Instruction Detection in SCADA/Modbus Network Based on Machine Learning

Haicheng Qu[1], Jitao Qin[1](✉), Wanjun Liu[1], and Hao Chen[2]

[1] Institute of Software, Liaoning Technical University, Huludao 125105, China
quhaicheng@lntu.edu.cn, lgc_qinjitao@sina.com,
liuwanjun39@163.com

[2] Department of Information Engineering, Harbin Institute of Technology,
Harbin 150001, China
hit_hao@hit.edu.cn

Abstract. Cyber security threats of industrial control system have become increasingly sophisticated and complex. In the related intrusion detection, there is a problem that intrusion detection based on network communication behavior cannot fully find out the potential intrusion. The Machine Learning is applied to seek out the abnormal of industrial network. First of all, the supervised learning methods, such as Decision Tree, K-Nearest Neighbors, SVM and so on, were adopted to deal with SCADA network dataset and related discriminated features. Next, an anomaly detection model is built using One-Class classification method, and the effect of the One-Class Classification method in the SCADA network dataset is analyzed from the recall rate, the accuracy rate, the false positive rate and the false negative rate. It is shown that the anomaly detection model constructed by the One-Class Support Vector Machine (OCSVM) method has high accuracy, and the Decision Tree method can commendably detect the intrusion behavior.

Keywords: Cyber security · Intrusion detection · Supervised learning
OCSVM

1 Introduction

1.1 Survey of Industrial Network Intrusion Detection

At present, most of the key infrastructure and industrial systems, such as oil and gas pipelines, water treatment systems, reservoir valve control, power grids and nuclear power plants, are all controlled by the Supervisory Control and Data Acquisition system (SCADA) [1]. These systems allow remote monitoring and remote access and control for geographically dispersed facilities, enabling operators in these facilities to monitor and control the entire system in real time. Traditionally, this control network is completely isolated in the application environment.

However, with the development of industrial control systems and traditional computer networks, industrial control networks have become fully open or semi-open networks in order to facilitate the control of industrial processes and the combination of industrial control networks and Internet networks. As the traditional industrial control

X. Gu et al. (Eds.): MLICOM 2017, Part II, LNICST 227, pp. 437–454, 2018.
https://doi.org/10.1007/978-3-319-73447-7_48

system is based on physical isolation, the main focus on data transmission real-time and accuracy, did not take into account the aspects of Cyber security, and lacking of appropriate protective measures. So in recent years for industrial control equipment, more and more attacks, especially in 2010, the outbreak of the complex malware Stuxnet virus, exposed the industrial network there are major flaws. In view of this, industrial safety research is of great significance. The intrusion detection technology as a Cyber security protection of the first step in industrial Cyber security has a very important sense.

Network intrusion detection, through the network data and behavior patterns, determine whether the Cyber security system is valid. Its purpose is to identify intruders, identify intrusion, and monitor successful security breakthroughs, and provide important security information. In fact, there is no essential difference between the industrial network and the traditional computer network; they check the network or system for violations of security policies and signs of attacks by collecting and analyzing network behavior, security logs, audit data, information available on other networks [2, 3], and information about key points in the computer system.

Industrial network intrusion is related to the structure of industrial control system. The current industrial control system [1] in the specific deployment usually involves the following network: enterprise office network (enterprise network or office network), process control and monitoring network (monitoring network), field control system.

Office network: According to the data of the monitoring network, the managers manage the enterprises and make the decision. Through the industrial management system for enterprise planning production, storage management, production scheduling and other processes, the managers conduct a unified deployment.

Monitoring network: The operator monitors and controls the field running equipment through the SCADA system.

Field control system: It includes distributed control system (DCS), programmable logic controller (PLC), and remote terminal unit (RTU). In these systems, the staff on the field device performs logical control, data sampling, instruction execution and other operations.

1.2 Advances in Related Fields

The study of industrial network intrusion detection does not have a mature theoretical system so far. Therefore, most of the researches on intrusion detection with industrial network are based on the traditional Internet network intrusion detection method.

In the paper [4], the intrusion detection method, combined with SCADA network and traditional intrusion detection, complete the industrial intrusion detection task; it also discusses the attack patterns of industrial network attackers and the vulnerability of industrial networks. The method of Kernel entropy [5] is used to ensure that the original data information loss is minimized in the process of dimensionality reduction, and then the intrusion detection model is constructed by artificial immune method, which avoids the high-dimensional space and the immune algorithm on the lack of data coverage shortcomings. However, the artificial immune algorithm cannot be detected for variant attacks in industrial networks. [6] It studies the function code characteristics in the Modbus/TCP protocol. Using the depth packet analysis technique, the function code in the protocol is parsed, and then the pattern matching method is used to match the rules to judge the abnormality or normal. It is ideal to use intrusion detection rules in real-time aspect, and could detect some of the attacks, but only the use of functional code analysis is not comprehensive enough.

Machine Learning and Deep Learning in recent years for the development of network intrusion detection technology provides a new impetus. [7] It is a combination of stratified misuse detection and anomaly detection combined with intrusion detection. The algorithm uses C4.5 to construct the misuse detection model, divide the normal training data into multiple subsets, and then use the subset data to create multiple OCSVM models. The process not only improves the detection accuracy, but also reduces the training and testing time. However, ignores the possible relevance of the subset. In this paper [8], the approach of isolating forest is put forward in view of the fact that the number of abnormal samples is small and the majority of the data are isolated. It constructs the isolation tree, divides the data space by the isolation tree, and judges the abnormality according to the path length of the tree structure. The method has the characteristics of low computational cost, fast operation speed, high detection accuracy of most abnormal points, and has been widely used in the field of abnormal network identification, factory production process abnormal judgment and financial data field.

In paper [9], PLS (Partial Least Squares) and CVM (Core Vector Machine) were adopted to construct the intrusion detection model in order to address the feature selection and large sample data. Experiments show that this method has high accuracy, low false alarm and false negative rate and high real - time performance compared with intrusion detection using SVM method. The SVM algorithm is combined with D-S evidence theory [10]. Similar to the method of ensemble learning, a number of SVM classifiers with different differences are combined by D-S evidence theory to obtain the final classification result. In paper [11], the PSO automatically finds the parameters that adapt to the SVM algorithm in a specific industrial control network, which improves the classification accuracy of SVM algorithm. Before, the dataset can be reduced dimension by KPCA method to improve the training speed of the classifier.

In recent research, Intrusion detection in industrial control network focus on ICS security schema which is mainly upon the traditional active defense solution. However, the ICS security schema could not be used to effectively forecast and control under the condition of high real-time and resource constraints. Therefore, in recent years, industrial network intrusion detection is the hotspot in this field. Machine learning and Data mining technology is able to find hidden attributes in the data. The machine learning method is applied to intrusion detection which is the trend of intrusion detection technology development in recent years.

This paper begins with the following sections. The first section introduces the industrial network and intrusion detection, and explains the research status of industrial network intrusion detection. In the second section, the principle of Machine Learning and One-class classification is expounded in detail. The third section is the experiment on the SCADA dataset. The final section is the experimental conclusion.

2 The Application of Machine Learning in Intrusion Detection

Machine learning algorithm [12], from the data itself, uses statistical laws and mathematical knowledge to dig out the information contained in the data. The traditional machine learning algorithm solves the problem of data classification, regression, and clustering and association rule learning and so on. In recent years, machine learning has achieved great success in text classification, natural language processing, machine translation, situational awareness, and image processing and computer vision. This also stimulates the development of machine learning methods to other areas.

2.1 Overview of Traditional Machine Learning Methods

In the SCADA dataset, intrusion detection can be seen as a multi-class approach in nature, and there are many mature machine learning algorithms in completing the classification task. In the following, we mainly from The Decision Tree, Logistic Regression, K-Nearest Neighbor and other algorithms to analyze the industrial network data.

(a) Logistic Regression
Logistic regression can be used as a regression algorithm, but also as a classification algorithm, which is a linear regression. It is a process that deals with a regression or classification problem, establishes a cost function, then solves the optimal model parameters by optimizing the method, and then tests to verify that the model we are solving is good or bad. The advantage is that it can be a linearly separable data set to construct a good linear model, and fast, easy to understand; of course, for non-linear data, will not complete the classification task.

(b) K-Nearest Neighbor
K-Nearest Neighbor is one of the traditional classification algorithms, and it is also one of the simplest classification algorithms. The idea of this method is that in the feature space, if the majority of K samples in the vicinity of a sample (that is, the nearest neighbor in the feature space) belong to the same category, the sample also belongs to this category. In the SCADA network data set, the normal data set has such a characteristic that the distribution of normal data sets is relatively concentrated. Using K-Nearest Neighbor algorithm to analyze SCADA network data set should be able to get a good classification effect.

(c) The Decision Tree
The Decision Tree algorithm is one of the data mining algorithms, with strong intuition. It is a typical classification method, the first data processing, and then uses the induction algorithm to generate the relevant rules, and finally use the decision to analyze the new data. On the other hand, the decision tree is an algorithm that classifies data according to rules. For SCADA Cyber security datasets, there is a clear correlation between some intrusion and rules, and the decision tree can mine the relevant rules and store them.

(d) Support Vector Machine (SVM)
Support vector machine algorithm is one of the best machine learning algorithms to solve small sample, non-linear and high dimensional data classification, and has the best classification effect and generalization ability. But for the large sample data, there is a long calculation time and low real-time performance. So in the actual production, using SVM algorithm is relatively small. But as one of the best classification methods, SVM algorithm can show a good classification effect on SCADA network dataset.

(e) Random Forest (RF)
Random forest is a kind of integration method, which integrates and evaluates the classification result of multiple weak classifiers, and finally obtains the whole classification effect. The random forest algorithm was proven to have a faster training speed and a higher accuracy rate in a variety of classification cases. Therefore, applying this method to the SCADA Cyber security datasets can improve the classification accuracy of a single classifier.

These approaches aim at helping the traditional IDS in detecting malicious activities and cyber-attacks threatening the critical infrastructures.

2.2 Anomaly Detection Algorithm Based on One-Class Classification

In the traditional intrusion detection, mainly divided into abnormal detection and misuse detection two categories. The anomaly detection is to establish a model for normal data, through which the intrusion event can be identified, but the method cannot accurately identify the specific category of intrusion; Misuse detection is to build a classification model for a variety of intrusion behavior, this way can detect the known intrusion category, and the higher accuracy; However, if there is a new type of intrusion, the algorithm will not be able to identify the intrusion. Misuse detection model training requires known intrusion data, belonging to the scope of supervision learning. And the anomaly detection comes from the normal data model, which belongs to the unsupervised learning category.

One-class classifiers learn the normal behavior modes of the studied system [13], and determine decision rules that accept as many normal samples as possible, and detect most of the outliers; through this classifier, the dataset is divided into parts that are similar to the model and portions that differ from the model. The single classification method belongs to unsupervised classification method, no need to label, only need to ensure that the training classifier data to meet the normal dataset requirements. From the detection method, the single classification method is one of the abnormal detection methods.

In the process of using traditional supervised learning methods, the linear indivisibility, which is the characteristic of this SCADA network dataset, is solved by using the kernel function method for the majority of machine learning algorithms. This method maps the data to a higher dimension space (Reproducing Kernel Hilbert Space) where the data can be linearly segmented. The data mapping process cannot be shown, so the use of a function conversion calculation. In this paper, the kernel function is Gaussian kernel function (also known as RBF function), as shown in formula (1).

$$k(x_i, x_j) = \exp\left(-\frac{\left\|x_i - x_j\right\|_2^2}{2\sigma^2}\right) \tag{1}$$

This chapter analyzes the classification effect of the One-Class classification the One-Class classification method based on kernel function, the One-Class classification method based on KPCA and the One Class Support Vector Machine (OCSVM).

(a) The One-Class classification based on Kernel function [13, 14]
In this method, the decision function in this approach is based on the Euclidean distance in the feature space between the samples and the center of the hyper sphere. Let c_n be the center of the data in the feature space estimated using all available training samples. And $c_n = (1/n)\sum_{i=0}^{n}\phi(x)$. The expression of the squared distance between any sample $\phi(x)$ and the center c_n can is given as follows:

$$\|\phi(x) - \phi(x_0)\|_H^2 = k(x,x) - \frac{2}{n}\sum_{i=1}^{n} k(x, x_i) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) \tag{2}$$

In this measure, after assessing the distance between all training samples and the estimation center, use the estimated threshold to evaluate the calculated sample distance. If the distance is farther than a predetermined threshold, the data would be treated as an outlier point. The above algorithm needs to estimate the threshold to evaluate. However, if the threshold is not appropriate, the classification effect will be reduced and performances become terrible.

(b) One-Class Support Vector Machine (OCSVM)
OCSVM [15] is an extension of traditional SVM in unsupervised learning. The algorithm aim to construct a hyper plane which can classify the normal sample set as much as possible. The model constructs a decision function based on the hyper plane to determine which side of each data is in the hyper plane. Figure 1. Two - dimensional Case of OCSVM Method displays the situation in the classification method in two-dimensional space. Based on the principle of separating the origin from the normal data as much as possible, the algorithm constructs the quadratic convex optimization function, such as formula (3).

$$\min_{\omega,\rho,\xi} \frac{1}{2}\|\omega\|_H^2 + \frac{1}{vn}\sum_{i=1}^{n}\xi_i - \rho \tag{3}$$

$$s.t. \langle\omega, \phi(x_i)\rangle_H \geq \rho - \xi_i, \xi_i \geq 0 (i = 1, \ldots\ldots, n)$$

Where the slack variables $\xi_i \geq 0$ penalize the excluded samples and the tunable parameter v represents an upper bound on the fraction of outliers. By solving the problem of nonlinear programming, the parameters ω and the parameter ρ are solved. And build the decision function on the sample:

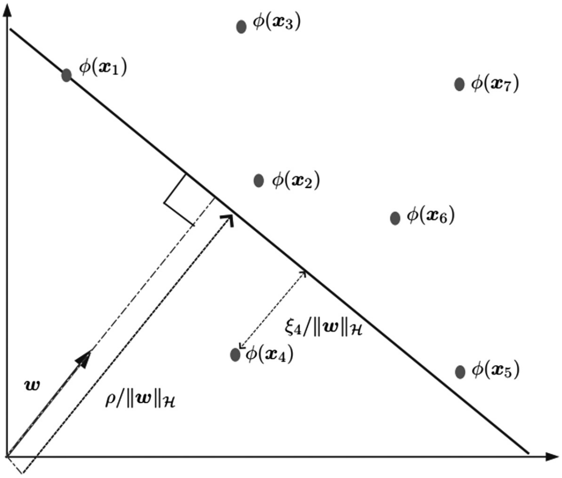

Fig. 1. Two - dimensional Case of OCSVM Method

$$f(x) = \langle \omega, \phi(x_i) \rangle_H - \rho \tag{4}$$

When $f(x) \geq 0$, the data become the normal sample; otherwise, the sample belongs to the outliers between the origin and the hyper plane.

(c) Based on KPCA One-Class Classification Method [13, 14, 16]
The KPCA is a nonlinear extension of the PCA (principal component analysis method) in a Kernel-defined feature space. KPCA extracts the subspace from the maximum variance of the data and performs dimension reduction and denoising by projecting the sample into the subspace. In the use of KPCA as anomaly detection, it introduces a new method of measurement that the data can be sorted by One-Class method. The use of KPCA classification method want to avoid the impact of noise in the data, but the calculation may be very large; this is a drawback of the algorithm. The measure is as shown in (5).

$$p(x) = \|\phi(x) - \phi(x_0)\|_H^2 - \|W\phi(x) - W\phi(x_0)\|_H^2 \tag{5}$$

where

$$f_l(x) = \sum_{i=1}^{n} \alpha_i^l \left[k(x, x_i) - \frac{1}{n} \sum_{r=1}^{n} k(x_i, x_r) - \frac{1}{n} \sum_{r=1}^{n} k(x, x_r) + \frac{1}{n^2} \sum_{r,s=1}^{n} k(x_r, x_s) \right]$$

Likewise, after using the above method to evaluate the distance between all training samples and the estimation center, the calculated sample distance is evaluated using the estimated threshold. If the distance is greater than a predetermined threshold, the sample is treated as an outlier. Such as the Sample One-Class Classification, it is critical to the KPCA Classification that the threshold setting.

3 Experiment

The experimental process is mainly used to introduce the supervised learning method and the single classification method, using the operating system for windows 7; the experiment is implemented in python, using the IDE for pyCharm; the library files used include: numpy, scikit-learn 0.18, scipy and so on.

3.1 Industrial Control Cyber Security Datasets

At present, there is no uniform standard for public datasets related to industrial control Cyber security. First, because of the industrial control Cyber security research has been developed in recent years, the industry does not have a unified dataset; second, due to the variety of industrial production processes and industrial configuration networks, it is inconvenient to have a unified data set that describes the prevailing industrial control Cyber security. Most of the research focuses on a network intrusion under a particular process of production.

Next, we introduce two of the most used industrial control Cyber security datasets: gas pipeline real dataset and storage tank dataset. The dataset includes network traffic, process control, and process measurement functions from a set of 28 attacks on two industrial control systems using the MODBUS application layer protocol. It can easily and effectively compare SCADA intrusion detection solutions.

SCADA Dataset Attack Types

The dataset apply the same type of attack [17], regardless of the normal type, a total of four major categories of 8 sub-class attacks, which contain reconnaissance attacks (RA), denial of service attacks (DOS), malicious function command injection (MFCI), malicious parameter command injection attacks (MPCI), malicious state command injection (MSCI) attacks, complex malicious response injection (CMRI) attacks, and original malicious response injection (NMRI) attacks. The following describes the contents of these attacks.

- The NMRI attack injects the response packet into the network, but lacks some information about the process being controlled and monitored; therefore some of the injected content may be invalid.
- The CMRI attacks are more complex than NMRI. Because they need to understand the physical processes of industrial control, try to mask the real state of the research process and influence the system's feedback control loops.
- The MSCI convey a malicious command to a remote field device to change the state of the physical process, driving the system from the security state to the critical state.
- The MPCI changes the controller parameter set points for sensors, actuators, and programmable logic controllers.
- The MFCI is used to affect the communication of the client server by modifying the command function field of the sent message.
- The DOS attacks try to stop the normal operation of the physical system and are designed to send the transmission speed faster than the processing speed, or by

sending an incorrect packet to change the industrial control system. This causes the operating system of the running program or target device to crash.

- The RA collects information about the control system network and builds network architecture to identify device features such as manufacturer, model, supported network protocols, and system address/memory mapping. The information collected can be used for other attacks.

The Content of SCADA Datasets

SCADA network datasets [13, 17, 18] organization has two parts: network traffic characteristics and effective content characteristics; the network traffic characteristics are related to SCADA network communications, and the effective content characteristics are related to the specific industrial control process. In the effective content characteristics, it includes the system of measured values, key system operating status parameters, system mode and other key information. One of the most important is the measured value attribute.

(a) Gas pipeline real dataset

The gas pipeline dataset contains key data for the natural gas pipeline control process, which contains 26 features. In the gas pipeline dataset, the core attribute is the pressure in the gas pipeline. After the study and analysis, the gas pipeline, there are three normal modes:

- The first mode is a relatively low pressure near 0.1 PSI.
- The second mode keeps the pressure near 10 PSI (between 9 and 11 PSI).
- The third mode maintains a pressure of about 20 PSI (between about 18 and 22 PSI).

In these three modes, the gas pipeline operates in normal condition, and most of the intrusion attacks are distributed around these normal datasets.

(b) Storage tank dataset

The reservoir dataset simulates the state of the liquid in the tank: When the liquid is located in the tank between the high and low alert position, the system is normal; when the highest alert or below the minimum alert bit, the system gives the corresponding alarm. The dataset contains 23 attribute values; some core attributes are the current water level measurement, the highest alert water level and the minimum warning level. In the storage tank system, the operator can change the corresponding system mode to change the maximum alert water level and the minimum warning level, so the attribute in the system mode is also the focus of attention in this research.

3.2 Result Analysis

The Classification of Supervise Learning Methods

Using the machine learning algorithm such as Decision tree, K-Nearest Neighbor and SVM to deal with two SCADA Cyber security datasets, the results of the classification of each method are shown in Table 1. Gas pipeline real dataset classification results and Table 2. Storage tank dataset classification results. In the evaluation of intrusion

Table 1. Gas pipeline real dataset classification results

		Logistic	KNN	DT	SVM	RF
Accuracy rate		65.41%	98.63%	98.66%	98.38%	98.77%
Recall	Normal	99.66%	99.42%	99.10%	99.45%	99.24%
	NMRI	90.88%	90.98%	92.87%	94.28%	92.44%
	CMRI	0.00%	99.85%	99.64%	98.40%	99.65%
	MSCI	0.00%	82.35%	93.22%	94.88%	94.88%
	MPCI	0.00%	98.05%	96.39%	97.43%	97.04%
	MFCI	0.00%	72.95%	93.37%	80.98%	95.46%
	DOS	0.00%	88.79%	93.03%	75.56%	90.80%
	RA	0.00%	99.19%	100.00%	99.43%	100.00%
Precision	Normal	96.85%	98.56%	98.85%	98.52%	98.87%
	NMRI	90.82%	97.44%	94.55%	88.34%	95.48%
	CMRI	0.00%	99.03%	99.23%	99.37%	99.03%
	MSCI	0.00%	96.99%	92.75%	97.38%	97.36%
	MPCI	0.00%	97.46%	97.12%	97.45%	97.56%
	MFCI	0.00%	100.00%	97.27%	100.00%	99.09%
	DOS	0.00%	99.63%	97.66%	99.86%	99.17%
	RA	0.00%	99.96%	100.00%	100.00%	100.00%

Table 2. Storage tank dataset classification results

		Logistic	KNN	DT	SVM	RF
Accuracy rate		67.20%	96.23%	98.24%	95.66%	97.08%
Recall	Normal	91.60%	97.15%	98.88%	95.85%	97.83%
	NMRI	0.00%	97.06%	97.57%	96.59%	97.23%
	CMRI	0.00%	73.00%	85.19%	82.25%	78.09%
	MSCI	0.00%	95.47%	95.42%	95.36%	95.68%
	MPCI	0.00%	96.81%	98.18%	98.42%	98.40%
	MFCI	0.00%	95.30%	100.00%	94.70%	99.55%
	DOS	0.00%	93.62%	99.19%	85.04%	98.46%
	RA	4.61%	99.98%	100.00%	99.50%	100.00%
Precision	Normal	72.06%	98.56%	98.89%	98.53%	98.32%
	NMRI	0.00%	97.44%	96.79%	98.49%	97.29%
	CMRI	0.00%	99.03%	85.19%	59.44%	72.65%
	MSCI	0.00%	96.99%	95.42%	96.95%	96.48%
	MPCI	0.00%	97.46%	98.18%	86.24%	98.98%
	MFCI	0.00%	100.00%	100.00%	100.00%	100.00%
	DOS	0.00%	99.63%	99.35%	99.34%	99.34%
	RA	4.62%	99.96%	100.00%	100.00%	100.00%

detection dataset classification effect, the use of three indicators: accuracy, precision rate and recall rate. From the above three evaluation indicators, it can be roughly described the classification effect of the classifier. In this experiment, the grid search method is used to find the optimal parameters of each classifier. The result is not only unique, but also can get the best classification effect.

According to the classification effect of these kinds of machine learning methods, there are some conclusions:

(1) The accuracy of the Logistic Regression is the lowest of the comparison algorithm (65.41% in gas data, 67.20% in water data), it can be seen that the dataset is linear inseparable. But the normal data of the recall and precision are relatively high. Through the study of the classification of the data set (gas pipeline dataset), the algorithm does not separate the normal data according to the case of confounding matrix; Since most of the attack types are close enough to the normal data set, the linear model can only identify attacks that differ greatly from the normal dataset. In short, the dataset is linear inseparable; only the non-linear classification method can solve the problem.

(2) K-Nearest Neighbor, as a method based on distance measurement, in the gas and water data sets, has achieved very good results (98.63% and 96.23% respectively). And in each attack category of recall and precision, it also has a relatively high ratio. Essentially, by computing the "similarity" on the dataset, K-Nearest Neighbor identifies K samples with the most recent "similarity" as the same category.

Using KNN classifier in the SCADA dataset and the results description that: Normal datasets are usually aggregated extensively, while exception datasets are mostly distributed at the boundaries of normal datasets. Using distance classification can separate part of the attack behavior and normal behavior. Although the overall classification of the classifier is better, but the classification rate of individual attack is still relatively low. Through the analysis of the confusion matrix, it is found that some attack categories are missing, and these attacks may be difficult to distinguish from the normal data set, resulting in being classified as normal data. KNN has a good classification effect,it shows that by calculating the distance between the samples or similarity, based on the similarity and distance comparison, you can sort out most of the attacks.

However, there are some problems with the use of the KNN method in the SCADA network datasets:

- The dimensions of the two datasets in the sample are simple. If the data collected in other industrial control networks has a high dimension, the calculation cost of KNN will be very large. It can be considered that using some dimensionality reduction ways to reduce the cost of KNN model calculations; At present, the mainstream PCA, LDA dimensionality reduction method is essentially linear dimensionality, and it is necessary to find a non-linear dimensionality reduction method to reflect the distribution of data; Dimension reduction process often leads to loss of information in the original space. Therefore, the appropriate dimensionality reduction method also needs to keep the distance between the data in the original data space as much as possible; in this way, the results in the new dimension space will be close to the results in the original dimension space.

- Although KNN are good enough for both datasets, the invasive data in different industrial control processes is different, KNN effect may be different, and even be unable to detect.

Nonetheless, KNN's performance on SCADA datasets shows that methods based on distance metrics and similarity metrics are viable in intrusion detection, but the computational cost is related to the data form and may be required when the cost is high with some distributed computing framework.

(3) The accuracy of the Decision Tree classification algorithm in the two types of datasets is 98.66% (gas dataset) and 98.24% (water dataset). The Decision Tree also has a high percentage of recall and precision rate. Decision tree is a mining and storage rules of the structure, the data show that there is a certain correlation between the data set and the intrusion category, or there are rules related to the feature in the intrusion. Using the decision tree algorithm to establish the rules and the pattern matching, finally in the two datasets, can get a high accuracy. However, in the decision tree, each of its branch nodes is a feature; there is no relevant way to consider the impact of multiple related characteristics on the classification results.

In the SCADA dataset used, since the industrial control process in the dataset is simple compared with the real industrial control process, the rules contained in it are relatively simple, and the decision tree can effectively classify the dataset. In the traditional IT network intrusion detection, the use of regular pattern matching is also used in the intrusion detection process.

(4) In the result of the classifier processing, the effect of SVM on nonlinear separable data sets is obvious and Ensemble learning on data classification results as same as SVM. The Ensemble learning is fast and the training cost is small, and it is equally feasible to classify the evaluation results by integrating multiple classifiers. As a result of the advantages of the method, in the reality of the classification task, Ensemble learning becomes more popular.

(5) As shown in Fig. 2. Comparison of Multiple Machine Learning Algorithms for Gas Dataset And Fig. 3. Comparison of Multiple Machine Learning Algorithms for Water Dataset, the gas dataset, for example, for a certain attack (MFCI, MSCI), KNN and SVM on the category of detection precision rate is low; Water dataset, for CMRI attacks, both the recall rate or precision rate, the ratio is relatively low. From the figure reflects the situation can be seen, different industrial processes focus on different attacks, the more critical attacks are more difficult to detect out.

The Decision trees, K-Nearest Neighbors, SVM, and Random Forest methods have achieved the desired results in both gas and water datasets. The above machine learning algorithm automatically establishes the classification model, and classifies the normal data and the abnormal data from the dataset. But the above machine learning method is Supervise Learning, which means that the establishment of intrusion detection model

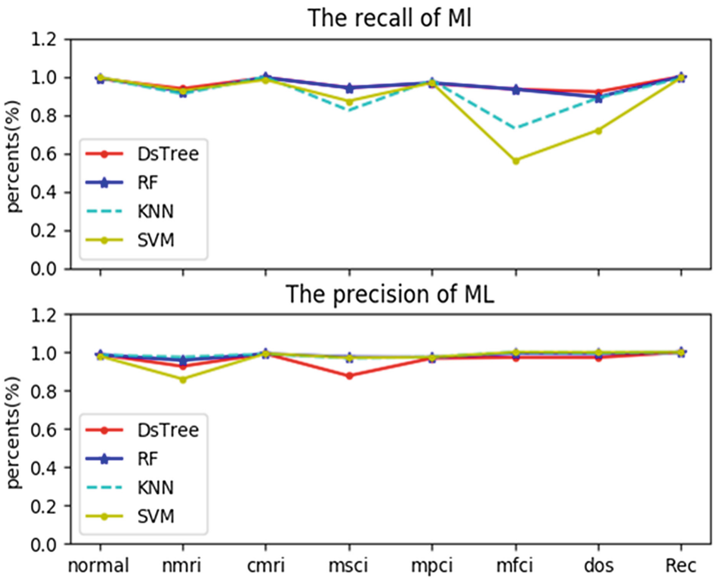

Fig. 2. Comparison of multiple machine learning algorithms for gas dataset

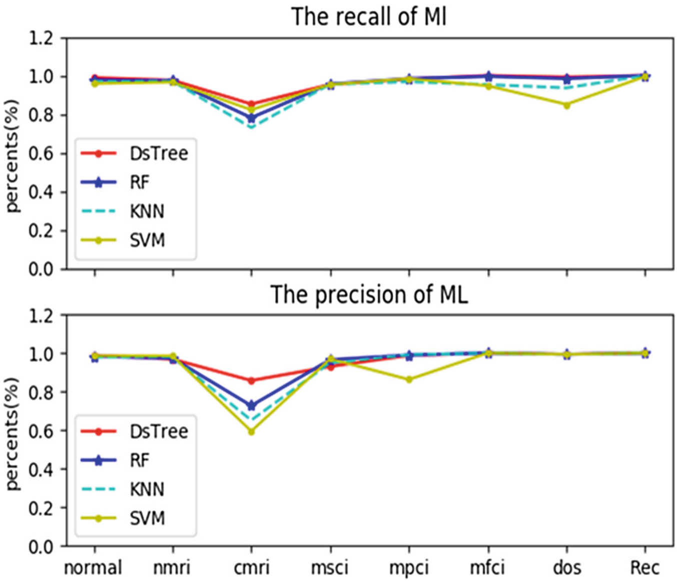

Fig. 3. Comparison of multiple machine learning algorithms for water dataset

on the premise that there is a large dataset of data, and each data intrusion category should be labeled. Otherwise, the above-mentioned algorithms cannot be used. Obtaining a calibrated data set is costly and almost infeasible, so the use of the algorithm has a lot of limitations.

In addition, the above classification, whether it is a multi-classification process or two classification process, there is little difference between the data item number in the different categories. However, if the number of data in each of these data sets is very different, for most of the supervised learning methods, there is a problem that the classifier is not adequately trained, which will seriously influence the classification results. In the real industrial control network, usually hundreds of normal data mixed with a few abnormal data; Normal data and abnormal data is seriously unbalanced, as for the use of supervised learning classification method, although the accuracy rate is high, the intrusion detection process used less. In short, the real problem leads to the traditional supervised classification method is not feasible in the real world. The use of appropriate unsupervised or semi-supervised approach is the main task of building an intrusion detection model.

The Results of One-Class Classification Experiments

In the One-Class Classification method, OCSVM, simple One-Class classification method, KPCA-based One-Class Classification method comparison results shown in Figs. 3 and 4. According to the figure, the use of OCSVM, whether in the false positive rate (FPR) and false negative rate (FNR) and the accuracy rate, has a better

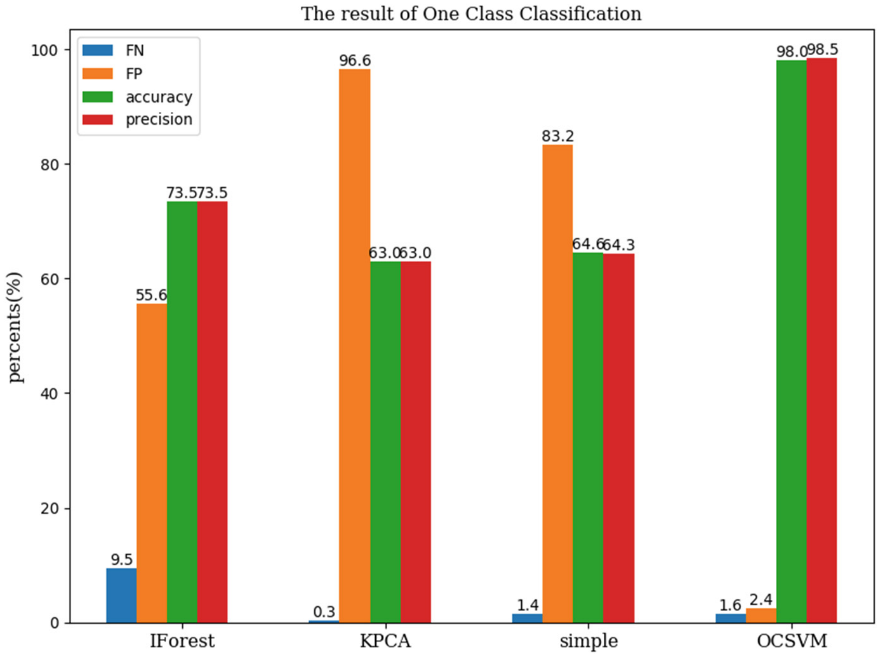

Fig. 4. Comparison of One-Class Classification of gas datasets

performance results. And this conclusion applies to two SCADA datasets in the sample. If you want to get better results, you need to adjust the algorithm's learning rate and kernel function parameters.

The results of the KPCA classification method and the simple One-Class Classification method have poor results in the false negative rate (FNR). After analysis, the threshold is not appropriate to lead to poor classification results. Solving the above problems can deal with introducing a penalty function; or use traditional supervised learning methods, such as decision tree algorithms, to learn a variety of attacks on the model

In addition, there is a lot of computing process in KPCA and Kernel method of One-Class Classification, resulting in relatively high operating costs. Although the One-Class Support Vector Machine (OCSVM) also needs to calculate the kernel matrix, the existing OCSVM algorithm introduces a sparse method to find the support vector associated with the hyper-plane. Therefore, for the same size of the dataset, OCSVM computing costs are reduced.

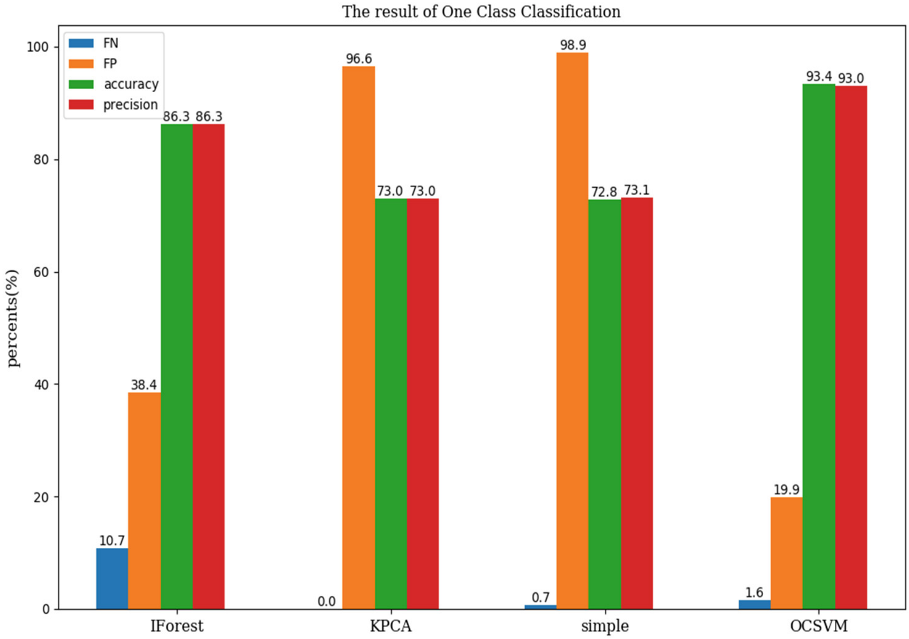

Fig. 5. Comparison of One-Class Classification of water datasets

Moreover, In Figs. 3 and 4, the comparison of the results of the algorithm application of the isolated forest (IForest) was added. The IForest algorithm [8] is used to mine anomaly data, such as attack detection and traffic anomaly analysis in Cyber security, and financial institutions are used to exploit fraud. The algorithm requires very low memory requirements and is fast and its time complexity is linear. It can handle high-dimensional data and large data, and can also be used as online anomaly detection. The algorithm is the most commonly used anomaly detection algorithm. The accuracy of the algorithm is 73.51%, the precision rate is 73.46%, the false negative rate (FP) is 55.62%, and the false positive rate (FN) is 9.49% (see the gas dataset result of IForest method). Compared with the OCSVM algorithm (the precision rate is 98.54%, the accuracy rate is 98.04%, the false negative rate is 2.40% and the false positive rate is 1.60%). The result is worse than the OCSVM algorithm. IForest is the integration of a number of randomly created isolation tree, the role of the tree is to continue to divide the data space, if the data tree where the number of layers, it is considered abnormal. The algorithm is the integration of a number of randomly created isolation tree, the role of the tree is to continue to divide the data space, if the data tree where the number of layers is shallow, and it is considered abnormal. It is also based on a distance measurement of the conversion, but the method does not use all the data, there may be errors; additionally, for industrial control intrusion, most of the structural form and the normal data is very close, the use of isolation tree space may not be accurate enough to determine. However, if you want to conduct online intrusion detection, the use of fast IForest algorithm is also appropriate.

The different results produced by the same algorithm in the gas and water datasets reveal a phenomenon: Different industrial control of the production process is different, even if the same type of attack, its focus is also different. Choosing the right algorithm requires manual testing, comparison and screening.

4 General Conclusion and Future Works

In the Instruction Detection in SCADA Network based on Machine Learning, the use of OCSVM method for abnormal detection can get a high detection rate, effectively limiting the false alarm rate and false negative rate; The limitations of a One-Class method based on a kernel approach lie in how to obtain the appropriate threshold and how to reduce the computational cost; IForest's application effect is not high in OCSVM, but it has real-time and can handle the characteristics of large-scale data, and it is appropriate to use this method in online learning. Using the supervised learning algorithm, especially the decision tree method, you can quickly learn, establish and store the learned intrusion rules, to achieve the purpose of multi-classification. Although the effect of KNN's use of distance metrics is similar to that of decision trees, the computational cost of KNN limits its use.

The intrusion in the industrial control network is different from the Internet network intrusion. The threat of the former focuses on the defects of using industrial control and the defects of industrial hardware, rather than looking for the defects of network communication. It is feasible to use machine learning and data mining to find the relationship between normal and abnormal.

In the course of the experiment, the use of One-Class Classification method for intrusion detection can only detect the existence of abnormal, and cannot find abnormal categories. From the experimental purpose and results analysis, the use of unsupervised learning for intrusion detection has limitations, applying the semi-supervised approach to the intrusion detection model is an improved aspect.

Acknowledgment. This paper uses the dataset for the intrusion detection and evaluation of industrial control systems proposed by the Key Infrastructure Protection Center of Mississippi State University in 2014. The author would like to thank T. Morris to create and share this dataset.

In addition, this work was supported in part by the general project of scientific research of the Education Department of Liaoning Province under grants L2015216 and other foundations under grants FJ1603, 20160092T.

References

1. Knowles, W., Prince, D., Hutchison, D., et al.: A survey of cyber security management in industrial control systems. Int. J. Crit. Infrastruct. Prot. **9**, 52–80 (2015)
2. Shang, W., An, P., Wan, M., et al.: Research and development overview of intrusion detection technology in industrial control system. Appl. Res. Comput. **34**(2), 328–333, 342 (2017)
3. Yang, A., Sun, L., Wang, X., et al.: Intrusion detection techniques for industrial control systems. J. Comput. Res. Dev. **53**(9), 2039–2054 (2016)
4. Bartman, T., Kraft, J.: An introduction to applying network intrusion detection for industrial control systems. In: AISTech 2016, The Iron & Steel Technology Conference and Exposition, 16–19 May 2016
5. Luo, Y., Chen, W.: On a network anomaly detection method based on kernel entropy component analysis and artificial immune. J. Southwest China Normal Univ. (Nat. Sci. Ed.) **41**(6), 119–124 (2016)
6. Wan, M., Shang, W., Zeng, P., Zhao, J.: Modbus/TCP communication control method based on deep function code inspection. Inf. Control **45**(2), 248–256 (2016)
7. Ayres, E., Nkem, J.N., Wall, D.H., et al.: A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. **41**(4), 1690–1700 (2014). Pergamon Press, Inc
8. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation-based anomaly detection. Acm Trans. Knowl. Disc. Data **6**(1), 1–39 (2012)
9. Wu, L., Li, S., Gan, X., et al.: Network anomaly intrusion detection CVM model based on PLS feature extraction. Control Decis. **32**(4), 755–758 (2017)
10. Li, H., Liu, Y.: A new kind of SVM intrusion detection strategy for integration. Comput. Eng. Appl. **48**(4), 87–90 (2012)
11. Wang, H., Yang, Z., Yan, B., Chen, D.: Application of fusion PCA and PSO-SVM method in industrial control intrusion detection. Bull. Sci. Technol. **33**(1), 80–85 (2017)
12. Zhou, Z.H.: Machine Learning. Tsinghua University Press, Beijing (2016)
13. Nader, P.: One-class classification for cyber intrusion detection in industrial systems. IEEE Trans. Industr. Inf. **10**(4), 2308–2317 (2015)
14. Nader, P., Honeine, P., Beauseroy, P.: The role of one-class classification in detecting cyberattacks in critical infrastructures. In: Panayiotou, C.G.G., Ellinas, G., Kyriakides, E., Polycarpou, M.M.M. (eds.) CRITIS 2014. LNCS, vol. 8985, pp. 244–255. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-31664-2_25
15. Shang, W., Li, L., Wan, M., Zeng, P.: Intrusion detection algorithm based on optimized one-class support vector machine for industrial control system. Inf. Control **44**(6), 678–684 (2015)

16. Hoffmann, H.: Kernel PCA for novelty detection. Pattern Recognit. **40**(3), 863–874 (2007)
17. Morris, T., Gao, W.: Industrial control system traffic data sets for intrusion detection research. In: Butts, J., Shenoi, S. (eds.) ICCIP 2014. IFIP Advances in Information and Communication Technology, pp. 66–78. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-45355-1_5
18. Shirazi, S.N., Gouglidis, A., Syeda, K.N., Simpson, S., Mauthe, A., Stephanakis, I.M., Hutchison, D.: Evaluation of anomaly detection techniques for SCADA communication resilience. In: Resilience Week (RWS) 2016, pp. 140–145. IEEE (2016)