Classification and Prediction of Network Abnormal Data Based on Machine Learning

Bin Ren[1,2*], Ming Hu[2], Hui Yan[1,2], Ping Yu[1,2]

1. School of Computer Technology and Engineering, Changchun Institute of Technology, Changchun, 130012, China
2. Jilin Province Science and Technology Innovation Center for Physical Simulation and Security of Water resources and Electric Power Engineering, Changchun, 130012, China
*corresponding author's email: renbin_ccit@126.com

***Abstract*—Network abnormal data detection is the main method used for network security situational awareness. In order to reduce the false positives and false negatives of intrusion data, it is proposed to preprocess the data used for prediction by K-means method. In the data prediction, the method of feed forward neural network is adopted, and different loss functions are used to improve the accuracy of prediction and reduce the loss value, and obtain better experimental results.**

Keywords- Machine Learning; Abnormal Data; Neural Network; Classification and Detection

I. INTRODUCTION

In the era when network attacks seriously threaten network security, intrusion data detection is one of the main methods used in network security to proactively discover whether there are violations of security policies and passive attacks. Traditional intrusion data detection methods have serious false positives and false negatives, while machine learning can find hidden and potential information and patterns from a large amount of data. At present, a large number of scholars are studying the relevant content of abnormal data detection. The literature [1] uses DARPA98 as the training set and test set to construct the decision tree classification algorithm; the literature [2] proposes the detection of intrusion data based on unsupervised clustering. The method uses the distance between the samples of the classless training set to generate the clustering model, and then determines the abnormal data according to the normal class ratio; the literature [3] uses the naive Bayes algorithm and the top-down recursive method to propose A new classifier is used for intrusion data detection. The literature [4] proposes a defense strategy model and an improved binary particle swarm optimization algorithm based on the model. The minimum key strategy set of the attack graph is obtained. The simulation experiment proves that the minimum key can be effectively obtained. The optimal solution of the strategy set and its comparison with the ant colony algorithm and the greedy algorithm prove that it is more efficient; the literature [5] proposes a communication-based APT based on the APT communication characteristics extracted from the international security company report. Attack detection method, in order to improve the detection effect of the method, it is also proposed to use the bloom filter to quickly filter and fine-tune the message. The matching two-layer communication feature matching algorithm is combined. The experimental results show that the method has higher detection rate and lower false positive rate. The literature [6] establishes a convolutional neural network model and applies it to network intrusion detection. The convolution kernel and the data are convoluted to extract the local correlation of the features to improve the accuracy of feature extraction.

In this paper, the sample data is verified and classified by K-means method to ensure the correctness of the subsequent data sets before prediction, which lays a foundation for neural network learning. For the data prediction part, this paper adopts multiple hidden layer neural networks, and compares the rules of its loss function to ensure the accuracy of data prediction.

II. RAW DATA SET PREPROCESSING AND CLASSIFICATION

A. Introduction to the raw data set

The KDD99 dataset is the data of the US DARPA98 dataset after preliminary data mining and pre-processing, a total of 5 million, each dataset is described by 41 eigenvalues, where 1-9 is the basic eigenvalue of the connection, used to describe The duration of the connection, the type of protocol, the type of service, etc.; 10-22 is the content eigenvalue of the connection, which is used to describe the data that may reflect the intrusion behavior; 23-31 is the statistic value of the connected time-based network traffic, used to describe Some associated attribute data exists between connections; 32-41 is the connected host-based network traffic statistics feature value used to describe statistical data with the same host. The details are as follows:

2, tcp, smtp, SF, 1684, 363, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 104, 66, 0.63, 0.03, 0.01, 0.00, 0.00, 0.00, 0.00, 0.00, normal.

0, tcp, private, REJ, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 38, 1, 0.00, 0.00, 1.00, 1.00, 0.03, 0.55, 0.00, 208, 1, 0.00, 0.11, 0.18, 0.00, 0.01, 0.00, 0.42, 1.00, portsweep.

The first one is normal data, the second one is abnormal data, and the attack type of abnormal data is 4 categories, 39 subclasses, and the four categories are: denial of service attack (DOS), port monitoring or scanning (PROBING), From a remote unauthorized access host (R2L), a local unauthorized super-access user (U2R). In the KDD99 training data set, the total amount of data is 500,000, including NORMAL 97728, PROBE 4107, DOS 391458, U2R 52, R2L 1126, and a total of 22 categories of attack types.

978-1-7281-2632-6/19/$31.00 ©2019 IEEE
DOI 10.1109/ICRIS.2019.00078

B. Raw data set preprocessing

For later data clustering and prediction, the character eigenvalues in the KDD99 data set need to be digitized, including protocol type, network service type, network connection status, and attack type. In the data mining process, because the evaluation index, dimension and magnitude of the data are different, in order to eliminate the impact of such similarities and differences, data standardization and normalization are usually adopted. There are two common methods for data standardization.

The min-max method linearly converts the raw data, and the converted values are normalized to the range [0, 1]. The specific formula is expressed as:

$$X_{ij}^{'} = \frac{X_{ij} - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

In the above formula: $X_{ij}^{'}$ is the value normalized by X_{ij}, $X_{\min}$ is the minimum value in the sample space, the maximum value in the $X_{\max}$ sample space, and i and j are the subscript quantities.

The values normalized by the Z-score method conform to the normal distribution, that is, the standard deviation has a value of 1, and the mean value is 0. The specific formula is expressed as:

$$X_{ij}^{'} = \frac{\left(X_{ij} - \mu\right)}{\sigma} \tag{2}$$

In the above formula: $X_{ij}^{'}$ is the value normalized by X_{ij}, μ is the data mean of the sample space, and σ is the data standard deviation of the sample space.

C. Raw data set classification

The K-means method is one of the clustering methods for dividing objects similar to each other into multiple categories. The main basis of the division is the vector distance, and the distance formula usually adopts the Euclidean distance. The formula for solving the two-point distance in two-dimensional space by Euclidean distance is as follows:

$$c = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3}$$

In the above formula: x and y represent spatial coordinate points, respectively.

The algorithm steps of the K-means method are as follows:

Step 1: k data is selected from x data objects as an initial vector of a cluster center.

Step 2: calculating the distance value of each data from the k initial vectors using the Euclidean distance according to the k initial vectors determined in step (1).

Step 3: According to the calculation result of step (2), each data and its nearest vector are grouped into the same cluster.

Step 4: Recalculating the center vector of each cluster.

Step 5: Repeat steps (3) and (4) until the vector categorization of each cluster is less than 1%.

In order to speed up the experimental processing speed, this paper uses the method of [7] to remove the useless dimensions of clustering, which reduces the complexity of the algorithm, speeds up the clustering speed, improves the clustering effect, and the experimental results of the original data. As shown in Fig. 1:

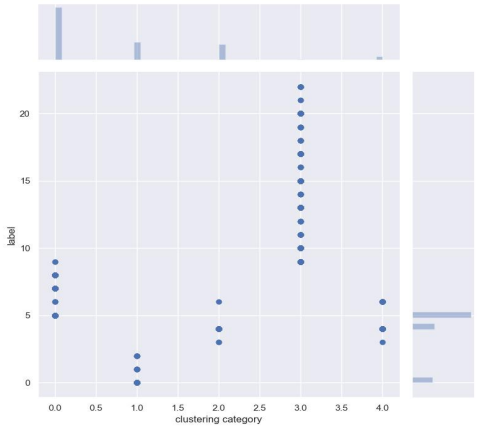

Figure 1. Scatter plot and histogram based on K-means clustering results

In the Fig. 1, the clusters are of 5 categories, which are 0, 1, 2, 3, and 4, respectively corresponding to DOS, NORMAL, PROBE, U2R, and R2L; their histograms and scatter plots show the quantitative trend, just like the original The trend of the number of categories has been constant The correctness of the method used in this paper is verified, which lays a foundation for the subsequent classification and prediction of newly collected data.

III. DATA PREDICTION METHOD AND EXPERIMENT BASED ON FEEDFORWARD NEURAL NETWORK

A. Data prediction method for feed forward neural network

In this paper, the feed forward neural network is adopted. The main advantage of the neural network is that it can solve the linear inseparable problem well. Each neuron is a linear classifier. The neural network is mainly composed of a neuron function, a loss function, an excitation function, and the like. For an n-dimensional vector, the input expression of its neuron is f (x_1, x_2, …, x_n), the output function is $f(x)=wx+b$, w is the weight, x is the input vector, wx is The inner product operation of two matrices, b is a real value. The loss function represents the difference between the fit and the true value. In order to obtain the appropriate w and b values, the loss function value is made as small as possible. The expression of the loss function is:

$$Loss = \sum_{i=1}^{n} \left| wx_i + b - y_i \right| \tag{4}$$

The excitation function is located after the neuron output function, and its main purpose is to add nonlinear factors to the output. At present, the main excitation functions are Sigmoid function and ReLU function. The Sigmoid function

finally projects the value of the output function to 0 and 1 values, where 1 means fully activated, 0 means completely inactive, and other values are between Between the two, its function expression is defined as:

$$z = wx + b, \quad f(z) = \frac{1}{1+e^{-z}} \tag{5}$$

The ReLU function is used for convolutional neural networks.

The neural network structure of this paper is shown in Fig. 2:

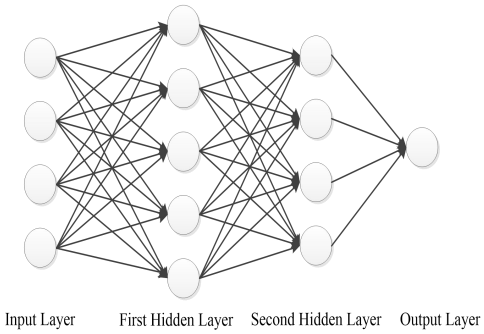

Figure 2. Multilayer neural network structure

The expressions of the hidden layer and the output layer in the structure are as follows:

$$z_h = w_h x + b_h, \quad y_h = \frac{1}{1+e^{-z_h}} \tag{6}$$

$$z_o = w_o y_h + b_o, \quad y_o = \frac{1}{1+e^{-z_o}} \tag{7}$$

Determine the values of the weights w and b according to the loss function [8]:

$$(w_h)^n = (w_h)^{n-1} - \eta \frac{\partial Loss}{\partial w_h} \tag{8}$$

$$(b_h)^n = (b_h)^{n-1} - \eta \frac{\partial Loss}{\partial b_h} \tag{9}$$

$$(w_o)^n = (w_o)^{n-1} - \eta \frac{\partial Loss}{\partial w_o} \tag{10}$$

$$(b_o)^n = (b_o)^{n-1} - \eta \frac{\partial Loss}{\partial b_o} \tag{11}$$

The above formulas η indicates the learning rate. The loss function uses two methods [9], which are the L2 regular loss function and the cross entropy loss function. The L2 regular loss function is the sum of the squares of the difference between the predicted value and the target value. The loss function has a better curvature performance near the target value, and the closer to the target, the slower the convergence. The cross entropy loss function is used as a logical loss function. When the predicted value is closer to 1, the loss function value is smaller.

B. Analysis of results

The operating system used in the experiment is WINDOWS 7, the data set is KDD99, the operating environment is PYTHON 3, and the library used for data prediction is TensorFlow. In this experiment, different loss functions were used, and different experiments were carried out to obtain the loss value and accuracy value. As shown in Figs. 3～6:

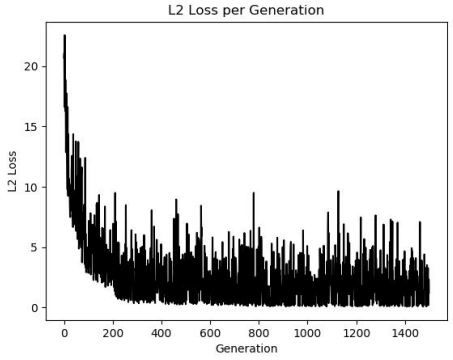

Figure 3. L2 regular loss function loss value

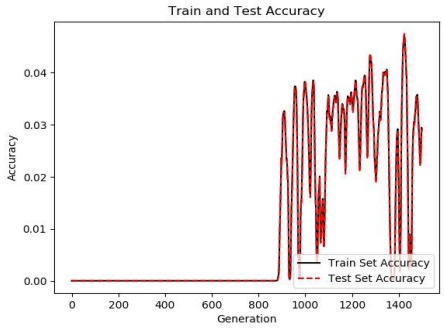

Figure 4. Test and prediction accuracy based on L2 regular loss function

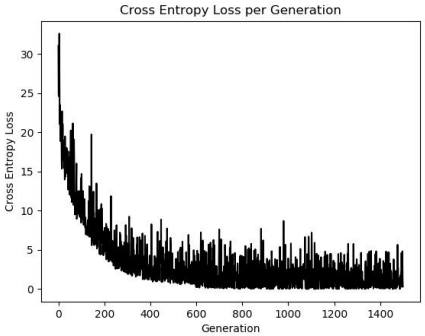

Figure 5. Loss value of the cross entropy loss function

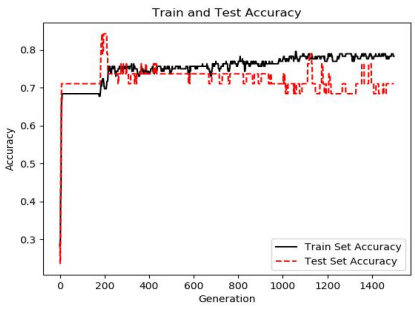

Figure 6. Test and prediction accuracy based on cross entropy loss function

It can be seen from the above figure that the loss value of the L2 regular loss function and the cross entropy loss function shows the relationship between the number of iterations and the jitter. As the number of iterations increases, the value of the regular loss decreases, and as the number of iterations increases, the jitter of the function It is also decreasing, except that the loss value of the cross entropy loss function has a range and degree of jitter that is better than the former.

As can be seen from the above figure, if the accuracy of the training data set is greater and the accuracy of the test data set is reduced, then the fit is over-fitting; if the accuracy of the test data set and the training data set are always increasing, then this fit is an under-fitting and requires continued training. In this experiment, the neural network algorithm is adopted, the convergence speed of the model is faster and more accurate, the accuracy of training continues to increase, and the cross entropy loss function shows that the accuracy of the test set sometimes increases slightly and sometimes decreases, while L2 The regular loss function fluctuates too much.

IV. CONCLUSIONS

Firstly, the K-means method is used to preprocess the data used for prediction, which reduces the prediction error rate of subsequent data. In the data prediction, the feed forward neural network method is adopted, and different loss functions are used to improve the accuracy of prediction and reduce the loss, and good experimental results are obtained.

REFERENCES

[1] Lee, et al. "Effective Value of Decision Tree with KDD 99 Intrusion Detection Datasets for Intrusion Detection System." International Conference on Advanced Communication Technology IEEE, 2008.

[2] LUO Min, WANG Lina and ZHANG Huaguo. "An Unsupervised Clustering-Based Intrusion Detection Method." Acta Electronica Sinica, 2003, vol.31, no.11, pp.1713-1716.

[3] Dalvi, et al. "Adversarial classification." Tenth Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, 2004.

[4] LIU Yuan, LI Qun and WANG Xiaofeng. " Improved PSO for network defense measures of weighted attack graph . " Computer Engineering and Applications, 2016, vol.52, no.8, pp.120-124.

[5] DAI Zhen, CHENG Guang. " Advanced persistent threat detection based on characteristics of communications." Computer Engineering and Applications, 2017, vol.53, no.18, pp.77-83.

[6] JIA Fan, KONG Lingzhi. "Intrusion Detection Algorithm Based on Convolutional Neural Network. " Transactions of Beijing Institute of Technology, 2017, vol.37, no.12, pp.1271-1275.

[7] JIA Fan, YAN Yan and ZHANG Jiaqi. "K-means based feature reduction for network anomaly detection. " J Tsinghua Univ, 2018, vol.58, no.2, pp.137-142.

[8] GAO Yang, WEI Zheng and WANG Juan. "Deep learning and TensorFlow. " 2017, pp.20-30.

[9] Gupta S, et al. "Deep Learning with Limited Numerical Precision." Computer Science, 2015, pp.1-10.