ATLANTIS PRESS
Advances in Computer Science Research, volume 87
3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019)

Research on Baseline Technology of Industrial Control Network Security based on Semi-supervised Learning

Yixiang Jiang[1, a], Chengting Zhang[2, b], Wenlong Jin [3, c]

[1]Bachelor of Ningbo University of Technology, Network Engineer, China Tobacco Zhejiang Industrial CO., LTD, NingBo 315000, China

[2]Cloud computing and smart factory, Zhejiang Tobacco Industry Co., Ltd., Ningbo 315000, China

[3]Computer application and enterprise informatization, Zhejiang Tobacco Industry Co., Ltd., Ningbo 315000, China

[a]jiangyxlunwen@sina.com, [b]titanbyron@126.com, [c]jinwenlong@zjtobacco.com

Abstract. With the rapid development of industrial control network, performance management and risk prevention based on network traffic data, especially abnormal traffic detection, have gradually attracted people's attention. However, the traditional flow detection method based on fixed baseline cannot adapt to the growing data and increasingly complex data types. It leads to inaccurate test results and false alarms, and also consumes a lot of manpower and resources. In this paper, a semi-supervised learning method is proposed to realize the self-construction of baseline and the automatic detection of abnormal index data.

Keywords: Semi-supervised Learning Method; Industrial control system; Baseline.

1. Introduction

At present, the importance of industrial control network has become increasingly prominent. With the development of computer and network technology, open industrial communication protocol, network facilities and general hardware and software are widely used in industrial control. Even with data exchange with the Internet and enterprise management information systems, attacks on industrial control networks are increasing rapidly. The network threat in the field of industrial control has greatly endangered the normal operation of industrial control, and the vulnerability of industrial control systems has gradually emerged. The outbreak of "Shenzhen" virus in 2010, the "flame" supervirus in 2012, and the Havex virus specially for industrial control system in 2014 have brought huge losses to users. It also threatens national security. Analysis of the following 2015.The impact of Ukraine's strong attacks that year showed that the cost of the attacks was declining and the impact of the attacks was worsening. In May 2017, Wanna Cry blackmail software became popular all over the world, affecting thousands of companies and public organizations in nearly 100 countries. Industrial control systems are designed to perform various real-time control functions. It will undoubtedly bring huge security risks and hidden dangers to the critical systems and infrastructure under their control. In order to avoid industrial security incidents, it is very important to effectively detect and prevent network attacks. [1]

In this regard, a dynamic baseline prediction model based on semi-supervised learning is designed. Specifically, this paper uses principal component analysis method to construct comprehensive indicators for preprocessing raw data, so as to improve interpretability. A semi-supervised classification model is constructed to realize the prediction of dynamic baseline values and to enable real-time warning when traffic is abnormal.

2. Research Background

2.1 Basic Principles of Safety Baseline

Safety baseline refers to the concept of "baseline". It is a basic reference in measurement, calculation or positioning. For information security, the security baseline can be regarded as the minimum set that can make the system run and the minimum performance and functional

requirements. The network security baseline is a set of minimum communications within a network to achieve the business objectives of industrial control systems. It is recorded as C. [2] Each element CI in set C is a communication. Each communication includes the following elements:

(1) Communication device pair (Es, Er): Each communication has and consists of only two devices, each of which is represented by its IP address. Es is the active party, using random ports to communicate. Er is the passive side, using fixed ports to provide services.

(2) Each E contains an IP address and a communication port P. If Es is a random port in practice, then P is an optional range of random ports, which is recorded as P < [IV, FV].

(3) Communication protocol CP: Communication protocol is the communication protocol needed to accomplish the goal of industrial control in the running state of the system. According to the above definition, OSIRM model based on TCP/IP protocol family, once given the networking mode of control network and application layer control protocol, self-organizing the framework of minimum set of CP, and combined with traffic analysis, can finally determine the minimum set of CP. A communication can be expressed as Ci ((Es (IP, P), Er (IP, P)), CP).

2.2 Basic Principles of Semi-supervised Learning

The model is trained with semi-supervised method by acquiring the data characteristics of network traffic. This method combines Rocchio and LIBLINEAR technologies. Rocchio algorithm is a related feedback algorithm, which appeared in the 1970s and was widely disseminated.

The algorithm represents each data class in the training set as a prototype vector. Each data D is expressed as vector d, so that D is the data of the whole training set, and Cj is the training set set set set in class cj. Using Rocchio method to construct prototype vector CJ for each class CJ to construct classifier

$$\vec{c}_j = \alpha \frac{1}{|C_j|} \sum_{d \in C_j} \frac{\vec{d}}{\|\vec{d}\|} - \beta \frac{1}{|D - C_j|} \sum_{d \in D - C_j} \frac{\vec{d}}{\|\vec{d}\|}$$

Among them, alpha and beta are the influencing parameters to adjust the correlation and irrelevance of training samples. In classification, for each test set network traffic data td, cosine similarity method is used to calculate the similarity between TD and each prototype vector. Which type of prototype vector is more similar to td, and the type of prototype vector is assigned to td. LIBLINEAR is designed and developed by Dr. LIN CJ for linear classification problems. When using LIBLINEAR, it is easy to process millions to tens of millions of data, because LIBLIN-EAR itself aims to solve the problem of large sample training. The idea of model training is to use Rocchio technology to select reliable normal network data from a large number of unlabeled network data, and then use LIBLINEAR technology to train the model. The algorithm is shown in Figure 1. [3]

Rocchio technology can be used to extract reliable negative sample data (expressed as RN) from unmarked network data (expressed as U). For example, the positive data record is P, and the pseudocode of the algorithm is shown in Formula 2.

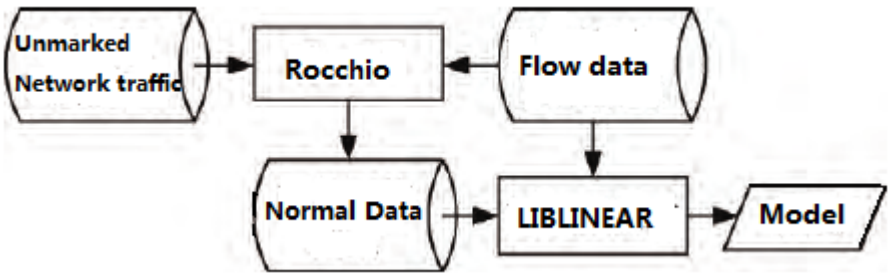

Fig.1 Algorithm processing flow

1.Unmarked network traffic data U is assigned to negative class and positive sample network traffic data P is assigned to positive class.

2 .$\vec{c} \leftarrow \alpha \frac{1}{|P|}\sum_{d \in P}\frac{a}{\|\vec{d}\|} - \beta \frac{1}{|U|}\sum_{d \in U}\frac{a}{\|\vec{d}\|}$

3 .$\vec{c} \leftarrow \alpha \frac{1}{|U|}\sum_{d \in U}\frac{\vec{d}}{\|\vec{d}\|} - \beta \frac{1}{|P|}\sum_{d \in P}\frac{\vec{d}}{\|\vec{d}\|}$

4. d do for each network traffic in U

5.if sim($\vec{c}$,$\vec{d}$)< = sim($\vec{c}$,$\vec{d}$)then

6.RN←RN$\cup$ {d}

In the sample-based active learning method, the negative sample data in unlabeled data set U usually contains many types. In vector space, it occupies a large area. The positive sample data usually belongs to the same type, and the coverage area is much smaller, as shown in Figure 2. Rocchio is a linear classifier. Assuming that there is a decision surface S that can distinguish positive samples from negative samples, the positive prototype vector is closer to the decision surface S than the negative prototype vector because of the vector superposition principle in Rocchio, and the negative sample data identified in this way has high purity.

Rocchio learning model is usually weaker than LIBLINEAR, and noise has a greater impact on LIBLINEAR technology. In order to better classify, this paper proposes a combination of Rocchio and LIBLINEAR. After extracting RN from U with Rocchio, LIBLIN-EAR is run with P and RN, and finally a better classification model is generated.

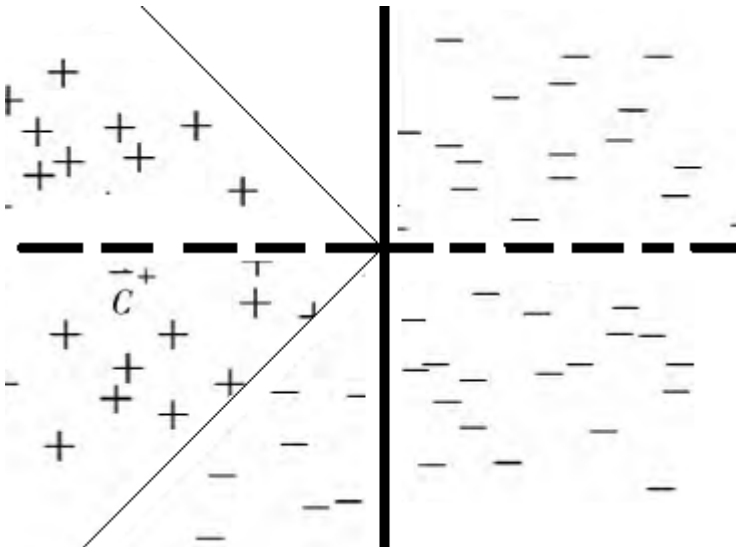

Fig.2 Classification model

3. Experimental Results and Performance Analysis

To verify the validity of the proposed model, this section applies it to real data and compares it with the dynamic baseline model proposed in the literature. [4]

3.1 Data Preprocessing

First, the original data is preprocessed. In order to make the baseline prediction more accurate, we distinguish the weekday data from the weekend data. For missing data, it can be considered that the numerical distribution of data on the same node is basically the same, so from a horizontal point of view, the missing values at a fixed node on a certain day are filled with the mean values of adjacent nodes in the previous days. If there are some missing data in the previous days, the longitudinal

distribution law of the data on that day is excavated, and the missing value is filled according to the distribution.

According to the characteristics of the data, we select two representative features for the original data, namely, transaction volume and average response time. Secondly, considering that the dimension levels of these two features are quite different, the maximum and minimum normalization is carried out to normalize the interval between the features [0,1]. For ml data sample set, the normalized point of Zi of the first sample point is ri, then RI = zmax-zi zmax-zmin. Here, Zmax and Zmin are the maximum and minimum values of the characteristic data, respectively. Similarly, each indicator (n in total) can be equally standardized. In this article, it may be noted that the standardized data is still T = xij}.

3.2 Parameter Selection

If the parameters of semi-supervised learning model are fixed, the performance of the whole model will be greatly reduced. Therefore, it is necessary to choose the best training days and parameters in the model as the best values. At the same time, the model only needs to consider the best parameter C and the best choice of training days. When the number of days is too small, the prediction results will be unstable. When the number of days is too large, because the data flow is time-sensitive, there is no reference value for the earlier data, which will affect the prediction results. Based on the above analysis and numerical experiments, absolute error is used as the test standard.

When choosing the parameter C of model (2), we use the grid search method. Firstly, based on experience, the range of the two parameters is determined to be {2-8, 2-7, ... 27, 28}. In this range, penalty parameters are optimized.

3.3 Prediction of Baseline Values and Determination of Early Warning Thresholds

Using semi-supervised learning technology, the baseline values of the new day are weighted by the previous baseline values and the real values of several days. The dynamic baseline based on semi-supervised learning method outperforms the fixed baseline. Firstly, absolute pair error P is used to determine the threshold value: P = xi, L + 1-yi |, where xi, L + 1 is the true value of the predicted day and Yi is the fitting baseline value. In the process of data processing, the error distribution of judgment index can be obtained by extracting the data of judgment index. Since all known data are assumed to be real and no abnormal data, the threshold can be determined according to the error distribution map. We normalize the value of the judgement index in the range of 0-1, as shown in Figure 3.

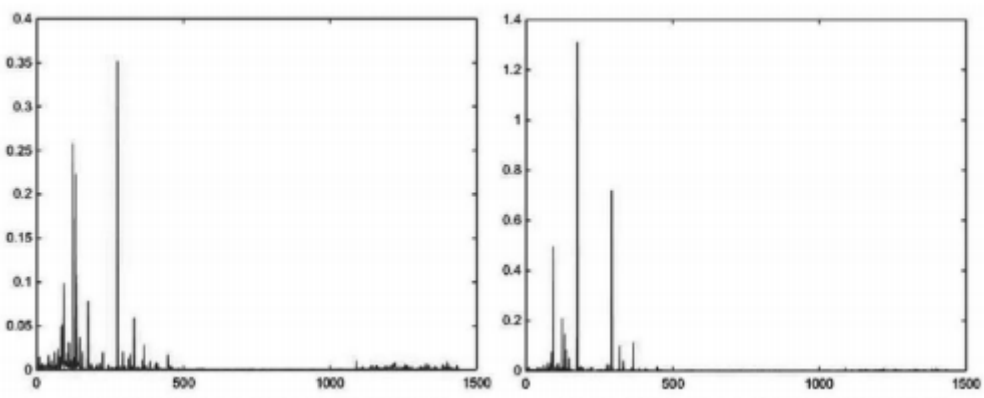

Fig.3 Prediction of baseline values

4. Summary

This paper presents a method to establish the safety baseline of industrial control network based on traffic analysis technology. This method can solve the dilemma of incomplete information and difficult coordination of resources when industrial control information security work is carried out by the owner unit of industrial control system. With this method, only the existing equipment is used to

construct the laboratory outside, complete the reverse analysis of the control protocol, or directly use the characteristic library of the industrial control network protocol mastered by the industrial control information security manufacturer, without coordinating the design units, integration units, industrial control equipment providers and other construction units, which greatly reduces the development. Threshold of information security in industrial control. At the same time, the bypass access control network of this method does not need shutdown cooperation, and does not affect the business continuity requirements in the production environment. In the process of establishing the network security baseline, the network communication situation of the host computer of the industrial control system can also be deeply analyzed, and it is possible to find malicious code that has been latent in the industrial control network.

Because of the universality of the network protocol of industrial control system, this paper has a certain general significance for other process control systems in the industry, even for other industries. In the long run, the maintenance of safety baseline is a long-term process. [5] The establishment of information security baseline of industrial control network mentioned in this method is the starting point of isolation, purification and monitoring of industrial control network around safety baseline. Enterprises also need to establish long-term baseline monitoring, change and maintenance mechanism to maintain in industrial system. When changing, adjust the baseline in time to ensure synchronization of the baseline control system.

References

[1]. Jolliffe I. Principal component analysis[M]. John Wiley &Sons, Ltd,2002.

[2]. Peng Yong, Jiang Changqing, Xie Feng, etc. Research on Information Security of Industrial Control Systems Progress [J] Journal of Tsinghua University (Natural Science Edition), 2012 (10):1396-1408.

[3]. Yang Chen, Ma Qin. Weaving Safety Net for Industrial Control System [J] Information Security and Security Communications Secrecy 2014, (6:36).

[4]. Jolliffe I. Principal component analysis[M]. John Wiley &Sons, Ltd,2002.

[5]. Van Gestel T, De Brabanter J, De Moor B, et al. Least squares support vector machines [M]. Singapore: World Scientific,2002.