Basic Intrusion Technology of Industrial Internet of Things—Based on Machine Learning

Yizhan Wu
School of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin, 300131, China
*Corresponding author’s e-mail: 175197.stu.hebut.edu.cn

Abstract. With the continuous advancement of technology and the continuous development of the Industrial Internet of Things, network security has become particularly important, especially in the detection of intrusion attacks. This article briefly introduces the Industrial Internet of Things, and the current network security threats faced by the Industrial Internet of Things, as well as analyzes different intrusion methods. Three machine learning algorithms are applied in this article to analyze and compare their effects in intrusion detection. Through experiments, a more suitable machine learning model for industrial IoT intrusion detection is selected.

1. Introduction
In recent years, in terms of information security, the industrial Internet of Things is full of crises. Hackers have used various methods to attack industrial IoT applications in order to damage the system or steal data. Therefore, in order to protect the Industrial Internet of Things from hackers, this article conducts machine learning-based intrusion detection model experiments. There are some introductions about the IIOT. Humans have carried out four industrial revolutions. The First Industrial Revolution is Watt’s invention of the steam engine, ushering in the era of machine manufacturing. Then Second Industrial Revolution comes to the popularization of electricity, which has realized mass production of products. Next, the Third Industrial Revolution is of the application of integrated circuits to realize automated production.

The concept of Industry 4.0 was first proposed by the German government in 2013, and it became a national strategy. Until January 31, 2017, the Industrial Internet Consortium (IIC) officially released the new Industrial Internet of Things reference architecture IIRA version v1.8, which is an upgrade to the first version launched a year and a half ago. The Industrial Internet of Things is composed of industrial systems and the Internet, which is the integration of advanced computing, analysis, and sensor technologies, as well as the process of industrial production and the Internet. It integrates manufacturing and monitoring manufacturing control, enterprise management, supply chain and customer feedback into an information system, which can intelligently process data from the same channel through the data center, thereby improving production efficiency, product quality and user satisfaction.

2. Hidden Dangers of Information Security Faced by Industrial IoT
Industrial Internet of Things has four underlying core technologies, including sensor technology, data acquisition technology, data transmission technology and data processing technology. There are still several unresolved problems and loopholes in these traditional technologies. However, with the
The continuous development of the Industrial Internet of Things, these problems have not decreased but gradually increased. With the combination of technologies, lots of new problems have arisen. The information security risks faced by the Industrial Internet of Things can be divided into three categories. The first category is security risks caused by the internal structural characteristics of the Industrial Internet of Things, and the second category is external intrusions from the Internet. The internal structure of the Industrial Internet of Things and the Internet of Things are similar, consisting of a perception layer, a network transmission layer, an intelligent processing layer and an application layer.

2.1. Perception Layer
The perception layer is the foundation of the Industrial Internet of Things. It is composed of a large number of sensors, RFID, and wireless sensor networks to realize the collection and transmission of underlying data. There are two possible threats, one is physical threats and the other is software threats. The physical aspect is that the hardware system of the perception layer is susceptible to man-made destruction, and once it is destroyed, it is prone to serious chain reactions. In terms of software, the node function of the perception layer is relatively single. Hackers can easily hijack the node and sneak into the industrial Internet of Things to perform data collection and information theft.

2.2. Transport Layer
The Industrial Internet of Things is a collection of multiple systems and multiple platforms. As many sensors are used, with their different interfaces, so in order to ensure the coordination and compatibility between them, a large number of network ports are opened, which makes data easy to be lost and stolen during transmission and this consumes a lot of resources.

2.3. Application Layer
The goal of the Industrial Internet of Things is to connect products and users. Therefore, the application layer is a layer that directly faces the application masses and management. The application layer is filled with a lot of data, so it is easy to cause data loss and leakage, such as customer personal information and product information.

3. External Network Attacks and Intrusion Methods
The Internet has always been a foundation of the Industrial Internet of Things, so in the Industrial Internet of Things, the threat of network intrusion cannot be avoided.

3.1. Common Ways of Network Intrusion
Network intrusions are not uncommon in our daily lives. Below are ten methods of network attacks that we may daily encounter.

| Threats                  | Description                                               |
|-------------------------|-----------------------------------------------------------|
| 1. Tapping              | Sensitive information transmitted on the network was tapped.|
| 2. Re-transmission      | Attackers resend some or all of the information obtained to the receiver.|
| 3. Forgery              | The attacker sends the forged information to the recipient.|
| 4. Distort              | Attackers modify, delete, insert and send communications.|
| 5. Unauthorized access  | Access the system through forgery, identity attacks, system vulnerabilities, etc.|

Table 1 Ten methods of network attacks
6. Denial of service attack

Attackers will slow down the system response or even paralyze it, thereby preventing legitimate users from obtaining services.

7. Behavior denial

Corresponding entity denies the behavior that has occurred.

8. Bypass control

The attacker discovers a system vulnerability or security breach.

9. Electromagnetic and radio frequency interception

Attackers extract information from electromagnetic radiation.

10. Personnel negligence

Authorized persons leak information to unauthorized persons for profit or due to carelessness.

3.2. Attack Methods Against Industrial Internet of Things

Although the core of the Industrial Internet of Things is a variety of nodes, the Internet is still a crucial tool in the application of Industrial Internet of Things, which leads to some of the current security threats facing the Internet will also be reflected in the Industrial Internet of Things [1].

3.2.1. Destroying the network system. This attack mode is to attack the core of the system through system vulnerabilities and service interference. There is also a large number of port scans to discover vulnerabilities in the system, then to carry out large-scale flooding attacks and dos attacks. In this way, the resources in the system will be continuously consumed.

3.2.2. Data-stealing attacks. This type of attack usually uses implanted viruses or sneaks into the service system through system vulnerabilities, and then steals user information and commercial secrets. This kind of attack is usually very covert and will not cause serious damage to the system like the destructive attack introduced in the first.

3.2.3. Host and control attacks. This kind of hijacked attack method is mainly to attack network nodes, and conduct a series of information theft and attack behaviors by intruding into the gateway node of the industrial Internet of things, and forcibly modify the configuration of software and hardware, such as monitoring and stealing user information and launching large-scale ddos attacks.

3.2.4. Camouflage attack. The most common phishing and disguised attack methods are those phishing websites that defraud users’ credit card passwords and other methods to steal money. In the Industrial Internet of Things, hackers can easily obtain the administrator’s user information through phishing websites, and then pretend to be an administrator to carry out a series of system attacks.

4. Intrusion detection technology based on machine learning

In the Industrial Internet of Things, in terms of detection technology, intrusion detection can be divided into two types, which are abuse detection and anomaly detection. The intrusion detection based on machine learning in this paper belongs to anomaly detection. If a large-scale unconventional phenomenon occurs in the system, the intrusion behavior of the intruder can be detected in time. In recent years, machine learning has developed rapidly, and it has excellent effects in intrusion detection. For example, Shi-Jinn Horng proposed a SVM-based intrusion detection in 2011 [2]. This method can detect dos attacks more effectively. In addition, Luo Yaofeng’s data model based on industrial control protocol was proposed in 2013 [3], which can use support vector machines for classification. In this article, three different machine learning methods are used to detect intrusions and compare the results.
4.1. Typical means of intrusion

In fact, there are only four most common intrusion methods in the Industrial Internet of Things, including monitoring, interruption, modification and disguise (see table 2).

Table 2: Four attack types in the IIOT

| Attack type            | Description                                                                 |
|-----------------------|-----------------------------------------------------------------------------|
| Monitoring            | Hackers can obtain any information in the Industrial Internet of Things through monitoring, including data of each node, user information and some commercial secrets. |
| Interruption          | Interruption refers to interruption of service, which paralyzes communication in the industrial Internet of Things. E.g. Denial of Service |
| Modification          | The hackers modify the information of the invaded target to achieve the purpose of the attack. |
| Fabrication           | Hackers pretend to be administrators, attack and steal through their identity. |

In this experiment, the selected data are industrial network data that are provided by the SCADA Laboratory of Mississippi State University in 2015 [4]. They used multiple attack methods such as intercepting data, malicious response injection, malicious command injection, denial of service, and detection. It can be seen from the following table 3.

Table 3: Seven attack means in the IIOT

| Means of attacks                  | Abbreviation | Threat type                        |
|-----------------------------------|--------------|------------------------------------|
| Normal                            | Normal(0)    | N/A                                |
| Naïve Malicious Response Injection| NMRI(1)      | Modification/Fabrication           |
| Complex Malicious Response Injection| CMRI(2)   | Modification/Fabrication           |
| Malicious State Command Injection | MSCI(3)      | Modification/Fabrication           |
| Malicious Parameter Command Injection | MPCI(4) | Modification/Fabrication           |
| Malicious Function Code Injection | MFCI(5)      | Modification/Fabrication           |
| Denial of Service                 | DoS(6)       | Interruption                       |
| Reconnaissance                    | Recon(7)     | Interruption                       |

There are totally seven attack methods and one non-attack method. The author sets the training set and test set separately, and then randomly process the number of attacks to achieve the real network environment.

4.2. Machine Learning Algorithm

A total of three commonly used machine learning algorithms for intrusion detection are applied to the article, namely C4.5 decision tree, naive Bayes algorithm and neural network model.

4.2.1. C4.5 decision tree. Decision tree is a predictive model consisting of three parts, which are decision node, branch and leaf node. The decision point represents a test experiment, and the branch and leaf nodes represent different experimental directions and experimental results. C4.5 is a classification decision tree algorithm, which is an important algorithm based on the improvement of the ID3 algorithm. Compared with the ID3 algorithm, it can select attributes through the information gain rate. ID3 uses information gain, which is entropy, while C4.5 uses information gain rate.

The first step is to calculate the information entropy, respectively calculating the information entropy of the node and that of the leaf node.
\[ \text{Ent}(D) = - \sum_{k=1}^{n} p_k \log_2 p_k \]  

(1)

\( p_k \) is the proportion of type \( k \) samples in sample set \( D \). The more uniform the probability distribution of different categories, the greater the information entropy; the more categories, the greater the information entropy.

In the second step, the author makes the difference between the information entropy of the node \( a \) and the leaf node to obtain the information gain.

\[ \text{Gain}(D, a) = \text{Ent}(c) - \text{Ent}(c / D) \]  

(2)

where \( \text{Ent}(c) \) is category entropy and \( \text{Ent}(c / D) \) is attribute entropy.

The last step is to calculate the information gain rate \( \text{Gain}_\text{ratio}(D, a) \), which is to divide the information gain by the classification information metric.

\[ IV(a) \] is the classification information metric.

\[ IV(a) = \sum_{\mathcal{V}} \frac{[D]}{[D']} \log_2 \frac{[D]}{[D']} \]  

(3)

\[ \text{Gain}_\text{ratio}(D, a) = \frac{\text{Gain}(D, a)}{IV(a)}. \]  

(4)

When making decisions, the author first selects attributes with high information gain, and then look for attributes with high information gain rate from the candidate attributes.

4.2.2. Naive Bayes.

Naive Bayes algorithm is one of the most commonly used classification algorithms. The reason why it is called the Naive Bayes algorithm is mainly because the basic principle of the algorithm is based on Bayes’ theorem, and the premise of its establishment is that these functions must be independent. Therefore, the theoretical basis of the Naive Bayes algorithm is a classification method based on Bayes’ theorem and conditional independence assumption. Although each piece of data is not independent of each other in the Industrial Internet of Things, the naive Bayes model is also applicable to the IIOT.

The first stage is preparation. delineating the training set, where the training set contains a samples \( i = 1, 2, ..., x \), and each sample contains \( b \) features.

The second stage is the training stage, which is to generate a classifier, mainly to calculate the frequency of each category in the training sample and the conditional probability of each feature attribute division of each category. The input is the feature attributes and training samples, and the output is the classifier. The author separates \( m \) classes }\{y_{1}, y_{2}, \cdots, y_{m}\} and \( n \) features for classification through the classifier.

As each feature is assumed to be independent in the naive Bayes model, the model formula is:

\[ p(y_{m} \mid x) = \frac{p(y_{m})p(x \mid y_{m})}{p(x)} = \frac{p(y_{m})\prod_{i=1}^{b}p(x_{i} \mid y_{m})}{\prod_{i=1}^{b}p(x_{i})} \]  

(5)

The third stage is the application stage. This stage is to use the classifier to classify the new data. The \( y_{m} \) at the maximum value is the judgment result.

\[ y = \arg\max_{y_{m}} \, p(y_{m} \mid x) \]  

\[ = \arg\max_{y_{m}} \, p(x_{1} \mid y_{m}) \cdots \arg\max_{y_{m}} \, p(x_{b} \mid y_{m}) \]  

(6)
The advantages of the naive Bayes model are simple calculation and high efficiency. However, the shortcomings are also obvious. For example, the classification performance is not high, and the conditions for the independence of the features are too harsh.

4.2.3. Convolutional neural networks.
Convolutional neural network is a feed forward neural network with deep structure and a convolution algorithm. Convolutional neural network can be divided into 5 layers, input layer, three hidden layers and output layer. The hidden layer includes a unique convolutional layer, a pooling layer and a fully connected layer, which are shown in figure 1 [5].

![Convolutional neural network structure diagram](image)

In the first step, the convolution layer applies multiple convolution filters (convolution kernels) to the image. Next, for each sub-region, the layer performs a set of mathematical operations with a specific spatial range and step value to generate a single value in the output feature map and perform nonlinear operations.

In the second step, the hybrid layer performs dimensionality reduction processing on the image data extracted by the convolutional layer, in order to reduce the dimensionality of the feature image, retain the maximum value, and discard other values. Finally, the fully connected layers classify the features extracted by the convolutional layer. In a fully connected layer, it is necessary to ensure that each node in the layer is connected to the corresponding node in the previous layer.

In the intrusion detection system of this paper, the use of two-layer convolution and pooling loss function is to define the cross-dimension to represent the distance between the predicted result and the real result. Convolutional layers use the same parameters: stride, S=1; zero padding, P=0. ReLU activation function keeps the volume of each layer unchanged. In order to prevent the model from overfitting, the author uses the L2 regularization method, and add the Dropout layer after the fully connected layer. In addition, the author uses the drop-out parameter d=0.5, and then get the classification probability through the Softmax layer. The last category is the category with the highest probability.

5. Experimental Results and Data Analysis
At the end of the experiment, the author used two criteria to judge the ability of the machine learning model to distinguish intrusion attacks, which are the precision rate and the recall rate. The precision rate is the ratio of the number of correctly classified samples to the total number of samples for a given data set. The recall rate is used to describe the ratio of the positive cases judged to be true in the classifier to the total positive cases.

\[
\text{precision} = \frac{\text{instances classified correctly}}{\text{total instances classified}}
\]  

(7)
Table 4 Statistics

| Category | C4.5 | NBM | CNN |
|----------|------|-----|-----|
|          | Precision | Recall | Precision | Recall | Precision | Recall |
| Normal   | 99.6%  | 99.8% | 98.6%  | 96.6%  | 99.8%  | 96.4%  |
| NMRI     | 95.6%  | 81.2% | 70.5%  | 3.8%   | 77.8%  | 84.4%  |
| CMRI     | 97.1%  | 84.6% | 71.6%  | 4.1%   | 86.2%  | 87.6%  |
| MSCI     | 97.4%  | 97.0% | 73.4%  | 14.4%  | 88.4%  | 87.5%  |
| MPCI     | 98.1%  | 96.9% | 92.0%  | 40.2%  | 91.6%  | 80.6%  |
| MFCl     | 99.1%  | 100%  | 93.4%  | 50.1%  | 94.5%  | 100%   |
| DoS      | 98.7%  | 95.8% | 80.7%  | 34.4%  | 93.4%  | 99.4%  |
| Recon    | 98.2%  | 97.9% | 72.8%  | 12.8%  | 90.1%  | 97.5%  |

Through data observation, in the experimental data, it is not difficult to find that the C4.5 decision tree model has the best precision and recall rate, followed by the convolutional neural network model, and the last one is the naive Bayes model. Although Naive Bayes is good in terms of precision, it can be said to be devastating in terms of recall.

6. Conclusion
This article briefly introduced and analyzed the security crisis faced by today's industrial Internet of Things at the beginning. In the following part, the author uses three machine learning models to simulate intrusion detection. Through input analysis, the C4.5 decision tree model is superior to the other two models, so the author reckons that it is more suitable for intrusion detection systems of the Industrial Internet of Things. The number of machine learning algorithms used in this experiment is relatively small, and more algorithms will be applied to intrusion detection in the future, and then the experimental results will be compared again.

Acknowledgments
Thanks to my professors and class teachers for helping me in this experiment, allowing me to go further in machine learning algorithms, network security and modeling. And I am very grateful to the teachers of the thesis group for revising and correcting my thesis, so that I can learn more about the writing requirements of the thesis.

References
[1] Zhang, M. (2017) Industrial Internet of Things Security Risk Analysis and Countermeasure Research, China Industry Review, pp. 44-46.
[2] Horng, S. J., Su, M. Y., Chen, Y. H., et al. (2011) A novel intrusion detection system based on hierarchical clustering and support vector machines [J]. Expert systems with Applications, 38(1): 306-313.
[3] Luo, Y. F. (2013) Research and design of intrusion detection methods for industrial control systems. Diss. Zhejiang University.
[4] Turnipseed, I. P. (2015) A new scada dataset for intrusion detection research [J]. Dissertations & Theses -Gradworks, pp. 120-135.
[5] Wang, J. J. (2019) Network Intrusion Detection System Model Based on Convolutional Neural Network, Journal of Hebei University of Economics and Business, March, Volume 19, Issue 1.