2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering

Research on Industrial Control Anomaly Detection Based on FCM and SVM

Wenli Shang
Shenyang Institute of Automation Chinese Academy of Sciences
University of Chinese Academy of Sciences
Shenyang,China
shangwl@sia.cn

Junrong Cui
Shenyang Institute of Automation Chinese Academy of Sciences School of Automation and Electrical Engineering, Shenyang Ligong University
Shenyang,China
cuijunrong@sia.cn

Chunhe Song
Shenyang Institute of Automation Chinese Academy of Sciences
University of Chinese Academy of Sciences
Shenyang,China
songchunhe@sia.cn

Jianming Zhao
Shenyang Institute of Automation Chinese Academy of Sciences
University of Chinese Academy of Sciences
Shenyang,China
zhaojianming@sia.cn

Peng Zeng
Shenyang Institute of Automation Chinese Academy of Sciences
University of Chinese Academy of Sciences
Shenyang,China
zp@sia.cn

***Abstract*—In order to solve the problem of virus and Trojan attacking the application layer network protocol of industrial control system, the rule of Modbus/TCP communication protocol is analyzed. An intrusion detection method based on clustering and support vector machine is proposed. The method combines unsupervised fuzzy C-means clustering (FCM) with supervised support vector (SVM) machine to calculate the distance between industrial control network communication data and cluster center. Partial data satisfying the threshold condition is further classified by support vector machine. Experimental results show that compared with the traditional intrusion detection method, this method can effectively reduce the training time and improve the classification accuracy without needing to know the class label in advance.**

Keywords—industrial control system, Modbus communication protocol, intrusion detection, fuzzy C-means clustering, supervised support vector (SVM)

I. Introduction

The traditional industrial control system is usually based on the factory area, which is independent of each other and has no physical connection with the outside world. But with the rapid development of the industry information and network technology, industrial control systems are increasingly using general-purpose hardware and software. The openness of industrial control systems is increasing. System security vulnerabilities and defects are easily exploited by viruses [1]. However, the industrial control system is also used in the national power, transportation, oil, heating, pharmaceutical and other large-scale manufacturing industry, once the attack will bring huge losses, so the need for effective ways to ensure industrial control network security[2].

There are many ways to protect the network security of the industrial control system. The most common way is to use firewall, log processing and other linkage. However, the firewall is based on third-party routing access control. It cannot detect attacks from within the system and can only act as a filter, which cannot effectively reduce the security risk of the system [3]. As an active defense technology, anomaly detection can detect the external attack and detect the internal attack of the system. It can integrate the protection, detection and response effectively, and provide more reliable guarantee for the safety of industrial control network[4,5].

Abnormal detection technology can be effectively applied in the industrial control system, domestic and foreign research scholars and experts also carried out a lot of research. Document [6] aiming at the problem of high dimensionality and low detection rate of intrusion detection, a method of dimensionality reduction based on data characteristics and a neural network method based on particle swarm optimization (PSO) are proposed respectively. However, when dealing with the characteristics of associated data, the effect is poor, and the feasibility of this aspect needs to be further explored. According to the problem of less intrusive samples and higher randomness in document [7], a model of intrusion detection algorithm based on support vector machines (SVM) is proposed. Although the detection rate has been improved, but the detection time is longer, it is not suitable for industrial control system environment. Document [8] aiming at the problem of fixed offset constant and detection threshold in intrusion detection, an improved CUSUM intrusion detection model is proposed. This model can meet the real-time requirement of industrial control system, but the stability is not high.

In view of the advantages and disadvantages of the above methods, this paper analyze the rule of Modbus/TCP communication protocol and studies the intrusion detection algorithm of industrial control network based on fuzzy C-Means (FCM) data preprocessing and SVM classification. This strategy combines unsupervised FCM and supervised SVM to realize the semi-supervised machine learning of industrial anomaly detection. This paper obtain the clustering center by fuzzy C-means clustering, calculate the distance between the communication data and clustering center. Partial data

National Natural Science Foundation of China "Research on anomaly detection and security awareness method for industrial communication behaviors"（61773368）and Foundation（6140242010116Zk63001）.

2324-9013/18/31.00 ©2018 IEEE
DOI 10.1109/TrustCom/BigDataSE.2018.00042

satisfying the threshold condition are further classified by SVM. The abnormal detection model of industrial control system is established, and the abnormal intrusion is detected in time, so as to realize the protection of industrial control system.

II. Modbus / TCP Protocol Analysis and Data Preprocessing

A. Modbus / TCP Protocol Analysis

Modbus is a strict application layer message transmission protocol, which communicates with other devices through the network (Ethernet) and obeys the master/slave mode. It is the world's first industrial bus protocol[9].

Modbus communication protocol is mainly used in serial link, and Modbus/TCP is generally used in Ethernet link [10]. Modbus/TCP is embedded in industrial Ethernet TCP/IP frames, in essence, Modbus/TCP messages are Modbus communications encapsulated in an Ethernet wrapper, and the message package is shown in figure 1.

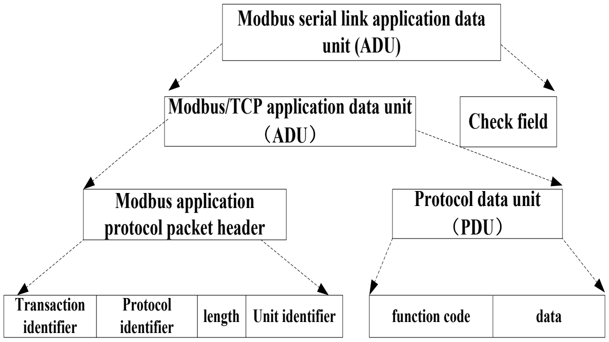
Modbus serial link application data unit (ADU)
Modbus/TCP application data unit (ADU)
Check field
Modbus application protocol packet header
Protocol data unit (PDU)
Transaction identifier
Protocol identifier
length
Unit identifier
function code
data

Fig. 1. Modbus packet encapsulation format

This paper focuses on Modbus/TCP application data unit (ADU). An application data unit mainly consists of Modbus application protocol (MBAP) message header and protocol data unit(PDU). MBAP is used primarily for the identification of application data units for Modbus/TCP. PDU includes function code and data, through different function code to achieve the operation of the server, data is used to store the data transmitted, and part of the data can also be used for functional code parameters [11,12].

B. Data Feature Extraction and Normalization

First, Modbus/TCP traffic data packets are crawled with wireshark. Then, the extracted data is pretreated. For each Modbus TCP / IP protocol has a variety of attributes, from which to extract the most characteristics of the data characteristics, and then remove the noise and redundant information, the data normalization, the establishment of intrusion detection model. Industrial data feature extraction is shown in table 1.

The extracted packets are sorted in chronological order and randomly divided into different sequences. To ensure the representation of the sample.

TABLE I. Industrial Special Table

sequence number	Industrial characteristics	attribute
1	source address	choose
2	destination address	choose
3	TCP packet header length	choose
4	function code	choose
5	source port number	choose
6	destination port number	choose
7	data length	choose
8	transaction identifier	choose
9	protocol identifier	choose
10	unit identifier	choose

The redundant data is removed, the data is normalized, and the data of the different units and dimensions are grouped into the unified form by the minimum and maximum standardization method.

$$v' = \frac{v - \min}{\max - \min}\left(\max' - \min'\right) + \min' \qquad (1)$$

Max and min represent the maximum and minimum values in the data set; max 'and min' represent the interval (min ', max') of the new space; v is the input vector; v' is the output vector.

C. FCM Data Preprocessing

The fuzzy C means clustering algorithm is to obtain the membership degree of each element by optimizing the objective function, so as to obtain a clustering algorithm which belongs to a certain class degree of each element [13]. The core idea of FCM algorithm is to assume that each data vector belongs to only one cluster, and the n data set vectors are divided into c clusters, and the clustering center of each cluster is obtained, so that the objective function is minimized [14,15].

Set X = {x$_1$, x$_2$,... x$_n$}, xi represents each data vector, and each data vector is k dimension, and the clustering center, membership degree and objective function are calculated by the following formula.

Clustering center:

$$v_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} x_i}{\sum_{i=1}^{n} u_{ij}^{m}} \qquad (2)$$

Membership degree：

$$\mathbf{u}_{ij} = \begin{cases} \left(\sum_{k=1}^{c} \frac{\left\| \mathbf{x}_i - \mathbf{v}_j \right\|^{\frac{2}{m-1}}}{\left\| \mathbf{x}_i - \mathbf{v}_k \right\|^{\frac{2}{m-1}}} \right)^{-1} & \| \mathbf{x}_i - \mathbf{v}_k \| \neq 0 \\ 1 & \| \mathbf{x}_i - \mathbf{v}_k \| = 0 \text{ 且 } k = j \\ 0 & \| \mathbf{x}_i - \mathbf{v}_k \| = 0 \text{ 且 } k \neq j \end{cases} \qquad (3)$$

objective function：

$$J = \sum_{i=1}^{n}\sum_{j=1}^{c}\left(u_{ij}\right)^{m}\left\|x_i - v_j\right\| \qquad (4)$$

First, FCM clustering is used, and then the data is normalized after the industrial control network is clustered. The clustering center of the cluster is calculated and the data vector near the cluster center is considered to be classified correctly. Therefore, calculate the distance between each data vector and the cluster center, given the threshold λ, get training set A, the specific steps are as follows:

Step 1:FCM clustering to get the clustering center O of each cluster, all the normal clustering centers are labeled O+, all the abnormal clustering centers are labeled O-, the normal set is marked A+, and the abnormal set is marked as A-.

Step 2: for each data vector xi, and calculate the distance from the cluster center, if meet d（xi ,O+）<λ mark the data vector xi ∈A+, or mark xi ∈A-.

Step 3: Repeat step 2 until each data vector in data set X is marked into the collection.

Step 4: Training set A=A+ ∪ A-.

After the data pretreatment of FCM, the industrial data is divided into three parts. The first part is the data near the normal clustering center. The second part is the data near the abnormal clustering center. The third part is the data close to the normal and abnormal boundaries. For the first and second part of the data, it is easy to determine the normal and abnormal, you can detect attacks in time. For the third part of the data that is close to the normal data is also close to the abnormal data, it is not easy to detect. The support vector machine is used to classify the data, but some data is classified, and the training time is greatly reduced.

III. The Application of Support Vector Machine in Industrial Control Anomaly Detection

A. Support vector machine

The main idea of SVM is to map the input vector from the low dimensional space to the high dimension space, and then construct the optimal classification surface in the high dimensional space[16]. The main idea is shown below.

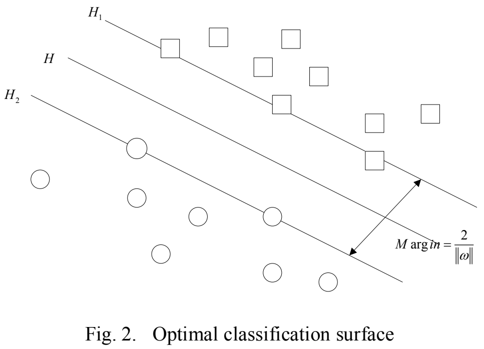

Fig. 2. Optimal classification surface

Square and circle represent two types of data, H is the classification line, H_1 and H_2 are parallel to H, and the distance between H_1 and H_2 is the sorting interval. The purpose of support vector machine classification is to ensure that the classification is accurate on the basis of the classification of the largest interval. H is defined as the optimal classification line, which is extended from two-dimensional space to high dimensional space, and the optimal classification line is extended to the optimal classification surface [17,18]. The optimal classification surface problem can be transformed into the following constraint problem:

$$\min \frac{1}{2}\|\omega\|^2 + c\sum_i^l \zeta_i \qquad (5)$$

$$s.t.\begin{cases} y_i\left(wx_i + b\right) \geq 1 - \zeta_i \\ \zeta_i \geq 0 \end{cases} (i = 1,2,\ldots,l) \qquad (6)$$

The kernel function realizes that the data in the low dimensional space can be separated linearly in the high dimensional space, and the radial basis function (Gauss kernel function, RBF) is selected.

$$k\left(x, x_i\right) = \exp\left(-g\left\|x - x_i\right\|^2\right) \qquad (7)$$

Largrange factor α_i is introduced to obtain the Largrange function of the above problem:

$$\phi\left(\omega, b, \alpha_i\right) = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{l}\alpha_i\left[y_i\left(wx_i + b\right) - 1\right] \qquad (8)$$

The dual function of upper form is obtained, and the partial derivative of ω and b is derived. The optimal solution of ω^* and b^* is obtained by using the dual principle, and the optimal decision function is received.

$$f(x) = sign\left(\sum_{i=1}^{N}\alpha_i^* y_j K\left(x \cdot x_i\right) + b^*\right) \qquad (9)$$

B. Data Feature Extraction and Normalization The Anomaly Detection Model of Support Vector Machine

Using support vector machines to build anomaly detection model is actually a two classification problem, which distinguish normal data from abnormal data. The anomaly detection model mainly consists of two parts: training stage and detection stage. In the training phase, the data near the cluster center is used as the training data set in the FCM data preprocessing.

In the detection phase, calculate the distance D1and D2 between the data vector in the B={B1， B2， …,Bi} and Clustering center O+ and O- in data preprocessing stage. Given the threshold ε, if $|D1 - D2| \phi \varepsilon$, then the data vector is close to a cluster center. If D1> D2, then marked as abnormal. If D1 <D2, marked as normal. Without the support vector machine algorithm re-classification can be judged, thus greatly reducing

the training time. If $|D1-D2| \pi \varepsilon$, indicating that the data vector near the middle of the classification, you need to re-classification through the support vector machine to determine. The block diagram of the SVM intrusion detection model is shown in figure 3:

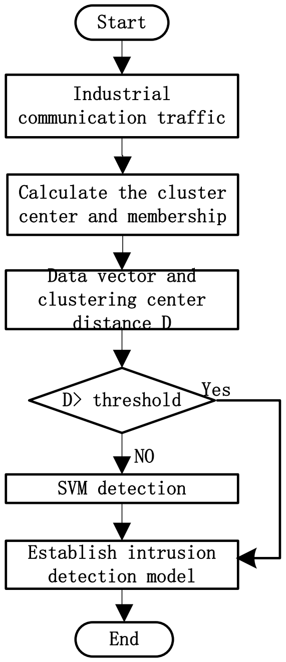
Start
Industrial communication traffic
Calculate the cluster center and membership
Data vector and clustering center distance D
D> threshold
Yes
NO
SVM detection
Establish intrusion detection model
End

Fig. 3. SVM intrusion detection program block diagram

The concrete steps are as follows:

Step 1: calculate the distance between each data vector Bi and the cluster center D1=distance (Bi, O+) and D2=distance (Bi, O-).

Step 2: If $|D1 - D2| \phi \varepsilon$, and D1>D2, mark the exception, D1<D2, mark normal. If $|D1-D2| \pi \varepsilon$, reclassified by SVM.

Step 3: repeat the above steps until B=Φ.

IV. Simulation Experiment Analysis

In order to validate the research of industrial anomaly detection based on FCM data preprocessing and SVM classification, a simulation experiment platform is set up to simulate the water level control in the actual production. The controller uses the Schneider M340PLC. The collection and control of the liquid level data are realized by PLC programming, and the network communication between the host computer and the host computer is realized through Modbus/TCP. When the water tank is running normally, the Modbus/TCP data in the industrial control network is captured, and the data is pretreated. Training data contained 350 communication data samples, the test data contained 150 data samples. The FCM data preprocessing and SVM classification are used to detect the industrial anomaly detection model. The experimental data are shown in Figure 4 and Figure5.

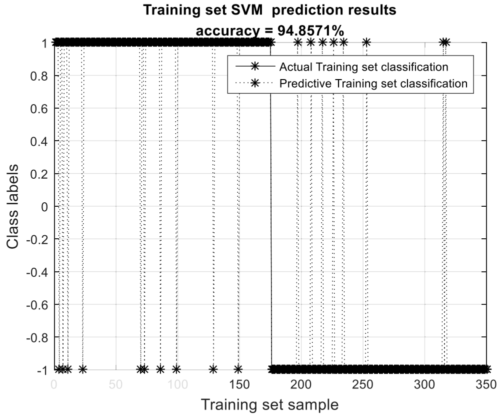

Fig. 4. training set results of SVM training data

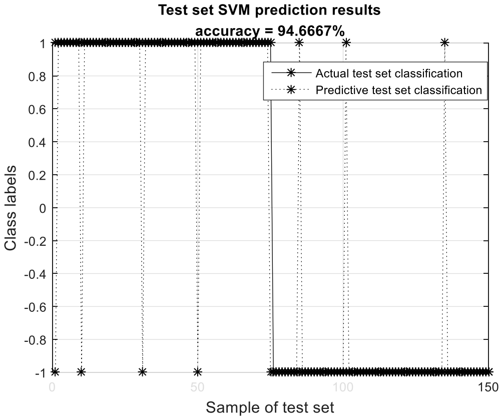

Fig. 5. Test results of SVM test data

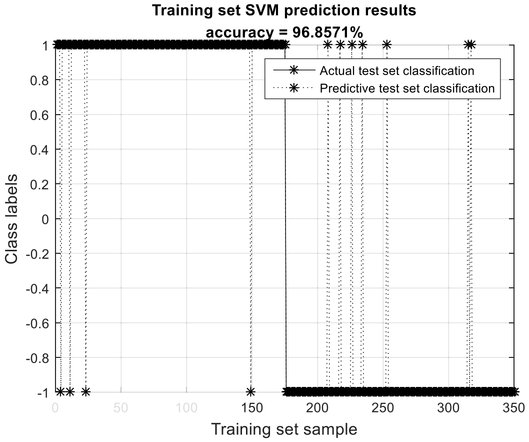

Fig. 6. FCM data preprocessing and SVM classification

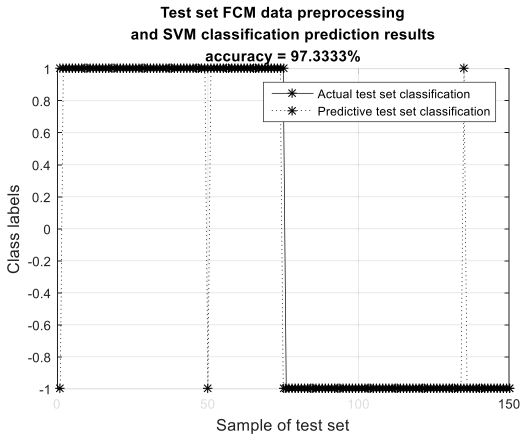

Fig. 7. SVM intrusion detection program block diagram

The simulation results show that the detection accuracy of the test set is 94.8571% by the SVM Classification, and the model time is 0.76647s. The detection accuracy of the test set based on FCM data preprocessing and SVM classification is 96.8571%, and the optimized test model time is 0.56155s. The accuracy of the training set is 97.3333% and the time of optimization test is 0.49516s by fuzzy C-means clustering algorithm. The accuracy of the training set based on FCM data preprocessing and SVM classification is 94.6667%, and the time of optimization test model is 0.68169s.The results of each experiment are shown in table 2.

TABLE II. EXPERIMENTAL RESULT

Algorithm	Training Data		Test Data	
	accuracy	*time/s*	*accuracy %*	*time/s*
SVM	94.67	0.6817	94.85	0.7665
FCM-SVM	97.33	0.4952	96.85	0.5616

For the analysis of Table 2, the training accuracy and test precision of the support vector machine algorithm are relatively low and have a long time to use in the detection model without FCM data preprocessing. The accuracy of model detection based on fuzzy c-means clustering data preprocessing and support vector machine classification is improved. And the time used is much lower than the detection accuracy of the support vector machine classification, which means that the detection precision is improved and the detection time is reduced.

V. CONCLUSION

In this paper, a new data preprocessing method based on FCM is proposed, and a model of FCM data preprocessing and SVM classification is established. This method can effectively classify the industrial data without the need to know the label in advance. Breaking the traditional need to know the limitations of category labels. In addition, this method can effectively apply the high efficiency of fuzzy C-means clustering algorithm and the high precision of support vector machine, and combine the clustering and support vector machine to achieve better application value in industrial intrusion detection.

REFERENCES

[1] Keke Gai, Meikang Qiu, Xiaotong Sun. A survey on FinTech. Journal of Network and Computer Applications, 2017, 103:262-273..

[2] Sadeghi AR, Wachsmann C, Waidner M, “Security and Privacy Challenges in Industrial Internet of Things”. In: 52nd ACM/EDAC/IEEE, Design Automation Conference (DAC), America, 2015, pp.1-6.

[3] Rathnayake K K C D, De Silva N H, Amarasinghe D J. “Future Firewall Security Enhancements”, Imperial Journal of Interdisciplinary Research ,2016.

[4] Keke Gai, Meikang Qiu. Blend Arithmetic Operations on Tensor-based Fully Homomorphic Encryption Over Real Numbers. IEEE Transactions on Industrial Informatics, 2017, PP(99):1-1.

[5] Kenkre, Poonam Sinai, Anusha Pai, and Louella Colaco. “Real time intrusion detection and prevention system.” 3rd ed Springer, Cham, 2015.

[6] Kuang F, Zhang S, Jin Z, “A Novel SVM by Combining Kernel Principal Component Analysis and Improved Chaotic Particle Swarm Optimization for Intrusion Detection”, Soft Computing ,2015, vol.5, pp. 1187-1199.

[7] Feng W, Zhang Q, Hu G, “Mining Network Data for Intrusion Detection through Combining SVM with Ant Colony Networks” Future Generation Computer Systems, 2014, vol.37, pp.127-140.

[8] Li S H, Liu J W, “An Improved CUSUM Intrusion Detection Method for Industrial Control System”, Application of Electronic Technology 2015, vol.09, pp.118-121.

[9] Keke Gai, Meikang Qiu, et al. Privacy-preserving multi-channel communication in Edge-of-Things. Future Generation Computer Systems, 2018, pp.190-200.

[10] CHEN Z H, “Application of Modbus TCP / IP Communication in Industrial Production”, Science and Technology Innovation and Application , 2016.

[11] Morris T H, Jones B A, Vaughn R B, “ Deterministic Intrusion Detection Rules for MODBUS Protocols”. 46th Hawaii International Conference, System Sciences ,2013, pp.1773-1781.

[12] Sethuraman C, Jood P, Srinivas K, “Remote Monitoring Energy Management System Using LonWorks and Modbus Communication Protocol” . Energy , 2015.

[13] Shang W, Cui J, Cong C, “Industrial Control Intrusion Detection Based on Semi-supervised Clustering Strategy”, Information and Control, 2017, vol 46 (04): pp.462-468.

[14] John V, Mita S, Liu Z, “ Pedestrian Detection in Thermal Images Using Adaptive Fuzzy C-Means Clustering and Convolutional Neural Networks”. 14th IAPR International Conference, Machine Vision Applications ,2015,pp.246-249.

[15] Kalti K, Mahjoub M A, “Image Segmentation by Gaussian Mixture Models and Modified FCM Algorithm”, Int. Arab J. Inf. Technol ,2014 , vol.1, pp.11-18.

[16] Kuang F, Zhang S, Jin Z,” A Novel SVM by Combining Kernel Principal Component Analysis and Improved Chaotic Particle Swarm Optimization for Intrusion Detection”, Soft Computing , 2015, pp. 1187-1199.

[17] .Ding S, Y u J, Qi B, “An Overview on Twin Support Vector Machines”, Artificial Intelligence Review , 2014, pp. 1-8.

[18] Tang Y,” Deep Learning Using Linear Support Vector Machines”, ArXiv preprint arXiv , 2013, pp.1306.0239.