Intrusion Detection Model Based on Autoencoder and XGBoost

Yunxiang Kang¹, Minsheng Tan¹*, Ding Lin¹, Zhiguo Zhao¹
¹University Of South China School of Computer Science, Hengyang, Hunan, China
*20192021210916@stu.usc.edu.cn

Abstract. In recent years, machine learning algorithms have been extensive used for intrusion detection field. At the same time, these algorithms still suffered from low accuracy due to data imbalance. To improve accuracy of detection, an intrusion detection model based on Autoencoder (AE) and XGBoost (IDAE-XG) is proposed. The training algorithm and detection algorithm related to IDAE-XG are given. IDAE-XG constructs the training set with preprocessed normal data. Data preprocessing includes feature selection and feature grouping. Through detection, XGBoost is used to predict results, which effectively improves prediction accuracy. The superiority of the proposed IDAE-XG is empirically demonstrated with extensive experiments conducted upon CSE-CIC-IDS2018. The experimental comparison show that IDAE-XG performs better than the KitNet model in the test, and has achieved a great improvement in accuracy and recall rate.

1. Introduction

With the advancement and development of the computer and communication technology, network attacks can bring huge losses to users, and some losses and damages are even irrecoverable. Network intrusion detection has become an indispensable part of network defense facilities. All kinds of machine learning algorithms[1]and deep learning algorithms[2,3] are used in intrusion detection. According to detection performance, IDS can be divided into signature detection and anomaly detection[4]. Generally, the signature-based IDS based on known signatures. This method can effectively identify existing attacks in the signature database, but cannot identify unknown attacks and variants of known attacks[5]. IDS based on abnormal behavior detection attempts to learn network traffic behavior to classify traffic, and can detect unknown attacks[6]. Yang et al proposed an intrusion detection system combining CDBN and SCAE. CDBN and SCAE are used to detect abnormal attack characteristics and reduce data redundancy respectively.[7] Han, K, Wang, etc. used AE to extract features from the dataset.[8] But the AE itself was not used for intrusion detection. Mirsky et al. proposed an efficient detection of attacks on the local network without supervision, named Kitsune[9]. Tang et al. proposed intrusion detection model SAAE-DNN which employs SAAE to extract features.
and initialize the weights of the DNN[10]. However, The IDS does not take into account the problem of unbalanced datasets and cannot effectively detect attacks with few data samples.

Because the original network data contains a lot of redundancy and useless features, it is easy to appear dimensional disaster to reduce the accuracy, and the unbalanced distribution of the dataset can easily lead to model detection errors. This paper proposed an intrusion detection model based on Autoencoder and XGBoost (IDAE-XG). The application of XGBoost in the intrusion detection model not only produces high-precision detection results, but also has higher efficiency[11]. Firstly, IDAE-XG preprocess the original dataset. If dataset is unbalanced, upsample the dataset. Secondly, IDAE-XG uses random forest algorithm to obtain the best features. AP achieves feature grouping by classifying the best features. Finally, the intrusion detection model reconstructs the new data features by AE, calculates the RMSE of the old and new data, and classifies the RMSE by XGBoost.

2. IDAE-XG model

2.1. IDAE-XG structure

As shown in Figure 1, IDAE-XG is composed of two modules: data preprocessing and intrusion detection. It can effectively detect anomalies in network traffic.

Preprocess the original data, replace the default value and infinite value with 0. If abnormal data or normal data is less than 10% of the dataset, upsampling is performed. Random forest algorithm is used.
to obtain the best features and reduce the feature dimensionality of the data. AP algorithm divides the processed data into multiple feature subsets. The intrusion detection module consists of XGBoost and 3-layer AE. The input data of AE is \( f=\{f_1, f_2, f_3, ..., f_n\} \), all subsets of feature grouping. During the training phase, AE uses a subset of normal data for training, and the number of AE is equal to the number of subsets after feature grouping. During the detection phase, the RMSE of each AE is calculated and the average is sent to XGBoost.

2.2. IDAE-XG main algorithm

2.2.1. IDAE-XG-TA Train the AE with normal data, and get the hyperparameter of the AE. After calculating the RMSE value for each subset, the RMSE average \( \varphi \) is used as the threshold to determine whether it is abnormal. Therefore, in the detection stage, \( \varphi \) is the reference value, if it exceeds \( \varphi \), it is considered abnormal\[12\]. The choice of the threshold \( \varphi \) is critical to the performance of IDAE-XG. Model linearly transform the original data by min-max normalization and map the resulting values to \([0,1]\).

| Tab. 1 IDAE-XG’s training algorithm |
|------------------------------------|
| **IDAE-XG-TA:**                    |
| Input: Feature grouping subsets \( x(\text{training set}) \); |
| Output: threshold \( \varphi \); |
| 1 Initializing \( L(2): z = \text{zeros} (k) \); |
| 2 for each \( \theta _i \in L(1) \) do |
| 3 \( n_i = \text{norm} (x_i); \) //normalization: |
| 4 \( W_i, z_i = h^{(1)} \theta _i (n_i); \) //propagates forward |
| 5 \( W_i^\top, y_i = h^{(2)} \theta _i (z_i); \) //propagates forward |
| 6 \( \Delta s_i = b_{i}(x_i, y_i); \) //back-propagation |
| 7 \( \theta _i = \text{SGD}(W_i, s_i); \) //updating weights |
| 8 \( z[i] = \text{RMSE}(n_i, y_i); \) //reconstruction error calculation |
| 9 end |
| 10 \( z[i]’ = \text{norm} (z[i]); \) |
| 11 \( \varphi = \text{avg} (z[i]’); \) |
| 12 return \( \varphi \) |

2.2.2. IDAE-XG-D4 In the training phase, the AE is trained using the features of the normal data in the training set. And threshold \( \varphi \) is calculated during the training period. In the detection stage, 25% of the normal data and all abnormal data are input to the trained AE. XGBoost can be used as anomaly classifier. Since the AE is used to learn the feature representation of the normal data, the reconstruction error will be small, and the abnormal data has some different feature compared with the normal data, which cannot be well mapped to the normal data. In the low dimensional data space, it cannot be reconstructed well by the decoder, and it will have a relatively high reconstruction error. Regarding the reconstruction error as the data anomaly score \( s \), the higher the anomaly score, the greater the possibility of anomalies. By processing the threshold \( \varphi \) of the anomaly score, all the data in the test set can be accurately classified.
Tab. 2 IDAE-XG’s detection algorithm

| IDAE-XG-DA: |
|-------------|
| Input: Feature grouping subsets x(test set); |
| Output: normal or abnormal status; |
| 1 Initializing L(2): z = zeros (k) ; |
| 2 for each \( \theta_i \in L(1) \) do |
| 3 \( n_i = \text{norm} (x_i) \); // normalization |
| 4 \( W_i,z_i = h(1)\theta_i(n_i) \); // propagates forward |
| 5 \( W^T_i, y_i = h(2)\theta_i(z_i); // \text{ propagates forward} \) |
| 6 \( z[i] = \text{RMSE}(n_i, y_i); // \text{ reconstruction error calculation} \) |
| 7 end |
| 8 \( z[i]' = \text{norm}(z[i]); \) |
| 9 \( s = \text{avg}(z[i]'); \) |
| 10 if \( s < \phi \) |
| 11 return abnormal; |
| 12 return normal; |

3. Experiment and analysis

3.1. CSE–CIC–IDS2018

CSE–CIC–IDS2018 uses CICFlowMeter to extract more than 80 network flow characteristics. It contains 7 basic features and more than 70 functional features. But the dataset is extremely unbalanced. For example, the datasets Brute Force-XSS and SQL Injection have a large difference in the number of normal datas and abnormal datas. IDAE-XG adopts up-sampling to solve the problem of data imbalance and eliminate abnormally sensitive problems.

3.2. Experimental results and analysis

AUC is defined as the area enclosed with the coordinate axis under the ROC curve. Obviously, The closer the AUC is to 1.0, the higher the authenticity of the detection method. Fig.2 shows the AUC of IDAE-XG detection results. IDAE-XG has a high recall rate in most data sets, and its recall rate is mostly close to 1. However, the performance of the IDAE-XG is poor in the Infiltration dataset. Because the attacker invaded the intranet and became a normal user. An intruder sends HTTP request containing malformatted SQL statement to perform SQL injection. So it is difficult to detect for IDAE-XG.
Fig. 2 The AUC of IDAE-XG detection results

In intrusion detection, people expects the recall (recall = TP/(TP+FN)) to be as high as possible. Table 3 presents the results and comparison of IDAE-XG and KitNet in each data set. Bold means improved results. The recall rate of IDAE-XG is almost always above 98% in all datasets except the Infiltration dataset. And the recall of IDAE-XG is greatly improved over OriginalData (not upsampled), KitNet in dataset Bot, SQL Injection, Infiltration, Brute Force -XSS, Brute Force-Web, DoS attacks-Hulk. Therefore, The experimental results prove that IDAE-XG has significantly improved the recall rate and eliminated data imbalance.

| datasets                  | IDAE-XG | KitNet [8](%) |
|---------------------------|---------|---------------|
| Bot                       | 98.89   | 45            |
| SQL Injection             | 100.0   | 9.43          |
| Infiltration              | 29.80   | 23.37         |
| SSH-BruteForce            | 99.82   | 99.77         |
| FTP-BruteForce            | 99.96   | 100.00        |
| Brute Force -XSS          | 100     | 42.38         |
| Brute Force-Web           | 99.78   | 5.22          |
| DDOS attack-HOIC          | 100     | 99            |
| DDOSAttack-LOIC-UDP       | 100     | 99.94         |
| DoS attacks-SlowHTTPTest  | 99      | 99            |
| DoS attacks-Hulk          | 99.81   | 99.52         |

4. Conclusion
IDAE-XG consists of two modules: data preprocessing and intrusion detection. The three-layer AE structure and XGBoost produce high precision detection results. The upsampling algorithm can effectively solve the common data imbalance problem in intrusion detection. IDAE-XG performed better than some popular algorithms, for example KitNet algorithm. Because RF and AP are valid to merge similar features and obtain best feature subset. So IDAE-XG reduces the computation cost. After experiments and comparisons, IDAE-XG can detect most attacks accurately and eliminate data imbalance. One direction of future research tries to implement this model in an actual mobile network platform and evaluate the performance of intrusion detection in real scenarios. Another research direction is to consider active learning and adversarial training in this model.

Acknowledgments
This work was supported by Project 61403183 of the National Science Foundation of China, Project
18A230 of the Hunan Provincial Education Office Science Research of China. Project 20183350502 of the Hunan Provincial Finance Office Science Research of China.

References

[1] Rashid A, Siddique M J and Ahmed S M 2020 Machine and Deep Learning Based Comparative Analysis Using Hybrid Approaches for Intrusion Detection System 2020 3rd International Conference on Advancements in Computational Sciences (ICACS) doi:10.1109/icacs47775.2020.9055946

[2] Ferrag M A, Maglaras L, Moschoyiannis S and Janicke H 2020 Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study Journal of Information Security and Applications 50 p 102419 doi:10.1016/j.jisa.2019.102419

[3] Bu S J and Cho S B 2020 A convolutional neural-based learning classifier system for detecting database intrusion via insider attack Information Sciences 512 pp 123–136 doi:10.1016/j.ins.2019.09.055

[4] Yassine M, Mohammad S, Mamoun A, Youssef B Machine Intelligence and Big Data Analytics for Cybersecurity Applications Springer, Cham:2021-01-01

[5] Masdari M and Khezri H 2020 A survey and taxonomy of the fuzzy signature-based Intrusion Detection Systems Applied Soft Computing p 106301 doi:10.1016/j.asoc.2020.106301

[6] Masdari M and Khezri H 2020 Towards fuzzy anomaly detection-based security: a comprehensive review Fuzzy Optimization and Decision Making doi:10.1007/s10700-020-09332-x

[7] Yang L, Li J, Yin L, Sun Z, Zhao Y and Li Z 2020 Real-Time Intrusion Detection in Wireless Network: A Deep Learning-Based Intelligent Mechanism IEEE Access 8 pp 170128–170139 doi:10.1109/access.2020.3019973

[8] Han K, Wang Y, Zhang C, Li C and Xu C 2018 Autoencoder Inspired Unsupervised Feature Selection 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp 2941-2945 doi: 10.1109/ICASSP.2018.8462261

[9] Mirsky Y, Doitshman T, Eloveci Y, Shabti A 2018 Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection Proceedings of NDSS Internet Society doi:10.14722/ndss.2018.23211

[10] Tang C, Luktarhan N and Zhao Y 2020 SAAE-DNN: Deep Learning Method on Intrusion Detection Symmetry 12(10) pp 1695 doi:10.3390/sym12101695

[11] Y. Kawaguchi and Endo T 2017 How can we detect anomalies from subsampled audio signals? 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP) pp 1-6, doi: 10.1109/MLSP.2017.8168164

[12] Paulo Gil, et al. 2019 Outliers detection methods in wireless sensor networks Artificial Intelligence Review 52 p 4 doi:10.1007/s10462-018-9618-2