Research on Model-based Abnormal Traffic Detection Method

Jiaxin Han¹, Xiaowei Wang²

¹School of Computer Science of Xi'an Shiyou University, China
²School of Computer Science of Xi'an Shiyou University, China

*Corresponding Author: Xiaowei Wang
Email: 19211060572@stumail.xsyu.edu.cn, 859675296@qq.com

Abstract. Information security has become a concern of all walks of life, and anomaly detection can protect information security, so anomaly detection has become a research hotspot. In this paper, the principles of four commonly used model-based anomaly detection methods, namely, depth-based, distance-based, density-based and deep learning-based detection methods are introduced and their research status is reviewed. Analyzed the characteristics of the four methods, and finally pointed out the future development trend of anomaly detection methods and gave a conclusion.

1. Introduction
With the rapid development of science and technology and the explosive growth of information, the Internet has become an indispensable part of various industries and fields, and the security of network information is a vital issue for a country. Network traffic anomaly detection has penetrated into every aspect of life. In 2020, Li Yuan and other scholars put forward the problem of detecting abnormal network traffic in Internet of Vehicles flooding attacks[1]. By using Hurst self-similar parameter estimation method to calculate and detect data packet information, through continuous experiments, it is concluded that the method proposed in the article is The low, medium and high flooding attack intensity can effectively detect malicious nodes, and the effect is better. In 2021, Chen Yan scholars proposed the problem of detecting abnormal network traffic of ship's network server. By optimizing the abnormal characteristics of the network traffic of the ship, finding the best feature subset for denoising modeling, and getting suitable for the detection of ship network traffic anomaly Model [2]. In the same year, Meng Fanyu and other scholars proposed the problem of detecting abnormal network traffic in the military Internet of Things. By applying the convolutional neural network to the detection of abnormal network traffic, the NPCNN model was established by combining the characteristics of the selected data set. The experiment proved that the NPCNN proposed in the article has better performance and higher accuracy [3]. It can be seen that the detection of abnormal traffic in the network has been a hot issue of current research.

2. Abnormal flow

2.1 The network is abnormal
Network abnormalities refer to abnormal behaviors caused by network failures, network abuses, and network attacks. These abnormal behaviors are reflected in abnormal traffic in the network. Network...
abnormalities will cause the waste of network resources and make the performance of network equipment and terminal hosts decrease, which in turn causes more information security problems for network users. Refining network anomalies into different areas has different definitions. A relatively broad definition is "an observation that is far from other observations and is worthy of suspicion [4]", the detection of anomalies and analysis is very meaningful.

2.2 Anomaly classification
The anomaly classification adopted in the article is based on the classification of Ahmed et al. [5], which can be divided into three categories:

1. Point anomaly, which refers to the deviation of a specific data instance from the normal pattern of the data set, which can be called a point abnormal. For example, a person’s daily consumption of food and beverage is 100 yuan, but if the food and beverage consumption suddenly becomes 1,000 yuan on a random day, we can say that the data instance is a bit abnormal.

2. Context abnormality refers to the abnormality that appears when the data instance deviates from the context more than a certain degree in a specific context. Network traffic data has a certain periodicity and trend. For example, the passenger flow of buses in the morning on weekdays is more than the passenger flow in the morning on weekends. This is not abnormal, but a reasonable periodic change. When the number of bus passengers on weekend mornings is the same or more than that on weekdays, this is an anomaly, which is called a contextual anomaly.

3. Collective abnormality refers to the abnormal performance of the collection of similar data relative to the entire data set. These data instances are called collective abnormalities. For example, if the electrocardiogram of the human body has a low value for a long time, it means that there is some physiological abnormality [6], and a single low value will not be called an abnormal value.

2.3 Anomaly detection performance indicators
The indicators for anomaly detection are more complicated, because the data set used for anomaly detection has more normal data and less abnormal data, which is called an unbalanced data set. You can't just use a single correct rate to evaluate. Here we introduce three evaluation indicators: True Positive Rate (TPR), False Positive Rate (FPR), and Precision (P). TPR represents the ratio of the number of samples predicted to be abnormal and in fact abnormal to the actual abnormal data. The larger the TPR value, the better the performance. FPR represents the ratio of the number of samples predicted to be abnormal but actually normal to the actual normal number. The smaller the FPR value, the better the performance. P represents the ratio between the number of samples predicted to be abnormal and the number of samples predicted to be abnormal. The larger the value of P, the better the performance. These three indicators will have different emphasis in actual problems, and they will vary according to specific problems.

3. Model-based abnormal traffic detection method
Nowadays, there are many scenarios for abnormal traffic detection. According to different scenarios, different models can be established for detection. The article mainly introduces four model-based abnormal traffic detection methods, namely detection based on depth, distance, density, and deep learning method.

3.1 Depth-based detection method
The depth-based detection method is to use the location of the edge of the data point to find outliers, and determine the number of layers and the number of outliers based on actual needs. It was proposed by Tukey in 1997. Generally speaking, abnormal points are distributed on the edge and are relatively sparse. In the Fig.1, the depth of the outermost layer is set to 1. The abnormal points can be filtered out by setting the threshold value of abnormal depth. If the abnormal depth threshold is set to 2, then the depth of 2 and the depth below 2 are regarded as abnormal values.
The above-mentioned model is only suitable for two-dimensional and three-dimensional spaces, but the idea of the algorithm is still worth learning. By changing the way of calculating depth, low-dimensional space can be extended to high-dimensional space. Scholar Wang Jingxian combined the deep model and autoencoder to detect abnormal power data. Taking photovoltaic power generation system as an example, combining photovoltaic power generation attributes to expand the original collected data and train the deep auto-encoding network to obtain abnormalities. Detect the model, and the experiment proves that the model has a high accuracy rate [7].

3.2 Detection method based on distance
The distance-based detection method uses the distance between each point and its neighbors to measure whether a point is abnormal. The premise of using this detection method is that there are many nearby points around the normal point, while the points around the abnormal point are few and far away. The idea of the model is relatively simple, but the model can be extended to models such as grid-based distance models, nested loop-based models, and K-means. Sun Yuhao scholars use the distance correlation coefficient to integrate the GPR model to detect the collected satellite data. Experiments show that this method can detect anomalies in the early stage of satellite failure and reduce the false alarm rate [8].

3.3 Detection method based on density
The density-based detection method is similar to the distance-based method. The difference is that the density calculation is based on the surrounding density of the research point and the surrounding density in the neighborhood of the point. The relative density is calculated from these two densities, which is called Anomaly score. The larger the value of relative density, the greater the degree of abnormality. The premise of the establishment of this method is that the density of normal points is similar to the points in its neighborhood, and the density of abnormal points is quite different from the surrounding points. The density-based detection method solves the problems that the distance-based detection method cannot solve, such as the anomaly detection problem of certain data sets with different densities. Zhang Bowen proposed an anomaly detection method based on nuclear density fluctuation. According to the actual data set, the nuclear density fluctuation factor was defined and used as the detection index. The experiment proved that the algorithm has good detection effect and good robustness. Great [9].

3.4 Detection method based on deep learning
Deep learning has become a popular algorithm today and has been applied to many fields. Among them, the model used for anomaly detection is the autoencoder (Autoencoder), which is composed of two parts: an encoder and a decoder, as shown in Fig.2.
This model uses the encoder on the left to compress the input high-dimensional data into low-dimensional information. In the compression process, some irrelevant information and noise will be removed by the neural network. The decoder on the right decodes the information compressed by the encoder, and tries to restore it to the input data and output it. The use of deep learning methods to detect anomalies requires a lot of training and testing of the model. Although the early stage is time-consuming, the final detection effect is better. Scholars Bai Mingliang proposed a combination of deep autoencoder and support vector data description for abnormal detection of gas turbine high-temperature components. Experiments have shown that this method significantly improves the detection efficiency [10].

3.5 Comparison of characteristics of model-based methods
The four model-based anomaly detection methods mentioned in the article each have their own characteristics. The characteristics of these four methods are now compared and summarized, as shown in Table 1.

| Model type   | Theoretical basis                                                                 | Applicable scene                        | Representative algorithm |
|--------------|---------------------------------------------------------------------------------|----------------------------------------|--------------------------|
| Based on depth | The proportion of abnormal data points is small and the feature values of abnormal data points are very different from normal points | Financial transaction fraud, data cleaning, etc. | Isolated forest          |
| Based on distance | There are more spots around normal points, and few spots around abnormal points | Time series detection and consumption behavior analysis | K-means algorithm        |
| Based on density | The density of normal points is similar to that of its neighbors, and the density of abnormal points is quite different from that of surrounding points. Use the autoencoder to continuously train and test the optimization model to convert anomaly detection into a classification problem | Single or multi-dimensional data detection | DBSCAN algorithm         |
| Deep learning      |                                                                                   | Single and multiple classification problems | Neural network method     |

It can be seen from the table that the first three model methods have their theoretical assumptions and have certain limitations when used. The deep learning method is to feed data to the autoencoder model to convert the anomaly detection problem into a single classification or multi-classification problems, labeling is required in the early stage of model training, which requires more training times, but it is more convenient and quicker to perform anomaly detection after training.
4. Development Trends and Conclusion

4.1 Development Trend
There are more and more researches on anomaly detection. The topics of master's thesis published on CNKI in the past four years (2017-2020) contain related papers on "anomaly detection". Among them, 2017 There are 162 master's thesis on anomaly detection, 194 master's thesis on anomaly detection in 2018, 223 master's thesis on anomaly detection in 2019, and 249 master's thesis on anomaly detection in 2020. According to the statistical results:

(1) From 2017 to 2020, the number of papers has been increasing every year, indicating that anomaly detection has become a current research hotspot and has received widespread attention.

(2) Among the above-mentioned statistical master's theses, there are papers on anomaly detection that have touched on all aspects of life and industry, reflecting the theoretical value and application value of anomaly detection, and explaining the importance of anomaly detection from the side.

(3) Anomaly detection is a guarantee for information security. In the era of increasingly powerful artificial intelligence, information security is increasingly important. In future research, there will be more research on anomaly detection and a wider range of fields.

4.2 Conclusion
Anomaly detection is particularly important for people’s industrial production and daily life. The article lists four anomaly detection methods based on depth, distance, density, and deep learning, and explains the principles of these four methods, and each The comparison of the characteristics of the methods provides a reference for scholars who follow-up research on anomaly detection. It can be seen from the number of master's degree thesis counted on CNKI in the past four years that anomaly detection is still a research hotspot, and anomaly detection methods based on deep learning models are expected to become the mainstream of future research.

References
[1] Li Yuan, Xie Yizhen, Wang Yongjian, Jiang Hong. Traffic anomaly detection method for car network flooding attacks[J]. Journal of Nanjing University of Science and Technology, 2020, 44(04): 454-461.
[2] Chen Yan. Detection of network traffic anomaly of ship network server[J]. Ship Science and Technology, 2021, 43(04): 187-189.461.
[3] Kang Yanyun, Meng Fanyu, Feng Yongxin. A network traffic anomaly detection model for military Internet of Things[J]. Fire Power and Command Control, 2021, 46(02): 120-125+132.
[4] D. M. Hawkins. Identification of outliers[M]. Springer, 1980.
[5] M. Ahmed, A. N. Mahmood, J. Hu. A survey of network anomaly detection techniques[J]. Journal of Network and Computer Applications, 2016, 60: 19-31.
[6] J. Lin, E. Keogh, A. Fu, et al. Approximations to magic: Finding unusual medical time series[C]. 18th IEEE Symposium on Computer-Based Medical Systems (CBMS’05), 2005, 329-334.
[7] Wang Jingxian, Wu Jing, Long Chengnian, Li Shaoyuan. Abnormal power data detection based on deep autoencoder[A]. Chinese Society of Automation. Proceedings of China Automation Conference 2020 (CAC2020) [C]. Chinese Society of Automation: Chinese Society of Automation, 2020: 6.
[8] Sun Yuhao, Li Guotong, Zhang Ge. Satellite anomaly detection method based on distance correlation coefficient fusion GPR model[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(04): 844-852.
[9] Zhang Bowen, Liu Zhi, Sang Guoming. Anomaly detection algorithm based on nuclear density fluctuation[J]. Computer Engineering and Applications, 2021, 57(12): 132-136.
[10] Bai Mingliang, Zhang Dongxue, Liu Jinfu, Liu Jiao, Yu Daren. Anomaly detection of gas turbine high temperature components based on deep autoencoder and support vector data description[J]. Power Generation Technology, 2021, 42(04): 422-430.