NADSR: A Network Anomaly Detection Scheme Based on Representation

Xu Liu[1,2], Xiaoqiang Di[1,2,3](✉), Weiyou Liu[1], Xingxu Zhang[1], Hui Qi[1,2], Jinqing Li[1,2], Jianping Zhao[1,2], and Huamin Yang[1,2]

[1] School of Computer Science and Technology, Changchun University of Science and Technology, Changchun 130022, China
dixiaoqiang@cust.edu.cn

[2] Jilin Province Key Laboratory of Network and Information Security, Changchun 130022, China

[3] Information Center, Changchun University of Science and Technology, Changchun 130022, China

Abstract. Deep learning has been widely used for identifying anomaly network traffic. It trains supervised classifiers on a pre-screened numerical traffic feature dataset in the most cases, so the classification effectiveness depends heavily on feature representation. There is no unified feature representation method, and the current feature representation methods cannot profile traffic precisely. Therefore, how to design a traffic feature representation method to profile traffic is challenging. We propose a Network Anomaly Detection Scheme based on data Representation (NADSR). Data representation method converts raw network traffic into images by treating every numerical feature value as an image pixel and then creating a circulant pixel matrix for a traffic sample. It retains the traffic feature's spatial structure instead of padding empty pixels with constant values while directly reshaping a long feature vector into a pixel matrix. Experimental results verify the effectiveness of the proposed NADSR. It improves the overall detection accuracy compared with state-of-the-art methods, and also provides reference to solve security-related classification problems.

Keywords: Anomaly detection · Traffic feature · Data representation

1 Introduction

With the fast development of the Internet, network security has become increasingly challenging. The Report of National Computer Network Emergency Technical Response Center in 2018 points out that malicious network attack is still

Supported by the Department of Science and Technology of Jilin Province grant NO. 20190302070GX, the Education Department of Jilin Province grant NO. JJKH20190598KJ, Jilin Education Science Planning Project (GH180148) and Jilin Province College and University "Golden Course" Plan Project (Network protocol and network virus virtual simulation experiment).

G. Li et al. (Eds.): KSEM 2020, LNAI 12274, pp. 380–387, 2020.
https://doi.org/10.1007/978-3-030-55130-8_33

one of the most important network security issues and deserves public attentions [1]. Anomaly detection, as a necessary part of security defense, is used to identify anomalies [5,8] for further defense. Though network anomaly detection has obtained some achievements with the help of deep learning methods, it has occurred with some difficulties. How to represent traffic feature is critical [7]. If extracted features cannot be learned to profile traffic efficiently, the classification accuracy will be influenced.

Therefore, we propose a Network Anomaly Detection Scheme based on data Representation (NADSR). The main challenge to be addressed in this paper is the network traffic feature representation in the data preprocessing stage.

To solve this challenge, we design a traffic-to-image conversion method for representation learning in the anomaly detection. Generally, one raw traffic sample is represented as a long feature vector [14]. Some researches have explored to reshape a long vector into a pixel gray value matrix directly to obtain an image [2,10], but it might disrupt the spatial structure and further decrease classification accuracy [16]. Some symbolic network traffic features [6,7,15] are encoded by One-Hot encoder and word embedding technique, this will produce massive zeroes that influence layer-by-layer processing effect of deep learning [11]. Therefore, to avoid the spatial feature loss and reduce the number of zeros within the pixel matrix, we bundle the sparse discrete features that are encoded by One-Hot scheme first, and then design a novel representation learning method – re-circulation pixel permutation strategy (RPP) by creating circulant pixel matrix, which retains spatial structure of raw network traffic. Finally, the data representation method coupled with Convolutional Neural Network (CNN) is constructed in the proposed NADSR.

The experiments are carried on two public network traffic datasets: NSL-KDD [12] and UNSW-NB15 [9]. Experimental results verify the effectiveness of NADSR. The contributions of this study are summarized as follows:

(1) The discrete features bundling method reduces the number of zeros within the pixel matrix, which is helpful for CNN to learn the traffic feature.
(2) The proposed NADSR improves the overall classification accuracy to 81.4% and 94.9% on NSL-KDD and UNSW-NB15, respectively.
(3) The proposed representation learning method RPP is not only suitable for CNN but also for other detection algorithms.

The remaining of this paper is structured as follows. Section 2 states the main problem and Sect. 3 illustrates the main methodology. The experimental results are described in Sect. 4. Section 5 concludes the full paper and the future work.

2 Problem Description

In a general network anomaly detection problem, the detection algorithm aims to identify anomalies deviated from the normal network traffic, and then report to the security operator for further analysis. This paper focuses on identifying anomalies, which can be abstracted as a binary-classification problem. Benefit from deep learning technique, it can be solved by training an effective classifier on the labeled data.

In this paper, network packets are captured first by the tcpdump, and then the features are extracted. After that, they are encoded into the image pixel matrices to represent as images. The classification model, CNN is used in this paper and it is trained on the image dataset. Finally, CNN will be evaluated on the new coming test data.

The core work focuses on how to retain the spatial characteristics of traffic feature in the data representation stage and then ensure a high accuracy of the detection model that is trained on the represented images. In this paper, we design an image conversion method which assumes traffic features have structural relationship and then represents them into images by the designed Re-circulation Pixel Permutation (RPP) strategy.

3 NADSR: A Scheme to Network Anomaly Detection Based on Representation

In the proposed NADSR scheme, we first encode one record of traffic feature into a long vector, then discard useless features and bundle the discrete features, and subsequently perform normalization. With the normalized features, data representation method is designed to convert traffic features into images. The images are used for training classification model. This section outlines the main work. The data pre-processing is introduced first, data representation method is detailed subsequently.

3.1 Data Pre-processing

There are both numerical and symbolic features in the traffic feature. The symbolic features can also be seen as the discrete features, and a discrete feature can be seen as a binary feature. To eliminate the influence of symbolic features on traffic representation, we bundle the discrete features. The discrete features are encoded into 0–1 vector by One-Hot encoder first. Assume all the discrete features are the same weight, the 0–1 vectors are bundled further to obtain a decimal feature. For example, there are four discrete features, after the bundling process, four discrete features are reduced into one numerical feature. This bundling operation not only maintains unity of each feature, but also keeps combination among each discrete feature.

3.2 Feature Reduction

To optimize the remaining features, a feature filter is designed to remove useless features. As the dimensions of features are different, using standard deviation to compare discreteness of features is inappropriate, so the coefficient of variance C_v is introduced and defined as (1).

$$C_{vi} = \frac{\sigma_i}{\mu_i} * 100\% \tag{1}$$

where σ_i and μ_i are standard deviation and mean of i^{th} feature. Generally, a higher C_v indicates a higher discreteness, and the feature of higher C_v plays a more important role. Specially, when the mean μ_i is equal to zero, the corresponding feature will be seen as unimportant relatively.

3.3 Data Normalization

Data normalization can eliminate differences among different dimensional data, so it is therefore widely used in machine learning. Because features of different scales will result in the unreliability of training model, we normalize them in the same distribution. Rescale-Min-max normalization is designed in this paper as (2).

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} * (1 - a) + a \tag{2}$$

where x_{max} and x_{min} represent the maximum and minimum value of feature x_i, x_i and x_i' represent the raw feature and the normalized feature. To change the minimum value of the normalized, we re-scale the range of the normalized feature from $[0, 1]$ into $[a, 1]$ where indicator $a \in (0, 1)$.

3.4 Image Representation

To learn the deep characteristics of traffic feature automatically, we convert the feature vector of every traffic sample into a pixel matrix and then feed them into CNN as images. A Recirculation Pixel Permutation (RPP) strategy is designed. The RPP function is defined in (3) which is used to convert a long vector into a circulant matrix, where x_i is i^{th} sample of the dataset, and it is an original long vector with n elements. x_i' is obtained by moving every element $x_{ij} (j = 1, 2, \cdots, n)$ of x_i one unit forward every time, then x_i' is used to represent pixel values of the transformed image whose dimension is $n * n$. RPP not only retains the original spatial structure of sample, but also facilitates the detection algorithm to mine relationships among the adjacent features deeply.

$$x_i = [x_{i1}, x_{i2}, ..., x_{in}] \rightarrow \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{in} \\ x_{i2} & \cdots & x_{in} & x_{i1} \\ \cdots & x_{in} & \cdots & \cdots \\ x_{in} & x_{i1} & x_{i2} & \cdots \end{bmatrix}_{n*n} = x_i' \tag{3}$$

3.5 Classification Model

After the data representation, classification model is used to verify their effectiveness and adaptability. CNN, as a typical deep learning algorithm, has achievements in the area of image processing. We use CNN to evaluate the performance and effectiveness of the proposed NADSR. The workflow of the proposed NADSR is detailed in Algorithm 1.

Algorithm 1. NADSR workflow

```
Input: Dataset $D$ = training data, validation data, test data;
**Preprocess:** Feature Reduction, Data Normalization;
**Representation:**
**while** $x \in$ D **do**
    $x'(1,:) = x = [x_1, x_2, \cdots, x_n]$
    **for** $1 \leq i \leq n$ **do**
        $x'(i+1,:) = [x_{i+1}, x_{i+2}, \cdots, x_n, x_1, \cdots, x_i]$
    **end for**
**end while**
**Train** → **Validation** → **Test**
Output: acc_test, loss_test, and test report
```

4 Experiment

All experiments are conducted on an Ubuntu 16.04 LTS machine with Intel Xeon (R)W-2123, 3.6 GHz CPU, GeForce GTX TITAN Xp COLLECTORS EDITION GPU and 12 GB VRAM. Two public datasets, NSL-KDD [12] and UNSW-NB15 [9] are used for evaluation.

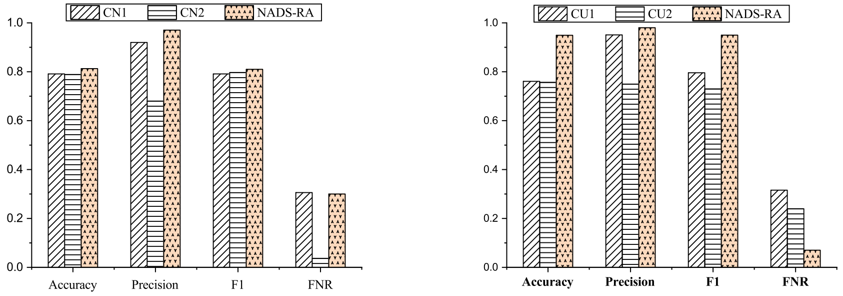
4.1 Data Representation Method Comparison

Figure 1 shows the comparison results with other researches that use different data representation methods. CN1 and CN2 represent the results obtained from [6] and [13], respectively on NSL-KDD test$^+$, and CU1 and CU2 indicate the results obtained from [13] and [3], respectively on UNSW-NB15. It can be found that, method used in previous research [13] obtains a lower FNR and a higher Recall than the others. In contrast, our result is better than the others in the perspectives of Accuracy, Precision and F1. Additionally, the accuracy rate in this paper is larger than [4] by nearly three percent. Though the results in research [6] are close to us, there is an over-fitting in its work by analyzing its confusion matrix.

In all, compared with other representation methods, the proposed data representation method is useful and competitive. It represents almost no distortion on the raw data. As a consequence, it can be regarded as supplying almost complete knowledge during representation. Its well performance can be contributed to the effective representation method and also the detection algorithm CNN, therefore, it provides the positive influence of data representation on the anomaly detection.

4.2 Detection Algorithm Comparison

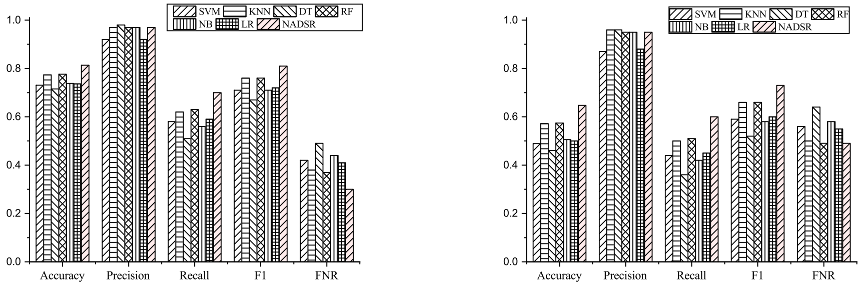
The proposed representation method RPP works in data preprocessing stage, so its function is to help the further detection algorithm to improve classification result. Therefore, we measure the effectiveness of the proposed data representation on other detection algorithms. Figure 2 shows the comparison results conducted on NSL-KDD. We compare six approaches, support vector machine (SVM), k-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), Naive Bayesian (NB) and Logistic Regression (LR), they are applied to test on the test$^+$ and test^{-21} of NSL-KDD.



(a) Comparison on NSL-KDD

(a) Comparison on UNSW-NB15

Fig. 1. Comparison result with other representation methods.



(a) test$^+$

(b) test^{-21}

Fig. 2. Comparison with other detection algorithms on NSL-KDD.

The obvious finding is that all evaluation metrics are close, and it suggests that the proposed data representation method is not only helpful for the CNN but also helpful for the other detection algorithms. Take a deep comparison, our method performs better than other detection algorithms. When performing data fitting, deep learning models can extract complex features than traditional machine learning models and mine hidden characteristics of the samples. Hence, deep learning models have a better representation learning ability than the shallow machine learning models [7].

4.3 Discussion

The classification based on image representation is still at the beginning stage and not mature. Though it might not be the best solution compared with other state-of-the-art approaches in some perspectives, it explores a new appliance mode for using powerful technique such as deep learning to solve the anomaly detection problem.

5 Conclusion

This paper proposes a network anomaly detection scheme based on representation, NADSR. It converts traffic features into images through a novel representation learning method. The proposed NADSR not only retains the original spatial structure of raw traffic features, but also propels the detection algorithm to learn the hidden knowledge. The experimental results suggest that NADSR is effective to improve overall accuracy. It also outperforms some other state-of-the-art representation methods and shows a robust adaptability on different detection algorithms.

The proposed representation method has improved classification effectiveness, but there still exists performance imbalance such as the rare anomaly is hard to detect. In the future, we will tend to study the data re-sampling method to solve the imbalance issue.

References

1. Summary of internet security situation in china in 2018, national computer network emergency technology processing and coordination center (2019). http://www.cac.gov.cn/2019-04/17/c_1124379080.htm
2. Blanco, R., Malagón, P., Cilla, J.J., Moya, J.M.: Multiclass network attack classifier using CNN tuned with genetic algorithms. In: 28th International Symposium on Power and Timing Modeling, Optimization and Simulation (PATMOS), pp. 177–182. IEEE (2018). https://doi.org/10.1109/PATMOS.2018.8463997
3. Khan, N.M., Madhav C, N., Negi, A., Thaseen, I.S.: Analysis on improving the performance of machine learning models using feature selection technique. In: Abraham, A., Cherukuri, A.K., Melin, P., Gandhi, N. (eds.) ISDA 2018 2018. AISC, vol. 941, pp. 69–77. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-16660-1_7
4. Kwon, D., Natarajan, K., Suh, S.C., Kim, H., Kim, J.: An empirical study on network anomaly detection using convolutional neural networks. In: IEEE 38th International Conference on Distributed Computing Systems (ICDCS), pp. 1595–1598 (2018). https://doi.org/10.1109/ICDCS.2018.00178
5. Kwon, D., Kim, H., Kim, J., Suh, S.C., Kim, I., Kim, K.J.: A survey of deep learning-based network anomaly detection. Cluster Comput. **22**(1), 949–961 (2017). https://doi.org/10.1007/s10586-017-1117-8
6. Li, Z., Qin, Z., Huang, K., Yang, X., Ye, S.: Intrusion detection using convolutional neural networks for representation learning. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.-S.M. (eds.) ICONIP 2017. LNCS, vol. 10638, pp. 858–866. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70139-4_87

7. Liu, H., Lang, B., Liu, M., Yan, H.: CNN and RNN based payload classification methods for attack detection. Knowl. Based Syst. **163**, 1–10 (2018). https://doi.org/10.1016/j.knosys.2018.08.036
8. Luo, X., Di, X., Liu, X., Qi, H., Li, J., Cong, L., Yang, H.: Anomaly detection for application layer user browsing behavior based on attributes and features, vol. 1069, pp. 1–9. Elsevier, Suzhou (2018). https://doi.org/10.1088/1742-6596/1069/1/012072
9. Moustafa, N., Slay, J.: UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: Military Communications and Information Systems Conference (2015). https://doi.org/10.1109/MilCIS.2015.7348942
10. Nsunza, W.W., Tetteh, A.Q.R., Hei, X.: Accelerating a secure programmable edge network system for smart classroom. In: IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, pp. 1384–1389. IEEE (2018). https://doi.org/10.1109/SmartWorld.2018.00240
11. Potluri, S., Ahmed, S., Diedrich, C.: Convolutional neural networks for multi-class intrusion detection system. In: Groza, A., Prasath, R. (eds.) MIKE 2018. LNCS (LNAI), vol. 11308, pp. 225–238. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-05918-7_20
12. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed analysis of the KDD cup 99 data set. In: IEEE International Conference on Computational Intelligence for Security and Defense Applications (2009). https://doi.org/10.1109/CISDA.2009.5356528
13. Vinayakumar, R., Alazab, M., Soman, K.P., Poornachandran, P., Al-Nemrat, A., Venkatraman, S.: Deep learning approach for intelligent intrusion detection system. IEEE Access **7**, 41525–41550 (2019). https://doi.org/10.1109/ACCESS.2019.2895334
14. Vinayakumar, R., Soman, K., Poornachandran, P.: Applying convolutional neural network for network intrusion detection. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1222–1228. IEEE (2017). https://doi.org/10.1109/ICACCI.2017.8126009
15. Wu, K., Chen, Z., Li, W.: A novel intrusion detection model for a massive network using convolutional neural networks. IEEE Access **6**, 50850–50859 (2018). https://doi.org/10.1109/ACCESS.2018.2868993
16. Xie, K., Li, X., Xin, W., Cao, J., Zheng, Q.: On-line anomaly detection with high accuracy. IEEE/ACM Trans. Netw. **26**(3), 1222–1235 (2018). https://doi.org/10.1109/TNET.2018.2819507