Research on Traffic Anomaly Detection Based on SDN Architecture

Xiangxiao Chen, Xin Cui*, Kangtao Wang, Qin Du
College of Computer Science and Technology, Shandong University of Technology, Zibo 255000, Shandong, China
DOI: 10.32629/aes.v4i1.1162

Abstract: With the development of science and technology, the network is also developing rapidly. The new software-defined network SDN (Software Defined Network) solves the defects of the traditional network architecture. It is regarded as the development direction of the future network, and its main feature is to separate and decouple the two modules of data forwarding and routing control of traditional network equipment. The SDN controller uses the standardized OpenFlow interface protocol to manage and configure devices from various manufacturers, improving the network horizontal expansion capability and optimizing the underlying infrastructure resources, and expanding the SDN network elasticity function. But the controllers are vulnerable to attacks that render the entire network inoperable. In order to solve this problem, machine learning abnormal traffic detection is proposed to solve the defect of low network abnormal traffic detection rate. Due to the large amount of data and high data latitude, several factors affect the efficiency and accuracy of machine learning. Therefore, it is necessary to reduce the dimensionality of data to improve the efficiency and accuracy of machine learning. S - FastICA (Fast Independent Component Analysis) is introduced. Fast independent component analysis dimensionality reduction algorithm, which uses a fixed-point iterative optimization algorithm, makes the convergence faster and more robust. Due to the low accuracy of the traditional Stacking model, the EIE-Stacking (E n semble in Ensemble Stacking) model is used to improve the base learner of the first layer. The improved model effectively improves the prediction accuracy. In order to verify the authenticity and effectiveness of the experiment, the KDDCUP99 data set, NSLKDD data set and UNSW-NB15 data set were used for experiments, and the S-FastICA (Softsign FastICA) algorithm was compared with the traditional FastICA algorithm. The machine learning model EIE-Stacking Compared with the traditional Stacking model, the experimental results show that the accuracy rate, F1 score, recall rate and precision rate are all improved, and the algorithm proposed in this paper is true and effective.
***Keywords*:** machine learning, network architecture, SDN, FastICA, ensemble learning, abnormal traffic detection

1. Overview

The software-defined network (SDN) is regarded as the development direction of the future network, and has been widely concerned by scholars since its inception. At present, SDN is used in many fields, such as data network centers, large-scale enterprise networks, campus networks, telecom operation networks, and Internet company business deployment, etc. The development prospects at home and abroad present a bright trend.

In recent years, scholars at home and abroad have successively discussed and studied the network architecture. Some progress has been made in SDN, followed by some new SDN network security issues.

Although SDN solves the problems of closed mode and many protocols in the traditional network architecture, it still has not given a reasonable solution to the problems of many network data characteristics, large data sets and high data dimensions. degree of research. In order to solve the problems of large quantity and high data dimension, in 2021, Chen and Haonan aimed at the low detection rate of the current IDS, applied the FastICA headline news algorithm to the intrusion detection system, and combined FastICA with the SVM algorithm [2] The experiment showed that the detection effect is obvious Higher than the PCA algorithm, made a little progress, and successfully brought FastICA [3] to the field of intrusion detection. In order to improve the accuracy of machine learning, integrated learning models such as Bagging model and Boosting model, and Stacking integrated learning model are proposed. Random forest and Adaboosting model are typical representatives of Bagging model and Boosting model respectively. This paper uses the Stacking ensemble learning method to improve the prediction accuracy by improving the base learner (weak learner) into an ensemble learning model, and has achieved good results through experimental verification. In order to further ensure the security of SDN network, a brand-new method is found, which is to reduce the dimension of feature value, so that machine learning can more accurately identify the characteristics of data traffic from the perspective of data, so as to give the most correct prediction value.

In this paper, the abnormal traffic detection module of machine learning is applied to the SDN controller to overcome the shortcomings of traditional complex networks that lack unified management; the low efficiency of traditional detection

modules is solved by machine learning; the S-FastICA+EIE-Stacking model is used to improve the machine The advantage of learned detection accuracy.

2. SDN-related technical theories

As the SDN technology is not yet fully mature, SDN security issues have attracted widespread attention from scholars. This paper mainly describes the research on abnormal traffic detection under the SDN architecture, that is, it proposes to increase the abnormal traffic detection module to ensure the security of SDN network. The abnormal traffic detection module uses machine learning technology to establish a model through machine learning open source standard data sets, predicts the abnormal traffic of the network through the model, and sends the correct strategy through the SDN controller, that is, the normal traffic passes, and blocks the abnormal traffic., Discard data packets to ensure the safe operation of the SDN network.

This paper focuses on how to improve the detection accuracy of machine learning models for abnormal network traffic. First, we improve the prediction accuracy of the model by improving the FastICA dimensionality reduction algorithm and processing the eigenvalues of network traffic. Second, by improving the machine learning Stacking model, the improved EIE-Stacking model enables machine learning to have higher accuracy and effectively identify abnormal traffic from multiple perspectives. Let the traffic anomaly detection module more accurately identify the abnormal traffic of the SDN network to ensure the security of the SDN network. The overall flowchart of SDN is shown in Figure 1.

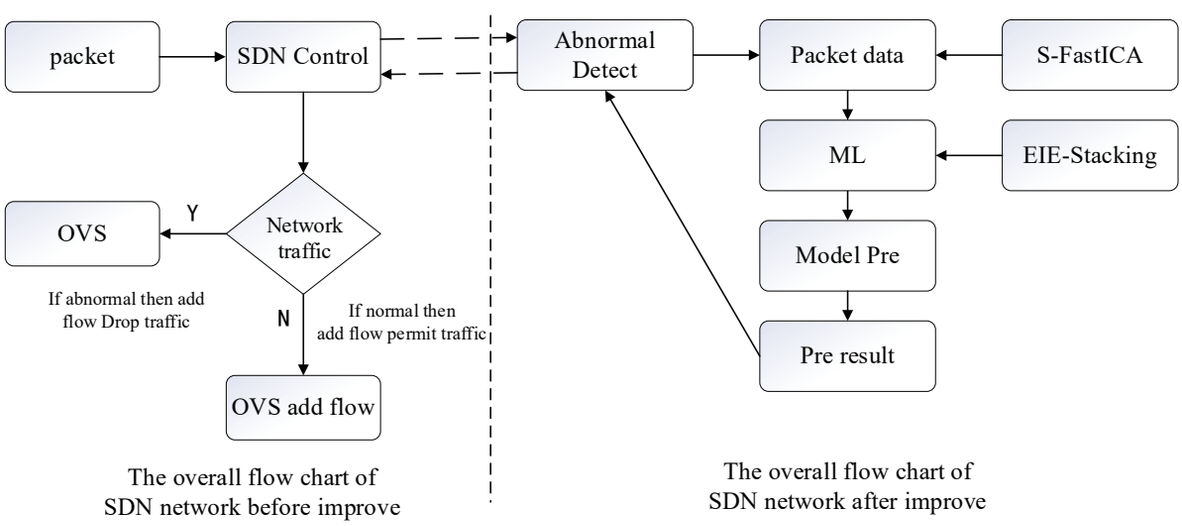

Figure 1. Overall of SDN flow chart

This paper provides a secure defense for the SDN architecture network by introducing a detection module. How to make the detection module quickly and accurately identify abnormal traffic is the focus of this paper. This paper conducts machine learning on SDN open source data sets, uses data feature values in network traffic and machine learning algorithm models to establish a machine learning model, and conducts innovative research on improving the prediction accuracy of the model. First of all, the SDN network traffic data set has many eigenvalues and large dimensions, which leads to low machine learning accuracy. This paper uses the dimensionality reduction algorithm FastICA to improve and innovate to reduce the feature dimension, so that machine learning can accurately predict network traffic. Secondly, use the integrated learning algorithm to stack Stacking machine learning models, combine other machine models, and combine different types of machine learning models into an integrated learner, thereby improving the accuracy of machine learning predictions.

3. Improved FastICA algorithm

Due to the fact that abnormal traffic attack packets cannot be completely detected, in order to improve the detection accuracy of the abnormal traffic module. We use the FsatICA dimensionality reduction algorithm to reduce the dimension of feature values by removing invalid and redundant information without reducing the amount of effective information to reduce the amount of machine learning calculations, reduce complexity, and improve accuracy, thereby effectively ensuring SDN network security. In this paper, the nonlinear function $g_{(x)}$ of the FastICA dimensionality reduction algorithm is optimized. The optimized $g_{(x)}$ function effectively separates the matrix w more accurately and faster, accelerates convergence, and separates independent variables. The flow chart of the separation matrix w is shown in Figure 2.

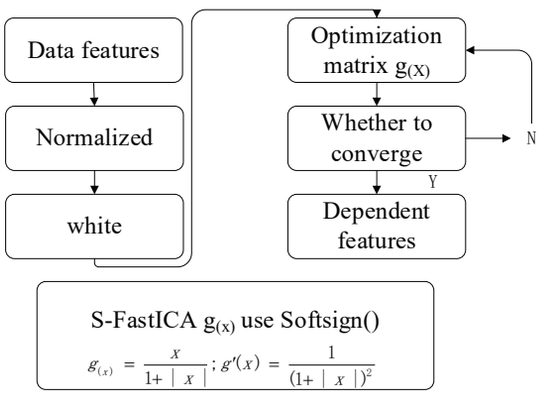

Figure 2. S-FastICA flow chart diagram

The implementation process of the algorithm:

(1) Form the original data set into an n-row, m-column matrix at $X_{n\times m}$.

(2) Zero-average each row of X (representing a feature), i.e., subtracting the mean of this row.

(3) Pre-process the data for whitening.

(4) Set the value of the parameter learning rate a.

(5) Solve for w at moment i, where the initial w can be assigned as a random matrix with the sum of each row being 1, and w is the unmixing matrix that separates the independent variables.

(6) The eigenvalue of column i — $s_{n\times 1}^{(i)}$ will be solved based on w and the formula $s_{n\times 1}^{(i)} = W_{n\times n} \cdot x_{n\times 1}^{(i)}$ obtained in the previous step.

(7) Repeat Step 4 and Step 5, and the eigenvalues $s_{n\times 1}^{(i)}$ of all columns are solved.

(8) The eigenvalues obtained at each moment are combined to obtain the final result $S_{n\times m} = [s^{(1)}s^{(2)} \cdots s^{(m)}]$, where $s^{(m)}$ is the mth independent component separated from $s_{n\times m}$.

The improved S-FastICA algorithm in this thesis is changed from the nonlinear cumulative distribution function to the Softsign() function from Logcosh(). The purpose of improving the nonlinear cumulative distribution function is to speed up the convergence of w unmixing matrix. Solve for independent components by w.

S-FastICA's improvements to FastICA include

The cumulative distribution function is changed from $g_{(x)} = \frac{1}{a} log(coshax)$ to $g_{(x)} = \frac{x}{1+|x|}$

The corresponding probability distribution function is changed to

$$g'(x) = \frac{1}{(1+|x|)^2}$$

Detailed description of the iterative formula of w:

The negative entropy approximation of $w^T x$ can be obtained from the best-fit condition of $E\{(w^T x)\}$. According to the Kuhn-Tucker condition, under the constraints of $E\{(w^T x)^2\} = \|w\|^2 = 1$, $E\{(w^T x)^2\} - \beta w = 0$.

Defining the left side of the equation as F, we obtain the Jacobi matrix $JF_{(w)}$ as $JF_{(w)} = E\{xx^T g'(w^T x)\} - \beta I$.

It is then transformed into a problem of roots of an equation, which we can solve by Newton's method: $x_{n+1} = x_n - \frac{f_{(x_n)}}{f'_{(x_n)}}$.

That is, $w_{n+1} = w_n - \frac{E\{(w^T x)^2\} - bw}{E\{xx^T g'(w^T x)\} - b}$

Multiply both sides by $\beta - E\{g'(w^T x)\}$, it is further simplified to $w_{n+1} = E\{xg(w_n^T x)\} - E\{g'(w_n^T x)\} \cdot w$, where $w = [w_1 w_2 \ldots w_m]$

In summary, the basic form of the one-time FastICA algorithm is as follows.

(a) Initialize the vector w;

(b) Let $W_{n+1} = E\{xg(w_n^T x)\} - E\{g'(w_n^T x)\} \cdot w$

(c) Let $W_n = \frac{W_{n+1}}{||W_{n+1}||}$

(d) If it does not converge, the steps (b) and (c) are repeated until w converges. By convergence, we mean that w is in the same direction before and after, i.e., their dot product is 1. One FastICA algorithm can estimate one independent component, and it needs m independent components, then it needs to do FastICA algorithm m times. The most obtained to $w = [w_1, w_2, \ldots w_m]$

Finally, when w is decomposed, all independent variables are separated out according to the formula $s_{n\times1}^{(i)} = W_{n\times n} \cdot x_{n\times1}^{(i)}$.

The Softsign() function is another alternative to the Logcosh() function. Like Tanh(), Softsign() is antisymmetric, decentralized, differentiable, and returns a value between -1 and 1. Its flatter curve and slower descending derivative suggest that it can learn more efficiently. At the same time, the derivative of Softsign() is a probability distribution function.

4. Stacking model improved EIE - Stacking model

Due to the low prediction accuracy of traditional machine learning models, in order to solve this problem in recent years, ensemble learning models have been introduced, such as the well-known Bagging ensemble learning and Boosting ensemble learning in the industry.

What this article introduces is that Stacking integrated learning can combine weak learners of different nature. learner to improve the overall prediction accuracy. The schematic diagram of the EIE-Stacking model is shown in Figure 3:

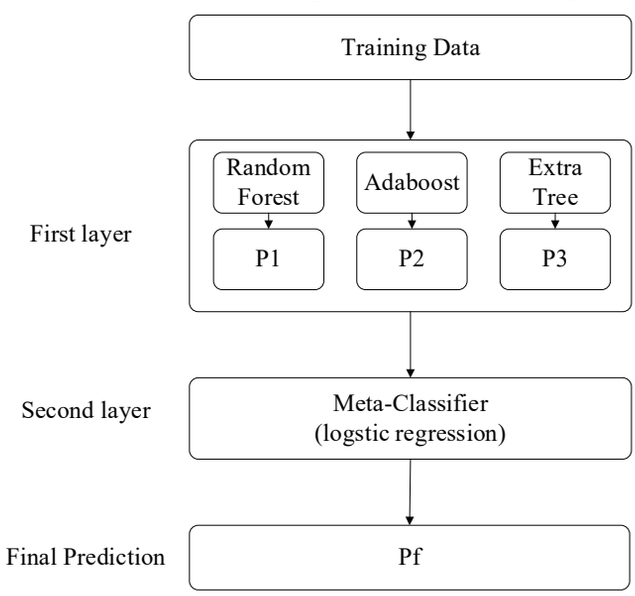

Figure 3. EIE-Stacking mode of machine learning

The principle of the improved machine learning model uses the ability of Stacking to combine different types of machine learning, and can improve the base learner to make the base learner an integrated learner, that is, the Stacking integrated learning model nested Bagging set learning model, Boosting integrated learning model The method, that is, the improved Stacking model combines three different integrated learning models of random forest, Adaboosting, and extreme forest to form a new machine learning model. The EIE-Stacking model has the characteristics of various models, and predicts abnormal traffic from different angles, which greatly improves the generalization ability of the model, and finally improves the prediction accuracy of the model as a whole.

5. Experiment and result analysis

In order to verify the effectiveness of the abnormal traffic detection method based on the SDN architecture proposed in this paper, a comparative experiment was carried out. The experimental environment is Scikit-learn 1.0.2, pandas 1.4.2, numpy 1.21.5, Mlxtend 0.21.0

In this paper, different data sets KDDCUP99, NSLKDD, UNSW-NB15 are used for experiments.

In order to verify the validity of the experimental results of abnormal network traffic detection and compare the detection performance of different models, it is necessary to evaluate the detection output value and the real value of the model. This paper selects the following performance evaluation criteria: accuracy rate (Accuracy), F1 score (F1_score), The recall rate (Recall) and the precision rate (Precision) evaluate the traffic anomaly detection model from different angles. Each evaluation function variable is explained as follows: TP means correctly predicting positive samples as positive; FN means incorrectly predicting positive samples as negative; FP means incorrectly predicting negative samples as positive; TN means correctly predicting negative samples as negative;

The evaluation formulas are as follows:

The accuracy rate is shown in formula (1):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The recall rate is shown in formula (2):

$$Recall = \frac{TP}{TP + TN} \tag{2}$$

The accuracy rate is shown in formula (3):

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

F1 is the harmonic mean of the precision rate and the recall rate. On the interval (0, 1), the higher the value, the better. The formula is shown in (4):

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

The experimental results data are compared with different algorithms and machine learning models on different data sets, among which Stacking integrates NB and KNN two major learning models. In Table 1 to Table 4, the four key performance indicators of AC, F1, Recall, and Precision are compared and analyzed. As shown in Table 1, Table 2, and Table 3.

Table 1. Metric of different algorithms in abnormal traffic detection based on Dataset KDDCUP99

Algorithm	AC	F1	Recall	Precision
ICA-MLP[5]	99.88%	N/A%	N/A	99.9312%
NB	95.62%	97.32%	99.28%	95.44%
KNN	99.972%	99.983%	99.982%	99.983%
Stacking	99.979%	99.987%	99.98%	99.99%
FastICA + NB	98.51%	99.06%	98.15%	100%
FastICA + Stacking	98.51%	99.06%	98.15%	100%
S-FastICA + NB	99.991%	99.994%	99.992%	99.9966%
S-FastICA + Stacking	99.994%	99.996%	99.994%	99.998%
S-FastICA + EIE-Stacking	99.997%	99.998%	99.996%	100%

The data set used in Table 1 is KDDCUP99, which is currently a more commonly used SDN network traffic detection experimental data set. The KDDCUP99 network abnormal traffic detection data set is used. When FastICA reduces the dimensionality to 21 eigenvalues, it can be seen from the performance indicators in Table 1 that the improved S-FastICA algorithm combines NB, KNN and Stacking for performance comparison, and using FastICA for dimensionality reduction is better than using no dimensionality reduction, the performance is about 5% higher; using S-FastICA dimensionality reduction is about 1.4% higher than FastICA dimensionality reduction. It shows that the S-FastICA dimensionality reduction algorithm effectively improves the performance; the performance of S-FastICA + Stacking is the best, and the prediction accuracy rate reaches 99.997%. The same improved S-FastICA+Stacking performance index is higher than the literature [5]. It is concluded that the performance of the improved S-FastICA + EIE -Stacking algorithm is better than other algorithms

on the data set KDDCUP99.

Table 2. Metric of different algorithms in abnormal traffic detection based on Dataset NSLKDD

Algorithm	AC	F1	Recall	Precision_
Xgboost[6]	98.7%	98.76%	99.11%	98.41
MKMC4[7]	97.66	98.34%	99.04%	96.97
FastICA + Stacking	98.621%	9 8.56%	98.41%	98.71%
FastICA+EIE-Stacking	99.14%	99.11%	98.87%	99.34
S-FastICA+Stacking	98.617%	98.56%	98.41%	98.71%
S-FastICA+EIE-Stacking	99.18%	99.14%	98.88%	99.41%

The data set used in Table 2 is NSL-KDD. From the performance indicators in Table 2, it can be seen that the performance index AC of the S-FastICA + EIE-Stacking algorithm is significantly higher than that of the algorithm before improvement, and is higher than the performance of literature [6,7]. It is concluded that the improved S-FastICA+EIE-Stacking algorithm in this paper has better performance than other algorithms in the data set NSL-KDD.

Table 3. Metric of different algorithms based on Dataset UNSW- NB15

Algorithm	AC	F1	Recall	Precision_
Bi-Directional LSTM[8]	99.67%	N/A	N/A	N/A
FastICA + Stacking	99.43%	98.82%	98.72%	98.91%
FastICA+ EIE-Stacking	99.65%	99.28%	98.95%	99.61%
S-FastICA + Stacing	99.50%	98.96	98.83%	99.09%
S-FastICA + EIE-Stacking	99.704%	99.37%	98.99%	99.76%

Table 3 is UNSW-NB15. The innovation point of this paper is S-FastICA combined with EIE-Stacking. The higher the value of the quality evaluation index, the better the detection effect. It can be seen that the performance index of S-FastICA combined with EIE-Stacking model is the best. Other related performance has also been improved. The performance is higher than that of BI-Directional LSTM literature [8]. Finally, it is concluded that the performance indicators of the innovative algorithm combined with the innovative model used in this paper are superior to those of the comparative literature, and have high application value.

6. Conclusion

This paper focuses on the research on abnormal traffic detection under the SDN architecture. In order to ensure the security of the SDN network, it predicts whether the traffic in the SDN network is normal or abnormal by improving the prediction accuracy of machine learning. The SDN controller issues policies based on the prediction results of the traffic detection module. This ensures the safe operation of the SDN network. In order to improve the prediction accuracy of the detection module, first: we use the improved S-FastICA algorithm to improve the prediction accuracy. Second: We combine other integrated learning models by innovatively using the characteristics of the Stacking model to form a nested two-layer model of integrated learning, effectively training and learning data features from different angles and dimensions, and innovatively using powerful EIE-Stacking model, greatly improving the prediction accuracy. Use the detection module to make the most accurate judgment, and send it to the SDN controller, and the SDN controller will make corresponding correct policies on the data packets in time, so as to protect the SDN network from abnormal traffic attacks, intrusion attacks, and correctly handle the SDN network traffic to protect the safe operation of the SDN network. Therefore, the research on abnormal traffic detection based on SDN network architecture designed in this paper can effectively detect network intrusion behavior. The focus of the next research is to further optimize the allocation of SDN controller resources under the special circumstances of insufficient or excess SDN controller resources.

References

[1] Zhang Y, Cui L, Wang W, et al. A survey on software defined networkingwith multiple controllers[J]. *Journal of Network and Computer Applications*, 1 February, 2018, 103: 101-118.

[2] Chen, Haonan, et al. "Research on Intrusion Detection of Industrial Control System Based on FastICA-SVM Method." International Conference on Artificial Intelligence and Security. Springer, Cham, 2021, 12736(7): 303–311
[3] Sheikh M S, Regan A. A complex network analysis approach for estimation and detection of traffic incidents based on independent component analysis[J]. *Physica A: Statistical Mechanics and its Applications*, 2022, 586: 126504
[4] Rashid M, Kamruzzaman J, Imam T, et al. A tree-based stacking ensemble technique with feature selection for network intrusion detection[J]. *Applied Intelligence*, 2022, 52(9):9768-9781.
[5] Liu Jinghao, Mao Siping, Fu Xiaomei. Intrusion Detection Model Based on ICA Algorithm and Deep Neural Network [J]. *Information Network Security*, 2019, 0(3): 1-10.
[6] Dhaliwal, SS; Nahid, A.-A.; Abbas, R. Effective Intrusion Detection System Using XGBoost. Information 2018, 9(7), 149-172. https://doi.org/10.3390/info9070149
[7] Flow Based Intrusion Detection System for Software Defined Networking using Hybrid Machine Learning Technique [J]. *International Journal of Innovative Technology and Exploring Engineering*, 2019, 9(2S2) 1026-1033. https://doi.org/10.35940/ijitee.B1108.1292S219
[8] Pooja TS, Shrinivasacharya P. Evaluating neural networks using Bi-Directional LSTM for network IDS (intrusion detection systems) in cyber security[J]. *Global Journal of Transformation*, 2021, 2(2): 448-454.