Data Balanced Algorithm Based on Generative Adversarial Network

I-Hsien Liu, Cheng-En Hsieh, Wei-Min Lin, Jung-Shian Li*
Department of Electrical Engineering / Institute of Computer and Communication Engineering,
National Cheng Kung University
No.1, University Rd., East Dist., Tainan City 701401, Taiwan

Chu-Fen Li
Department of Finance, National Formosa University
No.64, Wunhua Rd., Huwei Township, Yunlin County 632301, Taiwan
E-mail: {ihliu, cehsieh, wmlin}@cans.ee.ncku.edu.tw, jsli@mail.ncku.edu.tw, chufenli@gmail.com*
www.ncku.edu.tw, www.nfu.edu.tw

Abstract

In order to defend against malicious attacks, intrusion detection systems have introduced machine learning as a protection strategy. However, machine learning algorithms and datasets have a great influence on the effectiveness of the machine learning model. This study uses five algorithms which are Naïve Bayes, CNN, LSTM, BAT, and SVM to train the IDS machine learning model. We design a data-balanced method based on the GAN algorithm to improve the data imbalance problem of the IDS dataset.

Keywords: Anomaly Traffic Detection, Machine Learning, IDS Dataset, GAN, Performance Analytics

1. Introduction

As more data is transmitted on the Internet, the greater the risk of malicious network attacks, and even national security issues may arise. With the increasing popularity of artificial intelligence and the maturity of big data collection technology, machine learning has also been applied to intrusion detection systems. How to convert the huge amount of network traffic data into clean and highly recognizable machine learning datasets is a very important issue. There are many types of public IDS datasets. Some datasets are not compatible with current network attacks because of outdated attack methods and insufficient diversity. The content is not suitable for the research of modern artificial intelligence intrusion detection systems. Our research selects three representative IDS datasets as research objects based on attack diversity and data integrity, namely NSL-KDD [1], UNSW-NB15 [2] and CICIDS 2017 [3]. At the same time, we found through sensitivity analysis that traffic datasets with too large a gap between malicious traffic data and normal traffic data performed poorly. In order to overcome this problem, this research designed a data-balanced algorithm based on a generative adversarial network for the data distribution of the dataset, called GAN-BAL. The GAN-BAL algorithm is used to generate traffic datasets for training and evaluation. This dataset satisfies the diversity of attacks and data integrity and retains the heterogeneity of the data, which can achieve better results in model training.

2. Related Work

At present, most intrusion detection system researchers use artificial intelligence to defend against constantly changing and evolving malicious attack methods. The below explains common machine learning algorithms used in malicious traffic detection.

2.1. *Naïve Bayes classifier (NB)*

Naïve Bayes classifier (Naïve Bayes, NB) is a simple method for constructing classifiers. The theoretical basis is to assume that each feature is conditionally independent. According to Bayes' theorem, we can classify conditionally independent features easily. The Naïve Bayes classifier will calculate each data the conditional probability of each category, and then use the maximum a posteriori (MAP) estimation to determine the best classification method.

2.2. *Support vector machine (SVM)*

The support vector machine originated from the algorithm proposed by Vapnic et al. based on statistical mathematics in 1963 and it is designed to solve the problems related to regression analysis and statistical classification [4]. SVM is better than many traditional machine learning methods in classifying nonlinear and high-dimensional data. The core concept of the support vector machine is to classify data by finding a hyperplane. SVM searches for the closest data point to the hyperplane. The distance between the hyperplane and the point is called the support vector.

2.3. *Convolution neural network (CNN)*

Convolutional neural network is a combination of the convolutional layer and deep neural network. Convolutional neural network uses the convolutional layer and the pooling layer to achieve a method that can reduce data without losing too much information. Convolutional neural networks have achieved good results in image processing and speech analysis.

2.4. *Long short-term memory (LSTM)*

The short-term memory model is an improved model of the recurrent neural network. Hochreiter et al. published the paper for the first time in 1999 [5], which mainly improved the defects of the recurrent neural network for time series. LSTM is composed of four units: Input Gate), Output Gate, Forget Gate, and Memory Cell. It is mainly used to alleviate the problem of the disappearance of the gradient. The memory unit will record the data of the last state, and use 4 valves to determine whether the input or output data needs to be stored or output.

2.5. *BAT*

The BAT algorithm is a deep learning method proposed by Su et al. for network intrusion detection systems [6]. This algorithm is designed based on a two-way long and short-term memory model and attention mechanism and is used in the data processing. The convolutional layer is used for processing, and the two-way long and short-term memory model is used to learn the characteristics of each flow and obtain the vector corresponding to each flow. Then use the attention mechanism to perform feature learning on the sequence data composed of traffic vectors to obtain subtle features to achieve a better classification effect.

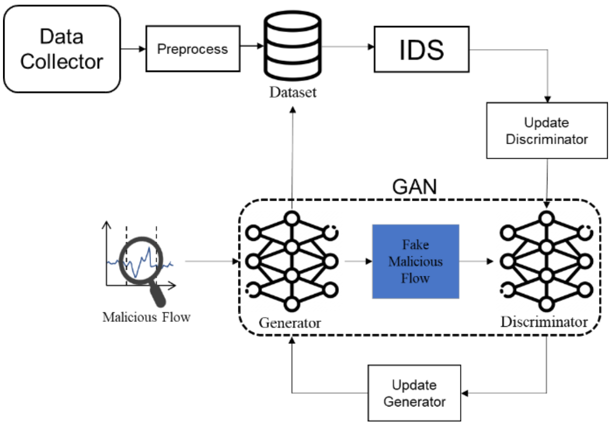

Fig. 1. System architecture.

3. System Structure

The structure of this research is introduced in this chapter. There are two stages: data collector and generative adversarial network data balance algorithm. The system architecture of the experiment is shown in Fig. 1.

3.1. *Data collector*

We collect the traffic in the real network set in our Lab and then use CICFlowmeter to transform the traffic into the statistics files. The CICFlowmeter is an Ethernet traffic Bi-flow generator and analyzer for anomaly detection, it can transform the pcap to the csv file. We call the traffic collected from the real network the true flow dataset.

We use the Cuckoo sandbox system to trigger malicious programs to create malicious traffic datasets. The malicious program samples are provided by the National High-Speed Network and Computing Center. The Cuckoo triggers malicious program samples through the network on the client-side, and Agent.py on the client-side will record various behaviors of the malicious program samples and send them back to the host on the user side. The main purpose of the above behaviors is to record the behavior of the malicious program samples without contacting malicious programs. For program samples, we analyzed the traffic generated by malicious samples and find the background traffic in the entire recording traffic is very small. The reason is that the Cuckoo system starts recording after the samples are triggered. The background traffic is so small that it can be ignored. Therefore, we call the traffic recorded by the Cuckoo system the malicious flow.

3.2. *GAN-BAL algorithm*

The core of the generative confrontation network is mainly divided into two parts: the generative model and the discriminative model. The generative model will generate fake traffic to the discriminant model during the training process. After the discriminating model discriminates the fake traffic true or false, the identification result will be fed back to the former, so that the generative model can improve the strategy of making malicious traffic. Then, the generative model will give the traffic generated after the modified strategy to the intrusion detection system which judges the authenticity of the forged traffic, and the discriminant model is used to improve the strategy of judging the traffic. After many times of updating, the generative model can generate fake traffic that is almost the same as the real traffic. The discriminant model will train a neural network that can identify the authenticity of the traffic, which is used to test the traffic generated by each generation model. On the other hand, the generation model will repeatedly improve its ability to "fake" from the feedback of each discriminant model. To fight against it, this is the concept of generating a confrontational network. We refer to the balanced data algorithm for generating adversarial networks designed by Huang et al. [7]. This algorithm uses Gaussian Noise as the training data for the generative model. In order to reduce training time and forge malicious traffic to be closer to the real malicious traffic. We change the input data of the generative model. We use the malicious traffic that triggers the malicious program sample to generate the malicious traffic training data. The flow chart of the algorithm is as follows Fig. 2. We call the improved algorithm the GAN-BAL algorithm.

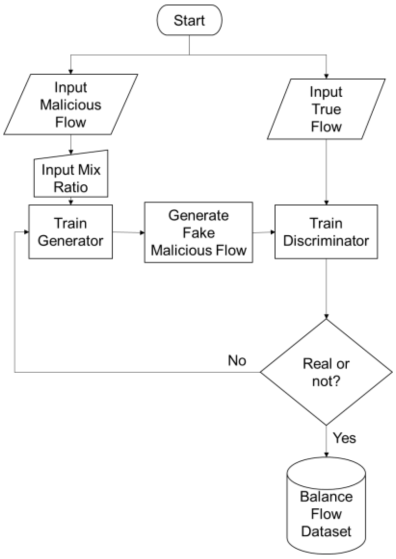

Fig. 2. The flow chart of the GAN-BAL algorithm.

4. Experimental Result

In order to verify whether the data balancing algorithm of this study improves the model results, we use supervised machine learning algorithms to evaluate its effects. We apply the GAN-BAL algorithm to the CICIDS 2017 dataset. We use five algorithms such as CNN, LSTM, BAT, SVM, and NB to perform model training on the original CICIDS 2017 dataset and the CICIDS 2017 dataset processed by the GAN-BAL algorithm to verify whether the algorithm improves the model results. The comparison results are shown in Table 1. As a result, the recall rate of CNN has increased by 20%, and the accuracy rate has increased by 4%; the recall rate of LSTM Increased by 10%, accuracy rate increased by 2%; BAT recall rate increased by 16%, accuracy rate increased by 3%; SVM recall rate increased by 15%, accuracy rate increased by 4%. Based on the above experimental results, although the use of the GAN-BAL algorithm proposed in this research will cause a loss of precision, the recall rate and accuracy rate are improved. The experimental results verify the

Table 1. Comparison of CNN, LSTM, BAT, SVM, and NB model training results

		Recall	Precision	F1-score	Accuracy
CNN	With GAN-BAL	0.991210	0.937769	0.963749	0.989934
	Without GAN-BAL	0.769745	0.998967	0.869503	0.957217
LSTM	With GAN-BAL	0.900657	0.947915	0.875400	0.918632
	Without GAN-BAL	0.800050	0.996801	0.887654	0.932155
BAT	With GAN-BAL	0.994952	0.953778	0.973930	0.992715
	Without GAN-BAL	0.824089	0.998934	0.903127	0.968218
SVM	With GAN-BAL	0.999636	0.992143	0.995876	0.998828
	Without GAN-BAL	0.816055	0.998819	0.898234	0.967708
NB	With GAN-BAL	0.165259	0.366583	0.227816	0.380103
	Without GAN-BAL	0.157882	0.368219	0.221004	0.352480

improvement of malicious intrusion detection using GAN-BAL algorithm.

5. Conclusion

This study analyzes the data sensitivity of the algorithm by evaluating the performance of each algorithm in the evaluation index of different datasets and proposes a data set balance algorithm for the unbalanced defects of the dataset. Compared with the research of intrusion detection system based on generative confrontation network proposed by Shahriar et al. [8], this research is developed for modern malicious attack behavior, and Huang et al. developed IGAN-IDS. The system uses noise as the input of the generative model [7]. This research uses the malicious traffic actually induced by the malicious sample as the input to reduce the training time and the basis for generating the malicious traffic.

Acknowledgements

This work was supported by the Ministry of Science and Technology (MOST) in Taiwan under contract numbers MOST 108-2221-E-006-110-MY3 and MOST 110-2218-E-006-013-MBK.

References

1. 1. The UCI KDD Archive, "KDD Cup 1999 Data," 28 10 1999. [Online]. Available: https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. [Accessed 1 6 2021].
2. 2. N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data set for Network Intrusion Detection systems," 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, 10-12 Nov., 2015.
3. 3. University of New Brunswick, "CICIDS 2017 dataset," University of New Brunswick, 1 7 2017. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html. [Accessed 15 6 2021].
4. 4. V. N. Vapnik, "An overview of statistical learning theory," IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 988 - 999, 1999.
5. 5. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation (1997), vol. 9, no. 8, pp. 1735-1780, 1997.
6. T. Su, H. Sun, J. Zhu, S. Wang and Y. Li, "BAT: Deep Learning Methods on Network Intrusion Detection Using NSL-KDD Dataset," IEEE Access, vol. 8, pp. 29575 - 29585, 2020.
7. S. Huang and K. Lei, "IGAN-IDS: An imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks," Ad Hoc Networks, vol. 105, p. 102177, 2020.
8. M. H. Shahriar, N. I. Haque, M. A. Rahman, and M. Alonso, "G-IDS: Generative Adversarial Networks Assisted Intrusion Detection System," 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 13-17 July, 2020.

Authors Introduction

Dr. I-Hsien Liu

He is a research fellow in the Taiwan Information Security Center @ National Cheng Kung University (TWISC@NCKU) and Department of Electrical Engineering, National Cheng Kung University, Taiwan. He obtained his PhD in 2015 in Computer and Communication Engineering from the National Cheng Kung University. His interests are Cyber-Security, Wireless Network, Group Communication and Reliable Transmission.

Mr. Cheng-En Hsieh

He received his B.S. degree from the Department of Communication Engineering, National Central University, Taiwan in 2019. He got the M.S. degree in National Cheng Kung University in Taiwan. His research focuses on network communication and cyber security.

Ms. Wei-Min Lin

She received her B.S. degree from the Department of Electrical Engineering, Yuan Ze University, Taiwan in 2020. She is acquiring the master’s degree in Department of Electrical Engineering / Institute of Computer and Communication Engineering, National Cheng Kung University in Taiwan.

Prof. Jung-Shian Li

He is a full Professor in the Department of Electrical Engineering, National Cheng Kung University, Taiwan. He graduated from the National Taiwan University, Taiwan, with B.S. in 1990 and M.S. degrees in 1992 in Electrical Engineering. He obtained his PhD in 1999 in Computer Science from the Technical University of Berlin, Germany. He teaches communication courses and his research interests include wired and wireless network protocol design, network security, and network management. He is the director of Taiwan Information Security Center @ National Cheng Kung University. He serves on the editorial boards of the International Journal of Communication Systems.

s

Prof. Chu-Fen Li

She is an Associate Professor in the Department of Finance at the National Formosa University, Taiwan. She received her PhD in information management, finance and banking from the Europa-Universität Viadrina Frankfurt, Germany. Her current research interests include intelligence finance, e-commerce security, financial technology, IoT security management, as well as financial institutions and markets. Her papers have been published in several international refereed journals such as European Journal of Operational Research, Journal of System and Software, International Journal of Information and Management Sciences, Asia Journal of Management and Humanity Sciences, and others.