7: Based on the information entropy and conditional information entropy, the mutual information can be calculated as follows:

$$I(F_k, F_{k'}) = H(F_k) - H(F_{k'}|F_k). \tag{8}$$

Definition 8: The degree of relevance of F_k can be calculated in the following manner based on the average mutual information of other attribute features and F_k, which represents the degree of correlation between the $k-th$ feature and other features:

$$R(F_k) = \frac{1}{M} \sum_{k'=1}^{M} I(F_k, F_{k'}). \tag{9}$$

The conditional correlation between $F_{k'}$ and condition F_k can be calculated as follows:

$$R(F_{k'}|F_k) = R(F_k) \frac{H(F_{k'}|F_k)}{H(F_k)}. \tag{10}$$

The redundancy $Red(F_k, F_{k'})$ can be calculated as follows:

$$Red(F_k, F_{k'}) = R(F_k) - R(F_{k'}|F_k). \tag{11}$$

On this basis, the importance of attribute features is defined as follows:

$$I_m(F_k) = R(F_k) - max\{Red(F_k, F_{k'})\}. \tag{12}$$

The weighted value of the attribute features is calculated in the following equation:

$$w_k = \frac{I_m(F_k)}{\sum_{m=1}^{M} I_m(F_k)}. \tag{13}$$

(13) satisfies $0 \le w_k \le 1$ and $\sum_{k=1}^{M} w_k = 1$.

2) SELECTION OF THE NUMBER L OF ADJACENT DATA POINTS

In current commonly-used algorithms, L is usually set as an empirical value, which leads to low detection efficiency and large calculation errors. This paper proposes to select the L value to obtain the best detection accuracy and efficiency.

The safety factor $s(X_i, l)$ of the data points is a monotonically-decreasing sequence, and the proof process is offered as follows: Given $s(X_i, l) = \frac{1}{\sum_{j=1}^{l} d(X_i, X_i^{(j)})}, 1 \leq i \leq n$, we know that the denominator is not 0, and the distance function $d(X_i, X_i^{(j)}) > 0$. It can be inferred from the above that the denominator continually increases and is a monotonically-increasing sequence about l, so it can be proved that the safety factor $s(X_i, l)$ is a monotonically-decreasing sequence.

When l is equal to a different value, the safety factor of the overall data point can be calculated as follows:

$$S(l) = \sum_{i=1}^{n} s(X_i, l), \quad 1 \leq l \leq n-1. \tag{14}$$

Therefore, $S(l)$ is also a monotonically-decreasing sequence of l.

Since the increase of $S(1) \geq S(2) \geq \ldots \geq S(n-1)$ and l will cause the decrease of $S(l)$, the relative change of decrease can be calculated as follows:

$$\triangle_L = S(l-1) - S(l), \quad 2 \leq l \leq n-1. \tag{15}$$

When $\triangle_L$ and l make the relative change reach its maximum value, the safety factor reaches its highest value, and l is chosen as the required L value.

3) SELECTION OF CLUSTER CENTERS

Different cluster centers produce different results. To improve the stability and accuracy of abnormal data detection, it is important to choose an accurate initial clustering center set. The method proposed in this paper selects the initial cluster center which has features of uniform distribution. The safety threshold δ is set as the critical value of the safety factor. When it satisfies $S(X_i, l) \geq \delta$, the data point X_i has a high safety factor. The set of data points with a high safety factor is represented by U – a data set used for selecting the cluster center. The initial cluster centers obtained by this method have a higher safety factor, which can improve the safety of anomaly detection. The algorithm for the selection of cluster centers is as follows:

Step 1. The data point with the highest safety factor selected from the data set U is represented by u_1, which is the cluster center of the cluster U_1.

Step 2. Select the data point farthest from u_1 in the data set U and denoted by u_2, which is the cluster center of the cluster U_2.

Step 3. Continue to select cluster centers according to step 1 and step 2. Finally, k initial cluster centers with high safety factors are obtained.

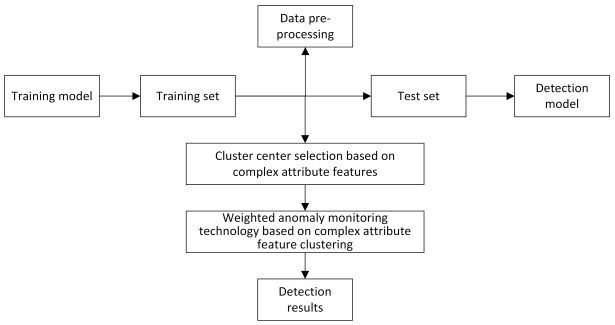


FIGURE 2. Security clustering exception detection method for complex attribute features.

C. ANOMALY DETECTION BASED ON COMPLEX ATTRIBUTE FEATURE CLUSTERING

The method of anomaly detection based on the clustering of complex attribute features proposed in this paper studies mainly the extraction and pre-processing of the useful data attribute features from abnormal traffic. The results are analyzed based on experience and actual conditions. The clustering algorithm used for abnormal traffic detection in the industrial IoT is generally a specific machine learning algorithm, which marks the cluster as normal or abnormal according to the distribution of data points in each cluster. The cluster center, cluster size, and average safety factors are used to formulate detection rules.

The detection diagram of complex attribute features is shown in Fig.2. First the traffic attribute anomaly detection pre-processes the data (i.e. standardizes calculation). Then the cluster center is selected to generate the clustering model. Finally, property anomaly detection is carried out according to the source/destination port, source/destination IP, a service type field, protocol type field, and logical input port.

The core idea of the selection algorithm of cluster centers is presented below. In the data set $X = \{X_1, X_2, \ldots, X_n\}$, each data object is $X_i = \{x_{i1}, x_{i2}, \ldots, x_{iM}\}$, which is an M-dimensional vector that has M number of attribute features. Taking into consideration the important attribute feature F_k and the weighted value w_k, the weighted distances and safety factors of all data objects are calculated first. Second, the uniformly-distributed cluster centers are selected. Finally, each data object X_i in the data set X is added to the cluster U_i where the nearest cluster center u_i is located. The optimal value of the arithmetic means and the sum of squared errors σ of clustering values is used to adjust the clustering center. Finally, when the clustering center no longer changes, the cluster $U_1, U_2, \ldots, U_k$ is generated. The pseudo code of the selection algorithm of cluster centers is as follows.

To improve the stability and accuracy of clustering, it is crucial to select an appropriate cluster center. The selection process of the cluster center is shown in Fig.3 (1) and (2). After generating four cluster centers, the fifth one is generated by the algorithm. During training, the operating flow of the system can be constructed through normal behavior. In the detection phase, the training set is compared to the test set, and the degree of deviation is calculated. If the deviation

Algorithm 1 Selection Algorithm of Cluster Centers

Input: X, δ, k
Output: The initial position of cluster center $u_1, u_2, \ldots, u_k$
Steps:
1. **While** $1 \leq i \leq n$ **do**
2. Calculate the safety factor $S(X_i, L)$ of each data point;
3. **If** $S(X_i, L) \geq \delta$
4. Add X_i to the set U of high safety nodes;
5. **EndIf**
6. **EndWhile**
7. Select the data point with the highest safety factor from U as the first cluster center u_1;
8. **While** $2 \leq i \leq k$ **do**
9. Select the data point furthest from u_{i-1} as the second cluster center u_i;
10. **EndWhile**
11. Return the initial position of cluster center $u_1, u_2, \ldots, u_k$

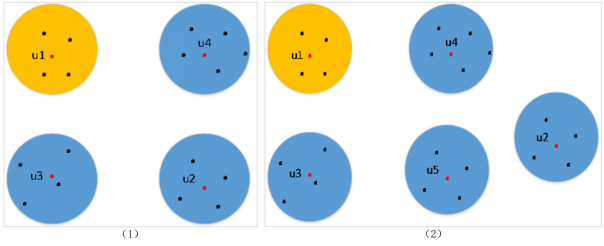


FIGURE 3. Cluster center selection.

exceeds a preset threshold, the current network traffic is marked as abnormal. The clustering algorithm proposed in this paper can detect new network traffic anomalies while ensuring effective detection.

In the anomaly detection of the industrial IoT, the detection model and the safety threshold are the keys to diagnosis. The main function of the detection model is to mark network traffic as normal or abnormal. The pseudo code of the weighted clustering algorithm for complex attribute features is as follows.

D. TIME COMPLEXITY ANALYSIS OF THE ALGORITHM

The secure clustering algorithm for complex attribute features proposed in this paper has good processing efficiency for large-scale data sets. The clustering algorithm has natural adaptability to high-dimensional data, so it is widely used in processing network traffic. The algorithm calculates the safety factor based on the similarity between the data, uses the safety factor and the arithmetic mean of the clusters as a reference to select the cluster center, and considers the priority of the data in the process of matching the data to the corresponding cluster. To improve the stability and accuracy of clustering, once the iteration termination condition is reached, the algorithm stops. At this time, the clusters show some similar features, but the clusters are independent of each other.

Algorithm 2 Weighted Clustering Algorithm for Multi-Feature Data

Input: X, δ, k;
Output: Clustering result $U_1, U_2, \ldots, U_k$
Steps:
1. Pre-processing the data set X, namely standardized calculation;
2. **While** $1 \leq k \leq M$ **do**
3. Calculate the weighted value w_k of the attribute feature F_k;
4. **EndWhile**
5. **While** $1 \leq i, j \leq n$ **do**
6. Calculate the Mahalanobis distance $d(X_i, X_j)$ between any two data in the data set X;
7. **EndWhile**
8. **While** $1 \leq i \leq n$ **do**
9. Calculate the safety factor $S(X_i, L)$ of data X_i;
10. Calculate the sum of safety factor $S(l)$ of all data points
11. **EndWhile**
12. $max = S(1) - S(2)$;
13. $index = 2$;
14. **While** $2 \leq l \leq n-1$ **do**
15. $\Delta_l = S(l-1) - S(l), 2 \leq l \leq n-1$;
16. **If** $\Delta_l > max$
17. $max = \Delta_l$;
18. $index = l$;
19. **EndIf**
20. **EndWhile**
21. Set index to the optimal number of adjacent data L;
22. Call the cluster center selection algorithm to calculate the cluster center $u_1, u_2, \ldots, u_k$;
23. **While** $1 \leq i \leq n$ **do**
24. Add X_i to the cluster U_i where the nearest cluster center u_i is located;
25. Calculate the arithmetic mean $\frac{1}{n_i}\sum_{i=1}^{n_i}(X_i)$ of all data in the cluster;
26. Adjust the cluster centers according to the arithmetic mean;
27. **If** the cluster center does not change anymore
28. break;
29. **EndIf**
30. **EndWhile**
31. Return Clustering result $U_1, U_2, \ldots, U_k$;

The time complexity of the secure clustering algorithm for complex attribute features proposed in this paper is $O(n*k*t*m)$, where n represents the number of samples in the data set, t represents the final number of iterations of the algorithm, k represents the number of final divided clusters, m represents the number of attribute features in each data point. In general, t, m, and k are constants by default and do not affect the time complexity. This indicates that the time complexity of the secure clustering algorithm can be simplified to $O(n)$, and the

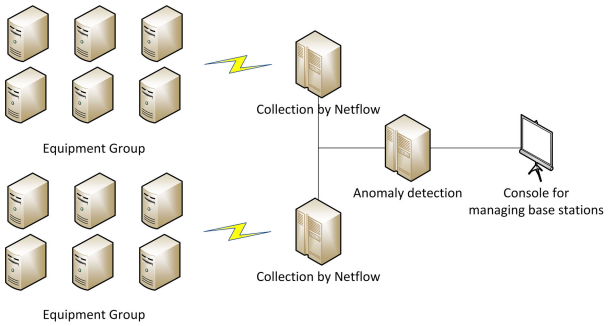


FIGURE 4. Simulation environment.

time complexity can be approximated as linear, that is, only related to the number of samples in the data set.

IV. SIMULATION OF A HIERARCHICAL DETECTION METHOD

A. ABNORMAL TRAFFIC DATA SET

This study simulated an operating environment of an abnormal traffic detection system in the network and used the Canadian Institute for Cybersecurity (CIC) data set to test the system. The CIC data set is a collaborative project between the Communications Security Establishment (CSE) and the Canadian Institute for Cybersecurity. Similar to the real-world data sets, the CIC-IDS-2018 data set contains benign data and the types of the latest attacks. The data set contains normal data for comparison, as well as the information on the abnormal traffic and traffic attacks, such as DoS, Heartbleed, Web attacks, Botnet, and infiltration of the network from inside. In the windows environment, the CIC Flow Meter feature extraction tool is used to obtain the flow set information. The abnormal flow information includes the host's IP address and the start and end times of the flow. Flow features include the source address and destination address of the data packet, data packet size, average data packet size, transmission time, average transmission time, etc.

B. ESTABLISHMENT OF THE EXPERIMENTAL ENVIRONMENT

An OPNET network simulation helps to avoid the risk of building a network in a real environment, and at the same time greatly reduces the experimental overhead. Thus, it is widely used in experimental network simulations. In this paper, OPNET was used to simulate the traffic frequency change at the time of abnormality in the network. The simulation environment was built as shown in Fig.4.

The experiment simulated a large amount of traffic in a network, resulting in its congestion and collapse after an anomaly occurs in the network. The traffic anomaly detection designed in this paper adopted a hierarchical detection method, using Netflow technology to collect traffic information, do pre-processing, extract seven-tuples from the original traffic, and establish a mapping between the seven-tuples and the original traffic in Redis. The seven-tuple included source address IP, destination address IP, source port number, destination port number, a protocol type field, a service type field, and a logical input interface. The flow frequency abnormality detection mainly detected whether the flow frequency fluctuated within the normal range. This article set a two-layer threshold interval and divided the traffic into three categories according to the frequency. When the traffic was dangerous, it reported directly to the management base station; when the flow was suspicious, the flow frequency detection was diagnosed as abnormal, and then the flow attribute detection was performed; when the flow was safe, it was skipped. Traffic attribute detection used a clustering algorithm to detect abnormal fields. If the traffic frequency detection and the traffic attribute detection were both diagnosed as abnormal traffic, the administrator was alerted.

TABLE 1. Experimental parameters.

Variable name	Interpretation
w_k	The weighted value of the $k-th$ attribute feature
$d(X_i, X_j)$	The distance between data X_i and X_j
L	Number of adjacent data points of data X_i
$S(X_i, l)$	Safety factor of data X_i
Δ	Safety threshold
K	Number of clusters
A_1	The number of undetected abnormal data points
A_2	The number of abnormal data points detected
N_1	Number of normal data points recognized
N_2	The number of normal data points misjudged as abnormal

C. SIMULATION OF ABNORMAL DETECTION OF TRAFFIC ATTRIBUTES

1) EXPERIMENTAL PARAMETERS

The CIC data set provides multiple types of abnormal traffic and normal traffic for comparison, which establishes the basis for evaluating the reliability of abnormal traffic detection. The experimental parameters of the CIC data set are shown in table 1.

2) EVALUATION INDEX OF DETECTION RATE

The security of traffic attribute anomaly detection can be evaluated with true-positive (TP) and false-positive (FP) cases. TP shows that a positive class is correctly judged as a positive class, and FP shows that a negative class is wrongly judged as a positive class. In the simulation, the TP rate was the ratio of accurately-identified abnormal data to the total abnormal data, and the FP rate was the ratio of normal data that was misjudged as abnormal to the total normal data, as shown below:

$$TP = \frac{A_2}{A_1 + A_2}. \tag{16}$$

$$FP = \frac{N_2}{N_1 + N_2}. \tag{17}$$

Based on the classification of attributes in different dimensions, this experiment selected five samples from each cluster to construct a training set and used the above samples to evaluate the clustering effect, as shown in Fig.5 and Fig.6. As a comparative experiment, the number of set attribute features was increased each time by 20, going from 0 to 100. The traffic attribute anomaly detection algorithm proposed in this paper was compared with the k-means algorithm and the

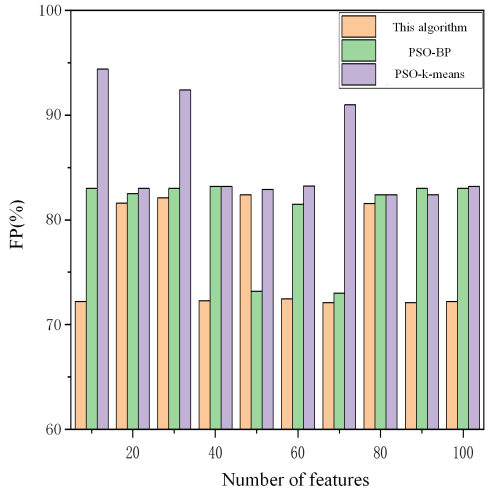


FIGURE 5. Comparative and analysis of FP rate.

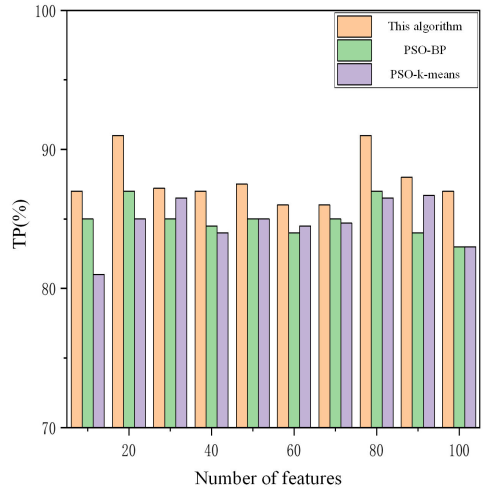


FIGURE 6. Comparative and analysis of TP rate.

Back Propagation (BP) neural network algorithm both based on the Particle Swarm Optimization (PSO).

Comparative experiments demonstrated that the traffic attribute anomaly detection algorithm proposed in this paper has the lowest FP rate. It is 0.97% lower than that of the k-means algorithm based on PSO, and it is 0.47% lower than that of the BP algorithm based on PSO. The proposed algorithm's TP rate is the highest. It is 3.08% higher than that of the k-means algorithm based on PSO and 2.82% higher than that of the BP algorithm based on PSO. In summary, the algorithm proposed in this paper has higher security for the anomaly detection of network traffic attributes.

The core of the clustering algorithm proposed in this paper is the safety factor of the data. Fig.7 shows the influence of data X_i and l on the safety factor of the algorithm in the same data set, where X_i is the number i of the data set, and l is the number of the adjacent data of X_i. The figure shows that with the anomaly detection algorithm proposed in this paper, for the same data X_i, the more the data is adjacent, the bigger its safety factor is. This indicates that the higher the information

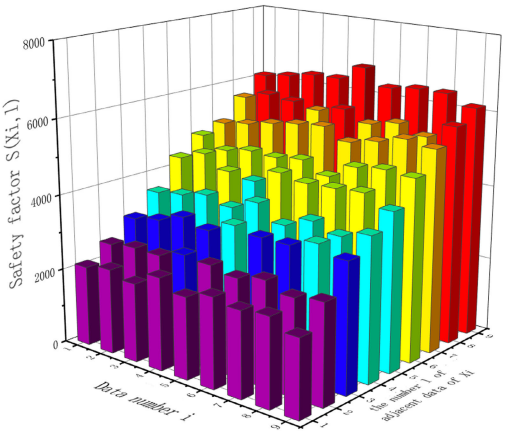


FIGURE 7. Influence of parameters on safety factor.

TABLE 2. Experimental parameters.

Attack types	Training time(s)	PSO-k-means(s)	PSO-BP(s)	This algorithm(s)
DoS	5733	5528	5261	5147
Probe	1486	1393	1207	1182
U2R	101	86	59	38
Average elapsed time	2440	2336	2176	2122

entropy of the data X_i, the more valuable it is. The larger the cluster formed in the clustering process, the safer the data X_i.

3) EVALUATION INDEX OF DETECTION TIME

The abnormal traffic detection of the industrial IoT requires a high speed. This paper used three abnormal traffic evaluations and detection times, comparing them with the k-means algorithm based on PSO and the BP algorithm based on PSO. The detection times are shown in Table 2.

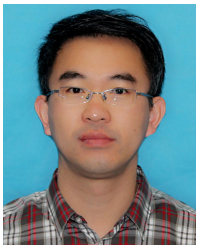
Three types of abnormal traffic (Dos, Probe, and U2R) were used to evaluate the detection time. As can be seen from the table, the average detection time of the k-means algorithm optimized with PSO is 2336s, and the average detection time of the BP algorithm optimized with PSO is 2176s. The algorithm proposed in this paper has the shortest average detection time of 2122s.

V. CONCLUSION

To detect traffic attributes, this paper designs a security clustering algorithm for complex attribute features. The algorithm calculates the security factor based on the Mahalanobis distance between data and the number of adjacent data points, gets the weighted value of data points according to the security factor, sets the security threshold, selects the security data points as the cluster center, and analyzes the data according to the priority of traffic attribute features. The weighted distance and safety factors are classified, and the data points are divided into corresponding clusters. The simulation experiment demonstrates that the anomaly detection method based on security clustering has higher detection accuracy and efficiency than other algorithms.

REFERENCES

[1] L. Khatibzadeh, Z. Bornaee, A. G. Bafghi, and Y. Yuan, "Applying catastrophe theory for network anomaly detection in cloud computing traffic," *Secur. Commun. Netw.*, vol. 16, no. 8, pp. 46–51, May 2019.
[2] A. P. Das, S. M. Thampi, and J. Lloret, "Anomaly detection in UASN localization based on time series analysis and fuzzy logic," *Mobile Netw. Appl.*, vol. 25, no. 1, pp. 55–67, Feb. 2020.
[3] A. R. Sadeghi, C. Wachsmann, and M. Waidner, "Security and privacy challenges in industrial Internet of Things," in *Proc. 52nd ACM/EDAC/IEEE Design Automat. Conf.*, Jun. 2015, pp. 1–6.
[4] Y. Liu, K. Tong, F. Mao, and J. Yang, "Research on digital production technology for traditional manufacturing enterprises based on industrial Internet of Things in 5G era," *Int. J. Adv. Manuf. Technol.*, vol. 107, nos. 3–4, pp. 1101–1114, Mar. 2020.
[5] X. Hu and J. Wang, "Detection algorithm of abnormal behavior in security for IIoT based on random inspection," *Int. Core J. Eng.*, vol. 6, no. 5, pp. 336–343, 2020.
[6] D. Kwon, H. Kim, J. Kim, C. S. Suh, I. K. Kim, and J. K. Kim, "A survey of deep learning-based network anomaly detection," *Cluster Comput.*, vol. 22, no. 1, pp. 35–38, 2019.
[7] J. Wurm, K. Hoang, O. Arias, A.-R. Sadeghi, and Y. Jin, "Security analysis on consumer and industrial IoT devices," in *Proc. 21st Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Jan. 2016, pp. 519–524.
[8] D. Patel, K. Srinivasan, C.-Y. Chang, T. Gupta, and A. Kataria, "Network anomaly detection inside consumer networks—A hybrid approach," *Electronics*, vol. 9, no. 6, p. 923, Jun. 2020.
[9] N. Dey, H. AE, C. Bhatt, A. AS, and S. SC, "Bigdata analytics in industrial IoT," in *Proc. Internet Things Big Data Anal. Toward Next-Gener. Intell.*, 2018, pp. 381–406, doi: 10.1007/978-3-319-60435-0_16.
[10] X. Yin, X. Chen, L. Chen, H. Li, and V. Milutinovic, "Extension of research on security as a service for VMs in IaaS platform," *Secur. Commun. Netw.*, vol. 18, no. 3, pp. 30–33, 2020.
[11] Y. Zhong, W. Chen, Z. Wang, Y. Chen, K. Wang, Y. Li, X. Yin, X. Shi, J. Yang, and K. Li, "HELAD: A novel network anomaly detection model based on heterogeneous ensemble learning," *Comput. Netw.*, vol. 169, Mar. 2020, Art. no. 107049.
[12] A. Nagaraja, U. Boregowda, K. Khatatneh, R. Vangipuram, R. Nuvvusetty, and V. S. Kiran, "Similarity based feature transformation for network anomaly detection," *J. Robot. Mach. Learn.*, vol. 55, no. 26, pp. 100–103, 2020.
[13] N. Chouhan, A. Khan, and H.-U.-R. Khan, "Network anomaly detection using channel boosted and residual learning based deep convolutional neural network," *Appl. Soft Comput. J.*, vol. 83, no. 35, pp. 77–79, 2019.
[14] S. Kim, W. Jo, and T. Shon, "APAD: Autoencoder-based Payload Anomaly Detection for industrial IoE," *Appl. Soft Comput. J.*, vol. 88, no. 56, pp. 96–98, 2020.
[15] T. Bodström and T. Hämäläinen, "A novel deep learning stack for APT detection," *Appl. Sci.*, vol. 9, no. 6, pp. 99–102, Feb. 2019.
[16] M. J. Zhao, A. R. Driscoll, S. Sengupta, R. D. Fricker, D. J. Spitzner, and W. H. Woodall, "Performance evaluation of social network anomaly detection using a moving window-based scan method," *Qual. Rel. Eng. Int.*, vol. 34, no. 8, pp. 1699–1716, Dec. 2018.
[17] Z. Elkhadir and B. Mohammed, "A cyber network attack detection based on GM median nearest neighbors LDA," *Comput. Secur.*, vol. 86, no. 53, pp. 73–78, 2019.
[18] F. Palmieri, "Network anomaly detection based on logistic regression of nonlinear chaotic invariants," *J. Netw. Comput. Appl.*, vol. 148, no. 3, pp. 68–73, 2019.
[19] S. Baek, D. Kwon, C. Sang Suh, H. Kim, I. Kim, and J. Kim, "Clustering-based label estimation for network anomaly detection," *Digit. Commun. Netw.*, vol. 33, no. 22, pp. 68–70, 2020.
[20] W. Han, J. Xue, and H. Yan, "Detecting anomalous traffic in the controlled network based on cross entropy and support vector machine," *IET Inf. Secur.*, vol. 13, no. 2, pp. 18–19, 2019.
[21] Y. Gormez, Z. Aydin, R. Karademir, and V. C. Gungor, "A deep learning approach with Bayesian optimization and ensemble classifiers for detecting denial of service attacks," *Int. J. Commun. Syst.*, vol. 33, no. 11, p. e4401, May 2020.
[22] Y. Chen, Y. Jiang, Z. Yang, Y. Wang, and P. Lin, "Research on network anomaly detection based on coordinated multiple algorithms," *IOP Ser., Mater. Sci. Eng.*, vol. 466, no. 1, 2018, Art. no. 012027.
[23] C. Seelammal and K. Vimala Devi, "Multi-criteria decision support for feature selection in network anomaly detection system," *Int. J. Data Anal. Techn. Strategies*, vol. 10, no. 3, pp. 334–350, 2018.



LIDONG FU received the B.E. degree from the Changsha University of Science and Technology, in 2000, and the M.E. degree from Northeast University, China, in 2008. He is currently an Associate Professor with the School of Information Science and Engineering, Shenyang Ligong University. His research interests include computer networks, geographic information, and the Internet of things.

WENBO ZHANG received the Ph.D. degree in computer science from Northeastern University, China, in March 2006. He is currently a Professor with the School of Information Science and Engineering, Shenyang Ligong University, China. He has published over 100 papers in related international conferences and journals. His current research interests include ad hoc networks, sensor networks, satellite networks, and embedded systems. He had been awarded the ICINIS 2011 Best Paper Awards and up to nine the Science and Technology awards, including the National Science and Technology Progress Award and the Youth Science and Technology awards from the China Ordnance Society. He has served as an Editorial Board for up to ten journals, including Chinese *Journal of Electronics* and *Journal of Astronautics*.

XIAOBO TAN received the B.E. degree from Liaoning Normal University, in 2000, and the M.E. degree from Northeast University, China, in 2006. He is currently an Associate Professor with the Communication and Network Institute and the School of information science and Engineering, Shenyang Ligong University. His research interests include wireless sensor networks and embedded systems.

HONGBO ZHU (Member, IEEE) received the B.Sc., M.Eng., and Ph.D. degrees from Northeastern University, Shenyang, China, in 2009, 2012, and 2020, respectively. He is currently an Associate Professor with the School of Information Science and Engineering, Shenyang Ligong University. His research interests include medical image computing and deep learning.