An Industrial Network Intrusion Detection Algorithm Based on Multifeature Data Clustering Optimization Model

Wei Liang, Kuan-Ching Li, *Senior Member, IEEE*, Jing Long, Xiaoyan Kui, and Albert Y. Zomaya

***Abstract*—Industrial networks are complex and diverse. Among existing intrusion prevention systems available, several of them have problems such as low detection accuracy rate, high false positive (FP) rate, and low real-time performance for impersonation attacks. To address such issues, it is proposed in this article an industrial network intrusion detection algorithm based on multifeature data clustering optimization model, where the weighted distances and security coefficients of data are classified based on the priority threshold of data attribute feature for each node in the network, given that the data modules in the industrial network environment are diverse and easy to diagnose, restore, and rebuild. The proposed algorithm can effectively improve the detection rate and real-time performance of detecting abnormal behavior for the multifeature data in industrial networks. The novel features are twofold, to rapidly select a node with high-security coefficient as the cluster center, and match the multifeature data around the center into a cluster. Experimental results show that the proposed algorithm has good superiority in terms of detection rate and time compared to other algorithms. In the industrial network, the detection accuracy of abnormal data reaches 97.8%, and the FP of detection is decreased by 8.8%.**

***Index Terms*—Clustering, industrial network, intrusion detection, multifeature, weighted distance.**

I. Introduction

With the deep integration and fast development of industry informatization, the industrial networks frequently suffer from illegal intrusion attacks, and the types of these attacks are becoming increasingly diverse and complex. The data in industrial networks have the characteristics of diversity, easy to diagnose, and rebuild. As a consequence, fast detection and prevention of abnormal behavior caused by the intrusion in the next generation of industrial networks have attracted great attentions from the world-wide governments and industrial companies.

Numerous attack events in the industrial network have been reported in recent years that caused severe consequences. For instance, the Stuxnet virus seriously threatened essential computing facilities of various countries in 2010 [1], such as the hydro-power stations and nuclear power networks. Iran suffered from the most severe attacks, as its nuclear power equipment was severely damaged. In due course, over 60% of personal computers and devices were attacked by the Stuxnet virus. In 2015, HawkEye RAT [2] intruded computer systems of enterprises to steal the core system access information. In the same year, the power system of Ukraine suffered from malicious attacks [3], which caused several hours of the power outage in a vast region. The Wanna Cry virus [4] swept all over the world in 2017, and affected many Chinese industrial companies, such as PetroChina. In summary, the security risks of the global industrial network are continuously increasing, and as a matter of fact, there happened several other attack events in addition to the abovementioned ones.

The basic structure of intrusion detection is shown in Fig. 1. Network activities that attempt to damage the integrity, confidentiality, and availability of industrial networks by illegal intrusion means can be treated as intrusion attacks in industrial networks. Intrusion detection systems can detect and trace the intrusion events actively, though impossible for network defense technologies. Therefore, the use of intrusion detection systems in industrial networks can remedy the deficiency of traditional

Manuscript received May 15, 2019; revised July 17, 2019 and September 3, 2019; accepted September 19, 2019. Date of publication October 11, 2019; date of current version January 16, 2020. This work was supported in part by the National Science Foundation of China under Grant 61572186 and Grant 61572188, in part by the Scientific Research Program of the New Century Excellent Talents in Fujian Province University, Fujian Provincial Natural Science Foundation of China under Grant 2018J01570, in part by the Hunan Provincial Natural Science Foundation under Grant 2016jj2058, and in part by the 13th Five-Year Plan of Education Science Program of Hunan Province under Grant XJK17BXX004. Paper no. TII-19-1896. *(Corresponding author: Wei Liang.)*

W. Liang is with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China, and also with the School of Opto-Electronic and Communication Engineering, Xiamen University of Technology, Xiamen 361024, China (e-mail: wliang@xmut.edu.cn).

K.-C. Li is with the Department of Computer Science and Information Engineering, Providence University, Taichung 43301, Taiwan (e-mail: kuancli@pu.edu.tw).

J. Long is with the College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China (e-mail: jlong@hunnu.edu.cn).

X. Kui is with the School of Computer Science and Engineering, Central South University, Changsha 410083, China (e-mail: xykui@csu.edu.cn).

A. Y. Zomaya is with the School of Information Technologies, University of Sydney, Sydney, NSW 2006, Australia (e-mail: albert.zomaya@sydney.edu.au).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2019.2946791

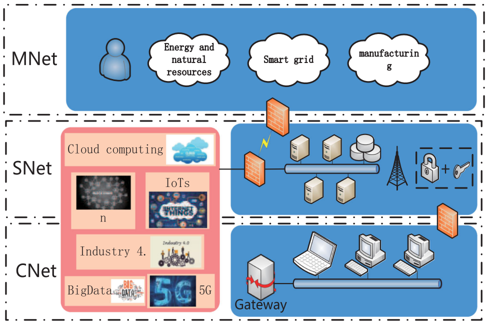


Fig. 1. Structure of an intrusion detection system in industrial networks.

network defense strategies, thereby perfecting the entire security system of industrial networks.

Industrial network detection techniques can be divided into four categories, namely, rule-based, neural network-based, support vector machine (SVM)-based, and clustering analysis-based intrusion detection. Rule-based detection one transforms the intrusion rules into related structures that performance relies on the integrity of the rule repository, which in turn depends on the integrity and real-time performance of audit records. They have a high accuracy rate and low false positive (FP) in detection; though, the detection speed is slow. The structure of artificial neural networks is similar to the synaptic connection of the human brain, as they are a nonlinear and self-adaptive information-processing system with many interconnected processing units. Neural network models are varied due to the differences in network topology, neuron characteristic, and learning rules. SVM-based intrusion detection techniques aim to address various issues in learning, classification, and prediction, as it transforms the input space into a high-dimensional space by the nonlinear mapping, and the optimal separating plane is established in the high-dimensional area. Finally, clustering analysis can discover the global distribution and the inner structure of each cluster in the clustering-analysis-based intrusion detection techniques. That is, the structure can recognize whether an individual belongs to a cluster, as it considers the sample data without classification indication. This technique aims to find an inner structure of a cluster that makes the sample in the same cluster similar, and samples among different clusters disparate, in which such a structure can be used to match and recognize the detected data. However, there are still several security issues in intrusion detection techniques to be addressed, as follows.

1) The use of the multifeature classification model in industrial networks does not occupy extra bandwidth and concealed, since the features of data in industrial networks are diverse. Any slight abnormal change in the network may easily lead to considerable security risks.
2) Intrusion attacks in industrial networks are concealed and include significant intrusion information, thereby causing substantial damage to industrial networks. It is challenging to prevent or eliminate the intrusion attacks with mixed types through the existing firewall, antivirus software, and network detection in industrial networks.
3) Existing intrusion detection techniques for industrial network belong to an active defense that has low detection efficiency and high rates of miss and false detection in a passive environment. As theoretical researches on the feature of industrial networks and the generality of intrusion signal are limited, potential security problems, such as security hole, still exist in industrial networks.

Advanced models in intrusion detection utilize the supervised learning algorithm. For such, plenty of training data with category mark should be provided to realize the correct classification of intrusion attacks. If the category mark is mistaken, an incorrect trained intrusion detection model will be generated. Clustering is the most common form of unsupervised learning, as it can analyze the global distribution and inner structure of each cluster in the data sample. The structure can recognize whether the sample belongs to a given category. In this case, the clustering analysis is introduced to intrusion detection and used for rapid and correct recognition of abnormal network behavior with the unmarked industrial network connection records.

The remaining of this article is organized as follows. Section II introduces related works and analyzes the shortcomings of existing algorithms, while definitions are introduced and the mathematical model is detailed in Section III, as well-presented the intrusion detection algorithm based on this model. The experimental results are discussed in Section IV, and finally, Section V concludes this article.

II. Related Work

Several attack tools can be used for network information hiding. Terrorist organizations or illegal attackers may employ them to hide specific secrets in various data carriers, so the intrusion attacks may arise when the carriers are under operation in industrial networks. The consequences of these attacks are unimaginable—paralyze the industrial economy or cause public panic. Thence, it is still difficult to deal with passive intrusion attacks with the current technologies. Countries such as Great Britain, Russia, Germany, and France have attempted to research concealed intrusion and intrusion detection techniques. Based on these issues, network security has brought attention, turning into research hotspot in most developed/developing countries. Nevertheless, the researches on the security mechanism of intrusion detection systems for industrial networks are still in its infancy, which is mainly focused on the prevention of active concealed intrusion attacks.

Many security issues in industrial networks need to be addressed [5], [6], such as whether node data has maliciously tampered after a node is successfully connected into an industrial network, whether the data is integrated, or how to accurately detect the abnormal behavior of intrusion. A protocol analysis method based on a decision tree is proposed, where each level of the protocol is decoded [7]. Intrusion detection is simplified into monitoring several fields and calling related functions. The method can accurately capture the intrusion signal with the protocol feature, so the performance of the intrusion detection system is greatly improved. In [8], authors proposed a machine learning-based intrusion detection method where the collected data can be divided into training sets and detection sets. Herein, a secure and reliable model is generated by using the training set, as the detection set is used for experimental analysis. In [9], [10], authors used neural network for misuse detection. The system detects intrusion attacks by searching the key code of attacks in a network flow. A multilayer perceptron is used to detect intrusion in the mainframe, including known attacks and unknown attacks. In [11], a data feature collection method is used to detect the abnormal parameter value in intrusion, by preset a secure threshold for each parameter.

Values higher than the threshold are treated as abnormal. This method can detect the abnormal parameter values in a single procedure yet may cause a high FP rate of detection. Gao *et al.* proposed an SVM-based intrusion detection model and discussed the process of establishing the model by using system call execution [12]. Another network intrusion detection technique based on passive learning was proposed in [13], where one-class SVM was used to establish the intrusion detection model. Such a proposed technique has good robustness through low security is achieved. The clustering method is applied in data connection [14], by assuming that clusters with fewer samples are probably abnormal clusters. Unsupervised learning is considered to learn the normal network behavior. Thereby, a small number of abnormal data is reserved to avoid the abundant abnormal data being gathered as a significant cluster in the training data set. After clustering, clusters exceeding a specific scale will be regarded as normal behavior. When a connection record is detected, the similarity between the record and the normal behavior is used to determine whether it is abnormal. Unsupervised clustering algorithm is proposed in [15], where parameters need not be set artificially and are unaffected by order of data input. The shape of a cluster is arbitrary and reflects the real data distribution, as the algorithm determines the constant data cluster by comparing the distance among training sets without any indication to the class.

A window-based feature extraction technique for intrusion detection is proposed in [16], with the feed-forward neural network and back-propagation training algorithms used to model the standard behavior edge of data feature in the network. However, it needs to learn the edge of normal behavior again before the appearance of new behavior, and a significant amount of time is consumed in the training stage. Several intrusion detection algorithms based on machine learning are proposed for industrial networks [17]–[19], in which techniques in machine learning are utilized to model the spatial, temporal correlativity of network data attribute. Such approaches are capable of improving the detection quality and decreasing the FP rate. However, the network attacks become complex due to the diversity and complexity of industrial network environment. In consequence, it is a big challenge to address the security and defense issues in industrial networks.

III. Intrusion Detection Algorithm Based on Multifeature Data Clustering Optimization Model

A. *Definition of Multifeature Data Clustering*

The definition of clustering given by Merriam–Webster is a multifeature classification technology based on statistics [20]. The multiple features are compared quantitatively to determine the individuals within a category; that is, it is an unsupervised learning method. Multifeature clustering requires a rule set to divide data into several categories, where the data within a category are similar, and those among categories are disparate.

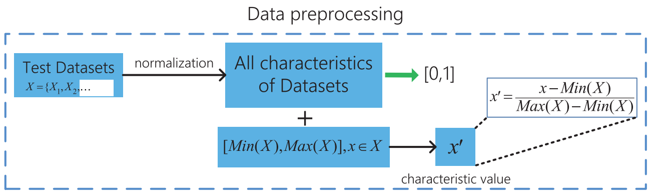


Fig. 2. Chart of data preprocessing.

Definition 1: Let $\mathfrak{R}$ be a cluster. A data set $X = \{X_1, X_2, \ldots, X_n\}$ is assumed, and the cluster $\mathfrak{R}$ of X is to divide X into K sets, $U_1, U_2, \ldots, U_K$.

1) $U_i \neq \phi, i = 1, 2, \ldots, K$;
2) $U_1 \cup U_2 \cup \ldots \cup U_K = X$;
3) $U_i \cap U_j = \phi, i \neq j, i, j = 1, 2, \ldots, K$.

Multifeature cluster $\mathfrak{R}$ involves three steps. 1) Data preprocessing: the attributes of data objects are usually different, thereby leading to a significant deviation in the iteration of the algorithm. The preprocessing procedure can effectively extract the features and standardize the data value, 2) Multifeature clustering: the data is allocated to several clusters due to specific rules. This procedure may involve the selection of the first cluster center and the number of multifeature clusters, and 3) Analysis of clustering result: the similarity is used to evaluate and analyze the clustering results.

1) Data Preprocessing: The multifeature data sets should be normalized in the range from 0 to 1 to improve the detection accuracy and detection efficiency. As shown in Fig. 2, the testing data set is denoted by X. $\mathrm{Max}(X)$ and $\mathrm{Min}(X)$ are the maximum and minimum feature value of X. Any $x \in X$ can be normalized to a new feature value x' by using (1)

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

2) Distance Measure of Multiple Features: The distance metric function is commonly used to measure the similarity of multifeature data. The shorter the distance, 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 two data objects in X that include M attribute features. The Mahalanobis distance is introduced in this article, and the correlation between data is used in the distance calculation through the covariance matrix, as in (2)

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

where S denotes the covariance matrix. When it is a unit matrix, the distance metric is similar to the Mahalanobis distance. The distance metric function can be used to evaluate the similarity of multifeature data clustering. As each metric is suitable for specific application scene, the clustering algorithm will generate different results in various cases. So thus, the closeness of data objects in multifeature data set X can be evaluated by similarity, namely the $n \times n$ symmetry similarity matrix, as in 3

$$\begin{pmatrix} 0 & d(X_1, X_2) & \dots & d(X_1, X_n) \\ d(X_2, X_1) & 0 & \dots & d(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ d(X_n, X_1) & d(X_n, X_2) & \dots & 0 \end{pmatrix}. \quad (3)$$

3) Evaluation Function of the Multifeature Clustering Effect: There are two ending conditions of the iteration for the multifeature clustering algorithm, namely, reaching the preset iteration times or achieving the best clustering effect. The optimal criteria of clustering are calculated by the evaluation function that will compute the results after each iteration. If it reaches the ending conditions, and the iteration is terminated and will continue until the result is optimally suitable, if otherwise. A general method to evaluate the clustering effect is the squared error criterion.

Definition 2: σ is the sum of square error, as (4)

$$\sigma = \sum_{j=1}^{K} \sum_{X_i \in U_j} \|X_i - u_j\|^2 \quad (4)$$

where u_j is the clustering center of the jth cluster U_j. The smaller value of σ denotes a better clustering result; hence, if σ is the optimal, the clustering is finished.

B. Mathematical Model

Let the data set be $X = \{X_1, X_2, \dots, X_n\}$. Each data object $X_i = \{x_{i1}, x_{i2}, \dots, x_{iM}\}(1 \leq i \leq n)$ is a M-dimensional vector with M attribute features. The kth attribute feature is denoted by $F_k = \{x_{1\,k}, x_{2\,k}, \dots, x_{nk}\}$. $\omega_k(1 \leq k \leq M)$ is the weight of the kth attribute feature.

Definition 3: For a complete graph G, each node is a point $X_i(1 \leq i \leq n)$ in X. The weighted value of the edge e_{ij} between node X_i and X_j can be calculated as (5)

$$W_{e_{ij}} = d(X_i, X_j) = \sum_{k=1}^{M} \sqrt{\omega_k (x_{ik} - x_{jk})^2}. \quad (5)$$

Definition 4: Let $N(X_i) = \{X_i^{(1)}, X_i^{(2)}, \dots, X_i^{(L)}\}$ be the collection of L points with the closest distance to X_i. The security coefficient $s(X_i, L)$ can be calculated by (6)

$$s(X_i, L) = \frac{1}{\sum_{j=1}^{L} d\left(X_i, X_i^{(j)}\right)}, \quad 1 \leq i \leq n. \quad (6)$$

Let the weighted value of node X_i be $W_{X_i} = s(X_i, L)$. K cluster centers $u_1, u_2, \dots, u_K(u_i \in X, i = 1, 2, \dots, K)$ can be selected. Next, each data object X_i in X will be added to the cluster U_i that has the closest distance to the cluster center u_i, $\sum_{i=1}^{n_i} X_i / n_i$ is calculated. Herein, n_i is the number of data objects in the ith cluster. Based on the principle of average value and σ optimization, the cluster centers $u_1, u_2, \dots, u_K$ can be adjusted. Lastly, K clusters $U_1, U_2, \dots, U_K$ will be generated until the cluster center is not changed. With the centroid, radius and the average security coefficient of each cluster, the detection rule $r = \{r_1, r_2, r_3\}$ is generated.

To improve the detection efficiency, reducing the FP rate and enhancing the stability of clustering, three aspects in the selection of ω_k, L and the cluster center $u_1, u_2, \dots, u_K$ are considered.

1) Selection of ω_k: In the selection process of ω_k, the principle of selecting attribution feature preferentially in the clustering procedure should be first considered. The testing data may have redundant or unrelated attributes that lessen the classification accuracy, increase the computation overhead and time. In this point, the evaluation method is given to measure the importance of attribute features. That is, the importance is transformed into a weight of ω_k, so thus those crucial attributes can be gathered in a collection to simplify the attribute features.

Definition 5: Let the kth attribute feature be F_k, and $p(x_{ik})$ the probability of F_k equal to x_{ik}. The information entropy is calculated by (7)

$$H(F_k) = \sum_{x_{ik} \in F_k} p(x_{ik}) \log \frac{1}{p(x_{ik})}. \quad (7)$$

Definition 6: When the value of attribute feature F_k is given, the conditional entropy of F_k can be denoted by (8)

$$H(F_{k'} | F_k) = -\sum_{x \in F_k} p(x) \sum_{y \in F_{k'}} p(y|x) \log p(y|x). \quad (8)$$

Definition 7: Based on the information entropy and conditional entropy, mutual information can be represented by (9)

$$I(F_k, F_{k'}) = H(F_{k'}) - H(F_{k'} | F_k). \quad (9)$$

Definition 8: The correlation degree of F_k can be represented by the average mutual information for F_k and other possible attribute features $F_{k'}(1 \leq k' \leq M)$, as calculated by (10)

$$R(F_k) = \frac{1}{M} \sum_{k'=1}^{M} I(F_k, F_{k'}). \quad (10)$$

The conditional correlation degree of $F_{k'}$ with the condition of F_k can be represented by (11)

$$R(F_{k'} | F_k) = R(F_k) \frac{H(F_{k'} | F_k)}{H(F_{k'})}. \quad (11)$$

Definition 9: The redundancy degree $\mathrm{Red}(F_k, F_{k'})$ is calculated by (12)

$$\mathrm{Red}(F_k, F_{k'}) = R(F_{k'}) - R(F_{k'} | F_k). \quad (12)$$

On the above basis, the importance of attribute feature is defined as (13)

$$I_m(F_k) = R(F_k) - \max\{\mathrm{Red}(F_k, F_{k'})\}. \quad (13)$$

Theorem 1: The weight $\omega_k(1 \leq k \leq M)$ is selected by (14)

$$\omega_k = \frac{I_m(F_k)}{\sum_{k=1}^{M} I_m(F_k)}. \quad (14)$$

It satisfies $0 \leq \omega_k \leq 1$ and $\sum_{k=1}^{M} \omega_k = 1$.

Proof: $\because I_m(F_k) \geq 0, \therefore 0 \leq \omega_k = \frac{I_m(F_k)}{\sum_{k=1}^{M} I_m(F_k)} \leq 1$

If $I = \sum_{k=1}^{M} I_m(F_k)$, then we have (15)

$$\begin{aligned}\sum_{k=1}^{M}\omega_k &= \frac{I_m(F_1)}{I} + \frac{I_m(F_2)}{I} + \cdots + \frac{I_m(F_M)}{I}\\ &= \frac{\sum_{k=1}^{M} I_m(F_k)}{I}\\ &= \frac{I}{I} = 1. \end{aligned} \tag{15}$$

■

2) Selection of L: In existing algorithms, L is usually set as an empirical value. It quickly leads to low detection efficiency and artificial errors, making the result uncertain. It is presented in this section a method to select L value for best detection accuracy and efficiency.

Theorem 2: The security coefficient $s(X_i, l)$ of industrial node X_i is a monotone decreasing sequence. The sum of $s(X_i, l)$ for all data objects when l is equal to different values can be calculated by (16)

$$S(l) = \sum_{i=1}^{n} s(X_i, l), 1 \leq l \leq n-1. \tag{16}$$

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

Proof: Given $s(X_i, l) = \frac{1}{\sum_{j=1}^{l} d(X_i, X_i^{(j)})}$ and $d(X_i, X_i^{(j)}) \geq 0$ in (9), we can deduce as $\sum_{j=1}^{l} d(X_i, X_i^{(j)})$ is also a monotone decreasing sequence of l. Hence, we have $s(X_i, 1) \geq s(X_i, 2) \geq \cdots \geq s(X_i, n-1) \geq 0, i = 1, 2, \ldots, n$, where $s(X_i, l), i = 1, 2, \ldots, n$ is a monotone decreasing sequence of l.

As $S(1) \geq S(2) \geq \cdots \geq S(n-1)$ is established, so thus the proposition of theorem 2 is proven. The increase of l will cause the decrease of $S(l)$. The relative change value of $S(l)$ can be calculated by (17) ■

$$\Delta_l = S(l-1) - S(l), 2 \leq l \leq n-1 \tag{17}$$

where $\Delta_l \geq 0$. If the value of l makes the relative change be the maximum, then it is denoted by L.

3) Selection of Cluster Centers: As shown in Fig. 3, different cluster centers generate different clustering results. In order to improve the stability and accuracy of abnormal data detection, it is crucial to select a finer initial collection of cluster centers that will provide a method to select the initial cluster centers with evenly distributed characteristics. The security threshold value δ is the critical value of the security coefficient. When $s(X_i, l) \geq \delta$ is satisfied, the data object X_i has high security coefficient, and the collection of high-security coefficient points can be denoted by U.

1) The node with the highest security coefficient is selected from X with (10), denoted by u_1;

2) A node with the highest security coefficient that has the longest distance to u_1 is selected from U, denoted by u_2;

3) The distances between each node u_j to the cluster centers $u_1, u_2, \ldots, u_i$ are $d(u_j, u_1), d(u_j, u_2), \ldots, d(u_j, u_i), j \neq 1, 2, \ldots, i$. The node that satisfies the condition in (18) is denoted by u_{i+1}

$$\max\left\{\min\left\{d(u_j, u_1), d(u_j, u_2), \ldots, d(u_j, u_i)\right\}\right\}. \tag{18}$$

4) The step (3) is repeated until all evenly distributed cluster centers $u_1, u_2, \ldots, u_K$ with high-security coefficients are generated.

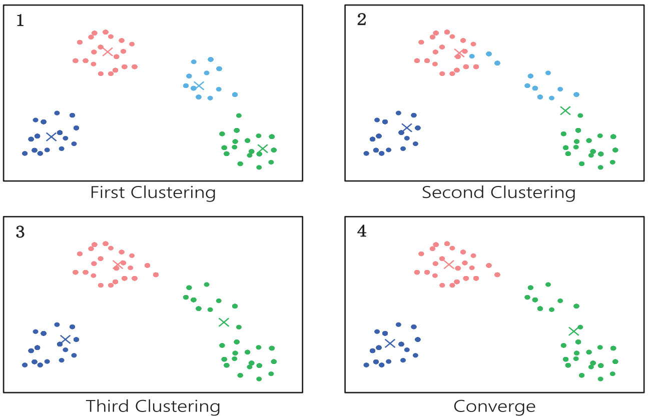


Fig. 3. Cluster centers selection in the multifeature data set.

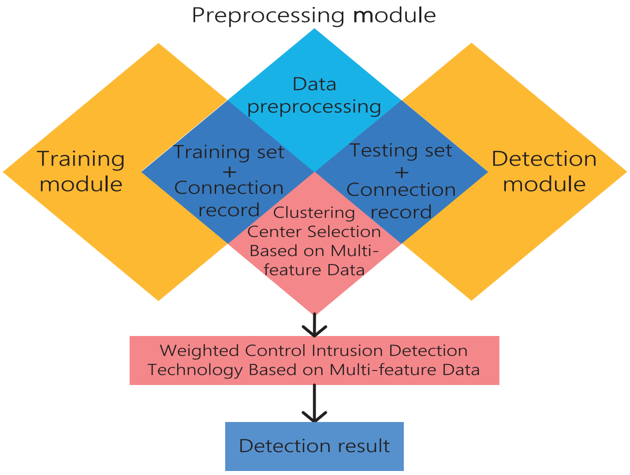

Preprocessing module
Data preprocessing
Training module
Training set + Connection record
Testing set + Connection record
Detection module
Clustering Center Selection Based on Multi-feature Data
Weighted Control Intrusion Detection Technology Based on Multi-feature Data
Detection result

Fig. 4. Chart of multifeature data detection.

C. Proposed Intrusion Detection Technique

The proposed algorithm mainly focuses on the extraction and preprocessing of useful data attribute features from the intrusion data. The results are analyzed based on previous experience and actual situation. The clustering analysis algorithm used in industrial network intrusion detection is to use a specific artificial intelligence algorithm for data clustering analysis. Based on the distribution of data objects in each cluster, the cluster can be marked as normal or abnormal. After that, the cluster center, radius, and the average security coefficient can be used to generate the detection rules. The chart of multifeature data detection is shown in Fig. 4.

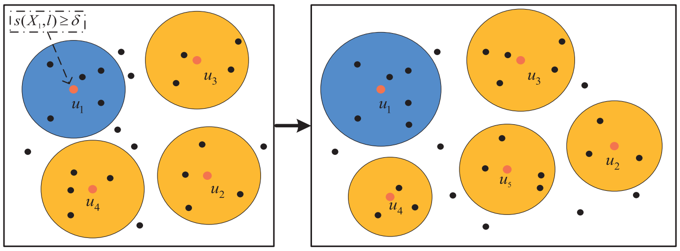


Fig. 5. Selection of cluster center for the multifeature data set.

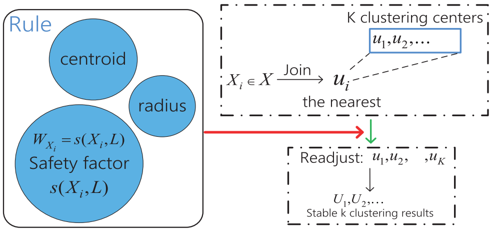


Fig. 6. Weighted clustering model for multifeature data.

Algorithm 1: Selection Algorithm of Cluster Centers.

```
Input:
    Data set $X = \{X_1, X_2, \ldots, X_n\}$;
    Security threshold $\delta$;
    Number of clusters $K$;
Output:
    Initial set of cluster centers, $u_1, u_2, \ldots, u_K$;
1:  Calculating the value of security coefficient $s(X_i, L)$
    for each $X_i$ with (6);
2:  if $s(X_i, L) \geq \delta$ then
3:      Adding $X_i$ into $U$;
4:  end if
5:  Selecting $u_1$ with the highest security coefficient
    from $U$;
6:  Selecting $u_2$ with the longest distance to $u_1$ from $U$;
7:  Calculating $d(u_j, u_1), d(u_j, u_2), \ldots, d(u_j, u_i)$,
    $j \neq 1, 2, \ldots, i$;
8:  $u_{i+1} = \max\{\min\{d(u_j, u_1), d(u_j, u_2), \ldots, d(u_j, u_i)\}\}$;
9:  Repeating 7 and 8 to generate initial set of cluster
    centers $u_1, u_2, \ldots, u_K$;
10: return $u_1, u_2, \ldots, u_K$.
```

Algorithm 2: Weighted Clustering Algorithm for Multifeature Data.

```
Input:
    Data set $X = \{X_1, X_2, \ldots, X_n\}$;
    Security threshold $\delta$;
    Number of clusters $K$;
Output:
    Clustering result, $U_1, U_2, \ldots, U_K$;
1:  Normalize data set with (1);
2:  Calculate attribute feature weighted value $\omega_k$
    with (14);
3:  for $1 \leq i, j \leq n$ do
4:      Calculate $d(X_i, X_j)$ with (5);
5:  end for
6:  Calculate $s(X_i, L)$ with (6), $S(l)$ with (15);
7:  if $\Delta_i = S(i-2) - S(i-1)$ is the maximum then
8:      $L = l$;
9:  end if
10: Repeat 6 and call algorithm 1 to retrieve
    $u_1, u_2, \ldots, u_K$;
11: Add $X_i$ into cluster $U_i$;
12: Calculate $\frac{1}{n_i} \sum_{i=1}^{n_i} (X_i)$;
13: Optimize cluster centers $u_1, u_2, \ldots, u_K$;
14: if Cluster centers are not changed then
15:     return $U_1, U_2, \ldots, U_K$;
16: end if
```

The kernel of the proposed algorithm is illustrated as follows. Let the data set be $X = \{X_1, X_2, \ldots, X_n\}$, and each data object $X_i = \{x_{i1}, x_{i2}, \ldots, x_{iM}\}(1 \leq i \leq n)$ is a M-dimensional vector with M attribute features. With considerations to the principle of important attribute feature F_k and weight value $\omega_k, 1 \leq k \leq M$, the weighted distance and security coefficient of each data object can be calculated. Next, the cluster centers with even distribution are selected. After that, each data object X_i in X will be added to the cluster U_i, which has the closet distance to the cluster center $\mathcal{U}_i$. With the principle of average value and σ optimization, the cluster centers $u_1, u_2, \ldots, u_K$ can be adjusted. Lastly, K clusters $U_1, U_2, \ldots, U_K$ are generated as far as the cluster center is not changed.

To improve the stability and accuracy of the clustering result, it is critical to select a sound cluster center. The selection algorithm of the cluster centers is shown in Algorithm 1.

The selection procedure of cluster centers is depicted in Fig. 5. In the training procedure, the system flow generated by normal behavior can be modeled. In the detection stage, the real data feature in the current industrial network can be compared to the model and generate the degree of deviation. If the degree of deviation exceeds the preset threshold, the current network behavior is marked as intrusion activity. The proposed selection algorithm of cluster center can effectively detect the new unknown attacks despite a high FP rate. Noting that the network intrusion node is different from the normal nodes mathematically and statistically.

Fig. 6 shows the weighted clustering model for multifeature data. The factors such as detection model and the security threshold are critical to the detection of the abnormal behavior in an industrial network. One important characteristic the proposed algorithm should have is to label as normal and abnormal behavior. Unfortunately, it will waste plenty of time to recognize the classification label of massive data. From this, the proposed model can effectively label the normal and abnormal data from the unmarked data that address the issues of low efficiency and accuracy. The pseudocode is described in Algorithm 2.

TABLE I
PARAMETERS IN EXPERIMENTAL ENVIRONMENT

Parameter	***Symbol***
The weight of the k-th attribute feature	ω_k
The distance between X_i and X_j	$d(X_i, X_i)$
The number of nodes closest to X_i	L
The security coefficient of X_i	$s(X_i, L)$
The security threshold	δ
The number of clusters	K
The number of undetected abnormal data	A_1
The number of detected abnormal data	A_2
The number of recognized normal data	N_1
The number of normal data that is marked as abnormal	N_2

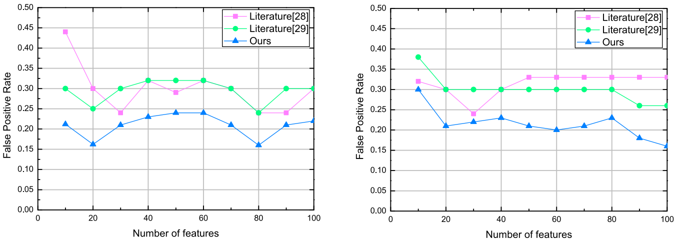


Fig. 7. Evaluation of FP under different thresholds.

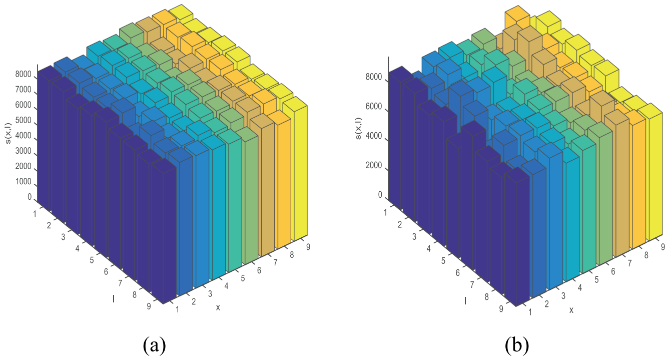


Fig. 8. Impact of security coefficient on the detection rate.

IV. EXPERIMENTS AND ANALYSIS

In this section, evaluations and analysis on the performance of the proposed algorithm are conducted, mainly involving metrics such as security, detection time, and detection accuracy.

A. Experimental Environment

The datasets NSL-KDD and KDDCU'99 are used in experiments by taking into consideration the clustering analysis for different data features [21]–[24]. The original record and label of parameter value are included in the dataset, to reflect whether the record is normal or abnormal. The difference in intrusion attacks, parameters, and data type in the data sets provide evidence to evaluate the reliability of intrusion detection. The parameters of the experiments are found in Table I. The experimental environment includes two parts: communication simulation in an industrial network and abnormal network flow detection. The former is implemented on a host machine with Windows OS, while the latter is realized on a virtual machine with Ubuntu OS. Both are connected via the serial port of the virtual machine.

B. Security

The security of intrusion detection algorithm [25]–[27] can be evaluated by using the true positive rate (TP) and FP rate, where the TP is the ratio of accurately recognized abnormal data objects and the total abnormal data objects, and defined as follows (19). Further, FP is the ratio of the number of normal data objects marked as abnormal and the total normal objects, as follows (20)

$$TP = \frac{A_2}{A_1 + A_2} \tag{19}$$

$$FP = \frac{N_2}{N_1 + N_2} \tag{20}$$

where A_1 is the number of undetected abnormal data, A_2 the number of detected abnormal data, N_1 the number of recognized normal data, and N_2 the number of normal data marked as abnormal.

Based on the classification in different dimensions of data features, this experiment selects five samples from each cluster to construct the training set and utilizes them for classification evaluation, which results are shown in Fig. 7. As an experimental setting, the number of features changes from 0 to 100 and the increment of each step 20, the proposed algorithm is compared to algorithms depicted in [28] and [29] with different security thresholds. From comparison performed, the proposed algorithm has the lowest FP rate. Yet, for the same data set, the proposed algorithm decreases the FP rate by 8.8% when compared to the algorithm in [28], demonstrating that the proposed algorithm has better security than other comparative algorithms.

In the dataset for experimentation, unbalanced data with the significant difference is selected for evaluation. Fig. 8 shows the effect of parameter on classification accuracy before and after multiple features are selected and noted that the change in the parameter value leads to a minimal change in the TP. As a consequence, the proposed algorithm is not sensitive to the parameter, given that parameter optimization is relative to the clustering of data features.

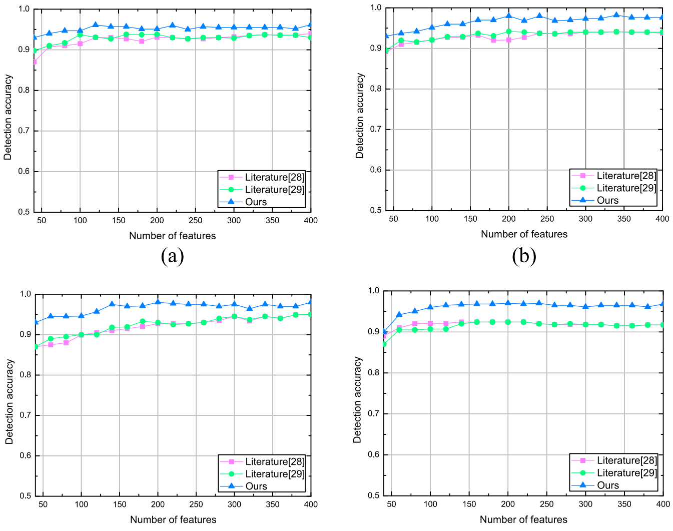
C. Detection Time

In an industrial network environment, the detection time is typically used to evaluate the real-time performance of intrusion detection algorithms. This experiment is conducted utilizing NSL-KDD data set, and six different attack methods are used to assess the detection time.

As shown in Table II, six attack methods are used to evaluate the real-time performance. The testing data is selected from NSL-KDD with several different features, and the algorithms in [28] and [29] are used for comparison, with the results of detection time listed in Table II. From the comparison,

TABLE II
EVALUATION RESULTS OF TEST TIME WITH MULTICLASS NSL-KDD DATASET (UNIT:MS)

Attack Methods	*Training Time*	*Algorithm [28]*	*Algorithm [29]*	*Ours*
Normal	9826	9712	8712	7712
PRB	1686	1567	1367	1055
Probe	1212	1106	1003	966
R2L	2438	2198	2084	1836
DOS	5866	5742	5226	5045
U2R	66	56	42	37
Average(%) Time Saving	-	3.6	5.7	7.8



(a) (b)

(c) (d)

Fig. 9. Evaluation of detection accuracy with different security threshold values.

the proposed algorithm has lower average detection time than algorithms in [28] and [29]. This is due to the multifeature data clustering model can detect the abnormal behavior of high-disguised class accurately and rapidly. Besides, observation of the training data set in real-time indicates that the number of training records for the feature data is too small. Therefore, the proposed model achieves good performance in detection time.

D. Detection Accuracy

Original data samples in industrial networks are usually high-dimensional, and to effectively apply a multifeature data clustering method to extract the abnormal data in the industrial network, high-dimensional data should be transformed into low-dimensional data for detection. This approach can reduce the storage of intrusion detection system yet decrease the computation complexity of model learning and weaken the noise, thereby clarifying the potential structure of data. As observed, the accuracy rate of detecting abnormal behavior in industrial networks is improved. Detection accuracy of each data in the training data set is an important metric to evaluate the performance of the proposed algorithm.

The data in the training data set is marked as normal or abnormal. In addition, the detection accuracy for the normal and abnormal data set is compared by using the securely weighted threshold, as shown in Fig. 9. The evaluation results of detection accuracy when the threshold is 0.2, 0.4, 0.6, and 0.8 are shown in Fig. 9(a)–(d), respectively. We observe that, with the increase of the feature value, the proposed algorithm achieves better detection accuracy than those of the comparative algorithms in [28] and [29] when the threshold is 0.6. For the data set NSL-KDD, the detection can reach to 97.8%. Consequently, the proposed algorithm can address the issues of low detection accuracy by using an empirical value.

V. CONCLUSION

With the rapid advancement of industrial network technology, intrusion attacks may be overlapped and disguised. As known, many intrusion detection algorithms have issues of low accuracy and high FP in detection. To address these issues, this article proposed an industrial network intrusion detection algorithm based on multifeature data clustering optimization model. Such an algorithm established a data clustering optimization model for the intrusion attacks in industrial networks, followed to the presentation of a cluster center selection algorithm for multifeature data and the intrusion detection algorithm. In the proposed algorithm, despite the trace procedure increases the training center of the model, the detection accuracy for the attacks with high overlapping and disguise degree dramatically improved. Experimental results showed that the proposed algorithm achieved higher detection accuracy and lower FP than other existing intrusion detection algorithms.

Several research institutes and organizations focus on the detection of abnormal behavior in the industrial network to resist attacks such as virus and Trojan, as well as to the evaluation and prediction of potential intrusion attacks in advance. The real-time and accurate detection of prospective network intrusion attacks is critical to ensure the security of future industrial networks. As intrusion detection systems are under operation, information such as system flow, behavior feature of each node, and data set of historical nodes can be used to determine whether nodes are illegal or systems under threatening and such issues are focused as future directions of this research.

REFERENCES

[1] R. Langner, “Stuxnet: Dissecting a cyberwarfare weapon,” *IEEE Secur. Privacy*, vol. 9, no. 3, pp. 49–51, May–Jun. 2011.

[2] Q. Tao *et al.*, “A cloud-based experimental platform for networked industrial control systems,” *Int. J. Modeling, Simul. Sci. Comput.*, vol. 9, no. 4, 2018, Art. no. 1850024.

[3] G. Liang *et al.*, "A cloud-based experimental platform for networked industrial control systems," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3317–3318, Dec. 2017.

[4] M. Savita and M. Patil, "A brief study of WannaCry threat: Ransomware attack 2017," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 5, 2017.

[5] W. Liang *et al.*, "A double PUF-based RFID identity authentication protocol in service-centric internet of things environments," *Inf. Sci.*, vol. 503, pp. 129–147, 2019.

[6] W. Liang and M. Tang, "A secure fabric blockchain-based data transmission technique for industrial internet-of-things," *IEEE Trans. Ind. Inform.*, vol. 15, no. 6, pp. 3582–3592, Jun. 2019.

[7] IDC, "Executive summary: Data growth, business opportunities, and the IT imperatives," The digital universe of opportunities: Rich data and the increasing value of the Internet of Things, Apr. 2014. [Online]. Available: https://www.emc.com/leadership/digital-universe/2014iview/executive-sum% mary.htm

[8] C. Jiang, H. Zhang, Y. Ren, Z. Han, K. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[9] R. Cunningham and R. Lippmann, "Detecting computer attackers: Recognizing patterns of malicious stealthy behavior," Purdue Univ.: West Lafaayette Cerias, Nov. 2000. [Online]. Available: http://www.cerias, purdue.edu/secsem/abstracts001.php

[10] R. P. Lippmann and R. K. Cunningham, "Improving intrusion detection performance using keyword selection and neural networks," *Comput. Netw.*, vol. 34, no. 4, pp. 597–603, 2000.

[11] J. Bigham *et al.*, "Test data for anomaly detection in the electricity infrastructure," *Int. J. Crit. Infrastructures*, vol. 2, no. 4, pp. 396–411, 2006.

[12] N. Gao *et al.*, "A lightweight intrusion detection model based on autoencoder network with feature reduction," *Acta Electronica Sinica*, vol. 45, no. 3, pp. 730–739, 2017.

[13] X. Deng, P. Jiang, X. Peng, and C. Mi, "An intelligent outlier detection method with one class support tucker machine and genetic algorithm towards big sensor data in internet of things," *IEEE Trans. Ind. Electron.*, vol. 66, no. 6, pp. 4672–4683, Jun. 2019.

[14] J. Yu, Y. Hou, and V. Li, "Online false data injection attack detection with wavelet transform and deep neural networks," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3271–3280, Jul. 2018.

[15] S. Ruoti *et al.*, "Intrusion detection with unsupervised heterogeneous ensembles using cluster-based normalization," in *Proc. IEEE Int. Conf. Web Services*, Honolulu, HI, USA, pp. 862–865, 2017.

[16] S. Naseer *et al.*, "Enhanced network anomaly detection based on deep neural networks," *IEEE Access*, vol. 6, pp. 48 231–48 246, 2018.

[17] J. Gao *et al.*, "On threshold-free error detection for industrial wireless sensor networks," *IEEE Trans. Ind. Informat.*, vol. 14, no. 5, pp. 2199–2209, May 2018.

[18] F. Sun, G. Huang, Q. M. Jonathan Wu, S. Song, and D. C. Wunsch II, "Efficient and rapid machine learning algorithms for big data and dynamic varying systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 10, pp. 2625–2626, Oct. 2017.

[19] S. Zhang, Y. Qi, F. Jiang, X. Lan, P. C. Yuen, and H. Zhou, "Point-to-set distance metric learning on deep representations for visual tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 187–198, Jan. 2018.

[20] *Cluster analysis*, Merriam-Webster Online Dictionary, 2018. [Online]. Available: http://www.merriamwebster-online.com

[21] *The NSL-KDD dataset*, NSL-KDD, 2018. [Online]. Available: http://nsl.cs.unb.ca/NSL-KDD/

[22] F. Jiang *et al.*, "Deep learning based multi-channel intelligent attack detection for data security," *IEEE Trans. Sustain. Comput.*, to be published.

[23] S. Raman *et al.*, "An intrusion detection system using network traffic profiling and online sequential extreme learning machine," *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8609–8624, 2018.

[24] H. P. Hamed, R. Javidan, R. Khayami, A. Dehghantanha, and K. R. Choo, "A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in IoT backbone networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 7, no. 2, pp. 314–323, Apr./Jun. 2019.

[25] J. Wu, Y. Zhang, and W. Lin, "Good practices for learning to recognize actions using FV and VLAD," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2978–2990, Dec. 2016.

[26] A. Liu, Y. Su, P. Jia, Z. Gao, T. Hao, and Z. Yang, "Multiple/single-view human action recognition via part-induced multitask structural learning," *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1194–1208, Jun. 2015.

[27] E. Bertino *et al.*, "Internet of Things (IoT): Smart and secure service delivery," *IEEE Trans. Internet Technol.*, vol. 16, no. 4, pp. 1–7, 2016.

[28] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Trans. Cybern.*, vol. 18, no. 2, pp. 1194–1208, Jun. 2016.

[29] N. Shone *et al.*, "A deep learning approach to network intrusion detection," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 1, pp. 41–50, Feb. 2018.