Research on Anomaly Detection of Train Communication Network Based on Long and Short-term Memory Network

Jianing Guo1*

1Department of Electrical Information, Shandong University of Science and Technology, Jinan, China
*1157110903@qq.com

Abstract. Aiming at the intrusion detection algorithm based on traditional machine learning, which can not effectively deal with the problems of large quantity, strong timing and high dimension in the train communication network system, this paper proposes an intrusion detection algorithm based on adaptive momentum estimation optimization of long and short-term memory networks. This article first introduces the principles and characteristics of the long and short-term memory network, and then preprocesses the train communication network information according to the characteristics of the train communication network mode, proposes a train communication network anomaly detection model based on the long and short-term memory network, and studies different parameters The impact on the performance of anomaly detection, and the comparison of multiple intrusion detection performance indicators with other traditional shallow machine learning methods, proves the effectiveness of the algorithm proposed in this paper.

1. Introduction

With the development of industrial control system networking and informatization, industrial control systems are increasingly using open software and standardized protocols, making the existing security loopholes of the system continuously exploited, and industrial control systems are facing more and more information security problems[1]. Traditional industrial control security technologies, such as user authentication, firewall and data encryption, have been unable to cope with the continuous innovation of network attacks, intrusion detection technology as the second line of defense of industrial control information security, has become a hot spot for domestic and international research[2-3], intrusion detection technology is to extract the data characteristics reflecting the system behavior, through the design of detection algorithms to classify the attack behavior and normal behavior data[4], various traditional machine learning methods, such as decision trees[5], support vector machines[6-7], neural networks[8], etc. in industrial control intrusion detection have achieved certain results.

Most of the traditional machine learning methods belong to shallow learning cannot effectively deal with the problem of large, time-series and high-dimensional network communication data in train communication network systems. Therefore, this paper proposes a method of using deep learning algorithms instead of traditional machine learning algorithms, using long and short term memory networks based on Adam optimization to learn normal and abnormal communication data, and verifies experimentally the effectiveness of the long The effectiveness of the intrusion detection system for train communication networks based on long and short term memory networks is experimentally verified.
2. Long and short term memory networks based on adaptive momentum estimation optimization

2.1. Short and long term memory networks
Recurrent Neural Network (RNN) is one of the deep learning methods, and its main purpose is to process sequence data. In theory, RNN can effectively process nonlinear time series, but in actual modeling applications of longer time series, there are problems of gradient disappearance and gradient explosion.

LSTM is a good solution to the long and long-term dependence of RNN, and is suitable for the processing of long period timing problems. The chain structure of LSTM is similar to that of RNN. The solution to long-term dependence is to add a gated mechanism to the internal structure and add a layer of memory cells, so that the gradient disappearance and gradient explosion in RNN can be controlled. The LSTM network structure is shown in Figure 1.

![Figure 1. LSTM network structure diagram](image)

2.2. Long and short-term memory network optimized based on adaptive momentum estimation
The essence of anomaly detection is a classification problem. Trains have high requirements for real-time and safety. The two-class classification can meet the requirements of train safety and real-time. LSTM is suitable for solving long-term dependence problems, and can well deal with dynamic system problems involving time series. The construction process of the LSTM-based intrusion detection model proposed in this paper is shown in Figure 2.

![Figure 2. Construction process of intrusion detection model for train communication network based on LSTM](image)
2.2.1. Loss function
The learning of the neural network is actually the process of calculating weights and biases. The loss function is used to measure the degree of difference between the model prediction and the real label. It is the benchmark for the neural network to find the optimal parameters. Its value is a non-negative function, and the value is Smaller indicates the better the model effect. This paper uses the cross-entropy loss function commonly used in classification problems as the loss function of the neural network. The formula is shown in (1).

\[ E = -\sum_k t_k \ln y_k \] (1)

\( E \) represents the cross-entropy error, \( k \) represents the k-th sample, and the label vector \( t_k \) represents the label of the correct solution. One-hot encoding is used, that is, only the value of the k-th dimension is 1, other elements are all 0, and \( y_k \) is the output of the neural network. That is, the model predicts the distribution. It can be seen from the formula that the value of the cross-entropy error is determined by the correct label of all samples and the corresponding predicted label.

2.2.2. Adaptive momentum estimation optimization algorithm
The loss function is complex and the parameter space is huge. The gradient method uses the gradient to find the minimum value or the smallest possible value, and uses the gradient information as the forward direction to continuously reduce the value of the function. In order to speed up the training of the weight matrix, the adaptive momentum estimation (Adam) algorithm is adopted as the gradient descent algorithm in this paper. Adam algorithm takes momentum as the direction of parameter update and can adaptively adjust the learning rate. It occupies less memory and the calculation efficiency of gradient update is higher. The Adam algorithm calculates the exponential weighted average \( M_t \) of the gradient \( g_t \) and the exponential weighted average of the gradient square \( g_t^2 \) as shown in the following formula.

\[ M_t = \beta_1 M_{t-1} + (1 - \beta_1) g_t \] (2)
\[ G_t = \beta_2 G_{t-1} + (1 - \beta_2) g_t^2 \] (3)

Among them, \( M_t \) is the first-order momentum term, which can be regarded as the mean value of the gradient, and \( G_t \) is the second-order momentum term. \( \beta_1 \) is the exponential decay of the first-order moment estimation, and \( \beta_2 \) is the exponential decay of the second-order moment estimation.

When \( M_o = 0 \) and \( G_o = 0 \), the values of \( M_t \) and \( G_t \) at the beginning of the iteration need to be corrected to approximate the true mean and variance. \( M_t^\wedge \) and \( G_t^\wedge \) are the corrected values.

\[ M_t^\wedge = \frac{M_t}{1 - \beta_1^t} \] (4)
\[ G_t^\wedge = \frac{G_t}{1 - \beta_2^t} \] (5)

Finally, the updated values of the Adam algorithm parameters are shown in equation (6).

\[ \Delta \theta_t = -\frac{\alpha}{\sqrt{\hat{G}_t + \epsilon}} \hat{M}_t \] (6)
Among them, $\alpha$ is the learning rate, a value will be set initially, and it can be attenuated during the learning process. $\epsilon$ is a relatively small constant, the purpose is to prevent the denominator from being zero.

3. Train communication network information processing

3.1. Data set construction
This article uses a combination of hardware and software to build a train communication network simulation platform to simulate the train communication network topology and normal operation. First, configure the train communication network to make the train communication network work normally. Second, obtain normal train communication data for a period of time, perform protocol deep analysis and network flow characteristic analysis on data packets to obtain network flow data, and save it to a file, and add the type field: "Normal", which means normal train communication data. Then, launch an attack on the train communication network, capture the train communication data during the attack, perform in-depth protocol analysis and network flow feature analysis on the data packets, and save them to a file, and use 0 or 1 for normal and attack network flow data. Finally, construct a normal data set and an abnormal data set.

3.2. Data stream processing
The detection model proposed above is based on a network stream, and the captured data packet format is a PCAP file, so the PCA file should be merged into a network stream first. Use the IP address and the communication identifier ComId to re-combine the data packets into a stream, with the start of the first data packet as the starting point of the data stream, and the end when there is no data packet communication for more than 10 seconds. For the TCP protocol, the starting point and the end point of the data stream are determined according to the characteristics of the protocol: SYN in TCP_flag is 1 to indicate that a new connection is created, and FIN is 1 to indicate that the connection is terminated.

3.3. Feature selection and pre-processing
According to the characteristics of the train communication network and protocol, and referring to the description of the network flow characteristics in the two general intrusion detection sets KDD99 and CICIDS2017, the statistical characteristics and basic characteristics that can express the network traffic are selected.

For encoding non-numerical features, this article uses one-hot encoding to digitize discrete data. One-hot encoding can diffuse discrete feature values into Euclidean space. The discrete feature value corresponds to a certain point in the space, so that the distance calculation is more reasonable.

The difference in the value of the feature will affect the training speed of the model and the accuracy of the classification. Therefore, it needs to be normalized. The formula is shown in (7).

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$  \hspace{1cm} (7)

In the formula, $x_{\max}$, $x_{\min}$, $x_i$, respectively represent the maximum value, minimum value and attribute characteristic value of the attribute characteristic.

4. Analysis of results
The parameters of the model constructed in this paper are as follows: the number of input layer nodes is the dimension of the training data after processing, the number of hidden layers of LSTM is 2, and the number of output nodes is 2, indicating that the binary classification result is abnormal or normal, and the number of iterations is set to 1000, The optimizer selects the Adam optimizer, the learning rate is 0.01, the initialization method is He_normal, the dropout algorithm is used to prevent
overfitting, the dropout probability is 0.2, the loss function for the binary classification problem is Binary_crossentropy, and the activation function is sigmoid. In order to obtain a model with better performance, the influence of learning rate and the number of neurons in the hidden layer on the performance of the model is studied.

4.1. Impact of learning rate on the performance of intrusion detection models
The learning rate was taken as [0.0001, 0.001, 0.01, 0.1] and the number of hidden layer neurons was taken as 50 for model training to obtain the detection rate and false alarm rate. As shown in Figure 3, as the learning rate increases, the detection rate DR and the false alarm rate FAR are constantly changing. In order to evaluate the performance of the intrusion detection system more accurately, performance=DR/FAR is calculated. As the initial learning rate increases, the performance first increases and then decreases. When it is 0.01, the detection rate is the highest and the false alarm rate is the lowest, and the best performance is obtained. Therefore, the learning rate was taken as 0.01 in the following experiments.

![Figure 3. The impact of learning rate on intrusion detection model and Performance](image)

4.2. Impact of the number of LSTM layer nodes on the performance of intrusion detection models
Based on the results of the previous experiment, the learning rate was taken to be 0.01 to investigate the effect of the number of nodes in the hidden layer K on the performance of the intrusion detection model. Based on practical engineering experience, we set the number of nodes in all LSTM hidden layers to K.

| K  | DR  | FAR  | Performance | ACC  |
|----|-----|------|-------------|------|
| 10 | 0.635 | 0.771 | 0.824       | 0.458 |
| 20 | 0.835 | 0.453 | 1.843       | 0.711 |
| 30 | 0.839 | 0.268 | 3.332       | 0.823 |
| 40 | 0.963 | 0.193 | 4.990       | 0.896 |
| 50 | 0.972 | 0.183 | 5.311       | 0.905 |
| 60 | 0.985 | 0.101 | 9.752       | 0.948 |
| 70 | 0.980 | 0.056 | 17.500      | 0.965 |
| 80 | 0.983 | 0.089 | 11.045      | 0.952 |
| 90 | 0.976 | 0.104 | 9.385       | 0.942 |
The experimental results are shown in Table 1. As $K$ increases, DR gradually increases, and FAR gradually decreases; when $K$ exceeds 60, DR and FAR tend to stabilize. With the increase of hidden layer nodes, the performance of the intrusion detection model improves, but when the number of nodes increases more than 70, the performance of the intrusion detection model begins to decline, that is, when the number of nodes is 70, the detection rate DR is 98.0%, that is, detection The number of normal samples that came out accounted for 98.0% of the total number of normal samples, and the false alarm rate FAR was 0.056, that is, about 5.6% of malicious samples were judged by the model as normal samples, and the accuracy rate was 96.5%, that is, the number of correctly classified samples accounted for the total number of samples. The ratio of 0.965 is 0.965. At this time, the intrusion detection system has the highest performance and the highest accuracy.

4.3. Compared with other machine learning algorithms
According to the above experimental results, we set the best parameters for training the intrusion detection system model: the learning rate is 0.01, the number of nodes in the LSTM layer is 70, and the number of hidden layers is 2. In order to compare the performance of various intrusion detection algorithms under binary classification problems, we use the open source machine learning and data mining software Weka to build support vector machines, Bayesian networks, KNN and SVM and other traditional machine learning models. Train the model through the same training set, and then use the same test set for testing. The experimental results are shown in Table 2: It can be seen from the table that the performance and accuracy of the LSTM intrusion detection model are better than other traditional machine learning algorithms.

| K   | DR(%) | FAR(%) | Performance | ACC(%) |
|-----|-------|--------|-------------|--------|
| Bayesian | 83.48 | 13.4   | 5.74        | 83.37  |
| SVM  | 89.89 | 7.39   | 9.38        | 91.28  |
| KNN  | 81.56 | 21.37  | 4.28        | 78.37  |
| LSTM | 98.03 | 5.58   | 17.57       | 96.47  |

5. Conclusion
This paper designs a train communication network intrusion detection algorithm based on Adam optimized LSTM, and establishes an anomaly detection model by learning the characteristics of normal and abnormal network flows. The data set is constructed to solve the problem that the train network traffic data is difficult to obtain. The algorithm was tested with intrusion detection evaluation indicators such as DR, FAR, and Accuracy to verify the effectiveness of the algorithm, and through comparative experiments, compared with traditional machine learning algorithms, it proved that the algorithm is effective in the detection of abnormalities in the train communication network. Better than traditional machine learning algorithms.

References
[1] J. Yu, E. Lee, S. Oh, Y. Seo and Y. Kim. (2020) A Survey on Security Requirements for WSNs: Focusing on the Characteristics Related to Security. IEEE Access, 8: 45304-45324.
[2] T. Duan et al. (2020) Intelligent Processing of Intrusion Detection Data. IEEE Access, 8: 78330-78342.
[3] X. Gao, C. Shan, C. Hu, Z. Niu and Z. Liu. (2019) An Adaptive Ensemble Machine Learning Model for Intrusion Detection. IEEE Access, 7: 82512-82521.
[4] K. Jiang, W. Wang, A. Wang and H. Wu. (2020) Network Intrusion Detection Combined Hybrid Sampling With Deep Hierarchical Network. IEEE Access, 8: 32464-32476.
[5] I. Delibalta, K. Gokcesu, M. Simsek, L. Baruh and S. S. Kozat. (2016) Online Anomaly Detection With Nested Trees. IEEE Signal Processing Letters, 23: 1867-1871.

[6] I. Ahmad, M. Basher, M. J. Iqbal and A. Rahim. (2018) Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection. IEEE Access, 6: 33789-33795.

[7] P. Tao, Z. Sun and Z. Sun. (2018) An Improved Intrusion Detection Algorithm Based on GA and SVM. IEEE Access, 6: 13624-13631.

[8] H. Yang and F. Wang. (2019) Wireless Network Intrusion Detection Based on Improved Convolutional Neural Network. IEEE Access, 7: 64366-64374.