Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/27/20230148

A review on machine learning methods for intrusion detection system

Man Ni

School of advanced technology, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China

man.ni20@student.xjtlu.edu.cn

Abstract. With the increasing access to the Internet and the development of information technology, concerns about computer security have been raised on a considerably large scale. Computer crimes contain various methods to undermine information privacy and system integrity, causing millions to trillions lose in the past few years. It is urgent to improve the security algorithms and models to perform as a thorough structure to prevent attacks. Among this prevention structure, an intrusion detection system (IDS) has played a vital role to monitor and detect malicious behaviours. However, due to the rapidly increasing variety of threats, the traditional algorithms are not sufficient, and new methods should be brought into IDS to improve the functionality. Deep learning (DL) and Machine learning (ML) are newly developed programs which can process data on a considerably large scale. They can also make decisions and predictions without specific programming, and these features are suitable to improve and enhance the IDS. This article mainly focuses on a review of ML methods used in IDS construction.

Keywords: machine learning, deep learning, IDS, intrusion detection.

1. Introduction

In the past decades, the development and usage of the Internet is increasingly growing. In 2022, the Internet has been widely used by 82% of the urban resistance World, and the percentage of usage in the rural area increased from 31% in 2019 to 46% in 2022 [1]. The Internet has possessed a dramatically large amount of utilization contemporarily. With the growth of the number of Internet users and devices, the damage caused by cybercrime becomes more catastrophic as well. The use of a computer or the Internet for criminal purposes is defined as cybercrime, and the commonly used attacks involve malware blackmail, invasion of privacy, identity theft, denial of service (DoS) and so on. In the field of economics, damages caused by cybercrimes are raised from $2 trillion in 2015 to $6 trillion annually by 2021. Cyberattacks were nominated by the World Economic Forum (WEF) as the third global risk for 2018, this status would not decline in the future [2].

Under the threats mentioned before, the requirement for reliable protection systems is acute and reasonable. On the operational level, Intrusion Detection System (IDS) is a complement to the firewall, whose main purpose is to detect malicious invasion which has a high possibility to undermine the functions of a network or system [3]. IDS can be categorized into two types in terms of deployment position: network-based IDS (NIDS) and host-based IDS (HIDS). While HIDS performs like a

DOI: 10.54254/2755-2721/27/20230148

stationing on a single complex like the host of one personal computer, NIDS can supervise the whole traffic from the network, including all the packets that pass through. Therefore, with regard to a large-scale network, the NIDS is more customarily used. Detection is fundamental to both types of IDS, and in respective of mechanism, the methods employed for identification can be sorted as two classifications: anomaly-based detection system and signature-based detection system. The main dissimilarity between them is that signature-based detection discovers known malicious intrusions, but anomaly-based detection can uncover new intrusions by diagnosing anything abnormal in the traffic. The definition of anomaly is based on the offset attributes from the baseline. Since supervisions by the means of signature detection depends on the updated database with malicious behaviour signatures recorded, it is rather lagging and not adequate to cover all the intrusions, especially the novel ones. Therefore, the implementation of anomaly-based detection is decidedly crucial. The root mechanism of anomaly-based detection is to build a baseline profile which symbolizes the behaviours outside the alerting range, by studying the traffic from the network. After this studying, anomaly-based IDS is well prepared to detect and supervise by juxtaposing the normal baseline and the current traffic [3].

However, since the data capacity of the Internet is excessively large, and the refreshing is rapid, it is impossible to manually set every baseline profile for anomaly-based detection. Therefore, the demand for an efficacious self-learning algorithm sprang. Machine learning (ML) performs as an algorithm consisting model included to artificial intelligence (AI), which is intended for data training, constructing models that are able to make selections and decisions by themselves, rather than depending on specific program commands [4]. ML has already manifested as a formidable approach in several domains, such as consumer services, control of logistics chains, biology and computer science [5]. Deep Learning (DL) is one sub-branch of ML, and adopting this algorithm to anomaly-based IDS has become a tren [6]. A huge amount of data is reckoned with by the employment of ML to perform supervising and detection of the network traffic [7]. Applying ML to early detections in the cyber security field is an effective way to reveal new attacks [8]. From 2009 to 2014, ML models are often used on detecting attacks from cloud security, malware, and malicious intrusions. After the raise of DL in 2013, this trend increased at a considerably large rate. And up to 2020, the usage of ML and DL is comprehensive in the computer security domain [9].

In this article, the main objective is to review deep learning and machine learning employed in intrusion detection systems. Section I aims to explain the concept and classification of IDS, the signature-based and anomaly-based IDS are addressed. And section II demonstrates the principles of deep learning and machine learning, with several frequently used techniques in detail. The proposed applications and innovations of machine learning to IDS are presented in section III, and section IV delivers the challenges met by the machine learning-based IDS. Finally, the conclusion contains a brief summary of the whole article.

2. Section I Intrusion detection system

The techniques of detection exist as the fundamental law of the intrusion detection system. Two types of IDS exist now in terms of detection mechanism: anomaly-based intrusion detection systems and signature-based intrusion detection systems. And AIDS and SIDS are the abbreviations of them, respectively. In this section, the mechanisms, and differences between these two detections will be introduced. And for each type, both advantages and disadvantages are presented.

2.1. Signature-based intrusion detection system

The term signature represents a pattern typifying one threat that has already been detected and recorded in the dataset. And the root mechanism of SIDS is to contrast the current network traffic with the malicious signatures in the dataset. If the patterns match, the alert would become positive, and the malicious behaviour can be alerted and blocked. However, the pattern must match with high precision, which means SIDS cannot detect an unknown threat, because every time SIDS performs as the monitor, it needs to go through the whole database to find if the traffic is malicious [3].

DOI: 10.54254/2755-2721/27/20230148

Since the comparison needs high precision, SIDS performs with high accuracy when detecting known threats, and this is the reason SIDS has been broadly deployed in the Internet of Things industry and many other fields [10]. The advantages of SIDS apparently focus on two aspects: wide employment and simple but effective detection of known attacks. But the disadvantages of SIDS are also ostensible. The lack of flexibility when detecting a new threat leads to a huge false negative ratio. And on account of the growing size of the malicious signatures, the search speed of one signature becomes slow and the requirement for updating is increasing. The most distinct challenge SIDS faces is to keep the database updated and arranging the limitation profiles of SIDS as a suitable manner [11].

2.2. Anomaly-based detection

Another type of IDS is AIDS. AIDS is based on anomaly which represents anything outsides the normal baseline in traffic, such as tremendous Telnet sessions within a short time interval and Heavy SNMP traffic. Before detection, AIDS would first study normal traffic in different time slots, and a baseline profile based on the studying would be produced. After the study, AIDS would begin to monitor the traffic with the profile as the standard. Besides the normal traffic, the AIDS profile can also be created based on a specific behaviour, such as the amount of user access workouts. But not like SIDS, this kind of behaviour is not a malicious one. The alert would be triggered once any abnormality is detected [3].

Compared to SIDS, AIDS satisfy the disadvantage of being unable to detect new threats, because, after the study, AIDS is able to detect a new malicious behaviour as long as this behaviour outranges normalcy. On the other hand, SIDS can only notify the specific malicious behaviour itself, like internet worms and viruses. But some other behaviours like Dos and buffer overflow, which cannot be represented by a signature, would misguide the SIDS. In this way, another advantage of AIDS is to notify behaviours like overflow by catching anomalies beyond the baseline. However, AIDS also has several disadvantages and the affirmation of the baseline profile is the predominant challenge. It is unachievable to ensure that every time the AIDS study, the traffic is at normal status. So, if when the traffic is an abnormal one during the study period, the baseline is meaningless and the whole AIDS would be devastated. And the process of studying is also time costing [3].

3. Section II Deep learning and machine learning

The basic principles of ML and DL would be covered and several techniques in particular support vector machine, decision trees, and neural networks would be explained with details in this section.

3.1. Machine learning and deep learning basics

Artificial intelligence (AI) comprises Machine learning (ML) , and the intention of ML is to train computers to make decisions and predictions without being explicitly programmed, by the way of going through a huge set of data, known as datasets [12]. ML can be ranked into three types: semi-supervised, supervised and unsupervised machine learning. The major difference is whether the target is known. If the targeted label and classes are not clear, it belongs to unsupervised ML, otherwise, it is supervised ML. And for semi-supervised ML, only a part of the data is labelled or that human intervention is needed during the training [13].

Among the subsets of ML, deep learning (DL) performs as a efficacious program. Compared to ML, DL has a larger set of datasets, and one surpassing feature is its data-hungry characteristic [14]. In this way, DL is more like a complement of ML, when the requirement of the dataset's size is far larger. However, ML and DL still share similar techniques and leading objectives, and they are both widely used in the computer security domain. For example, in 5G networks, they have both played an important role in improving anomaly-based detection [15]. Some practice experiments and innovations of ML and DL used on IDS is further discussed in the next section.

DOI: 10.54254/2755-2721/27/20230148

3.2. Some ML methods and techniques

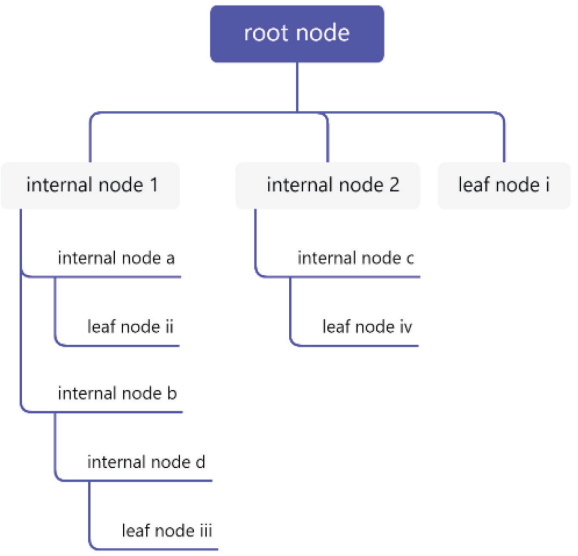

Figure 1. The construction of Decision Tree algoritm.

3.2.1. Decision tree. The decision tree is a construction which includes nodes and branches, and the target is to finalize decisions by going through steps such as splitting, stopping, and pruning. In the figure, every node is represented by a name, and regarding to the deployment location they are categorized into three types: root node, internal node and end node. The root node, or a decision node, represents a selection made to cause two or more nodes as its sub-nodes. The internal node or chance node is the node between the other two nodes, representing one possible choice. And the leaf node is also called an end node, representing one of the end decisions made by the decision tree. Every pathway from the decision node to the end node in the decision tree construction is one branch, which represents one outcome from the decision tree. The 'if-then' rule is one way to embody these pathways [16]. There are several important decision tree algorithms that have already been used in practice, in particular C4.5, ID3 and CART [17].

3.2.2. Neural networks. Artificial neural network, or ANN, performs as a simulation based on the biological nervous systems of a human brain, and the applications are already in practice, such as intrusion detection, data classification and optimization method. ANN has a high accuracy one information management and has the ability to handle a large scale of inputs, just like the human brain. A large number of elements perform like neurons in an interconnected way, to achieve specific solutions [18]. DL is actually ANN with complex multilayers, which link with each other [19]. This characteristic

DOI: 10.54254/2755-2721/27/20230148

enables DL to process information with broad variables and layers but in a unique basic network architecture.

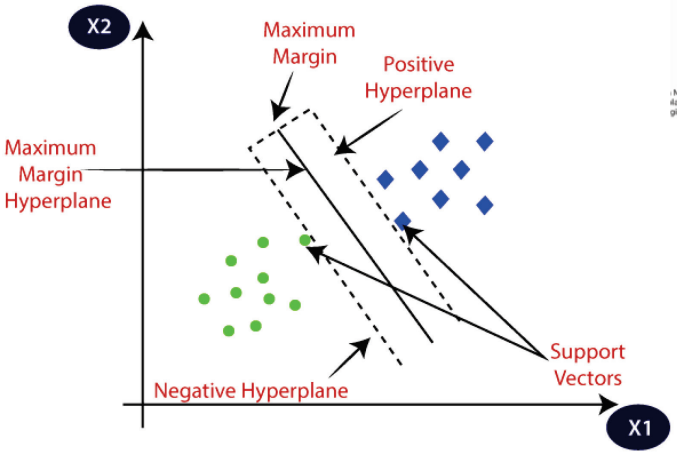

Figure 2. SVM.

3.2.3. SVM. As it is shown in figure2, the green and blue points are data points belonging to separate classes, and the two support vectors are the closest data points between the two classes. Locating the finest separating hyperplane between the two separate classes is the main intention if this algorithm, by the ways of maximizing the margin between the support vectors. And the Support Vector Machine, shorted as SVM, is a program with the capability of solve this problem [20]. For different kernel functions, SVM can be classified into linear and non-linear types. With regards of the detection modes, two forms one-class and multi-class are included [21, 22]. A huge amount of time is essential to train SVM as well.

4. Section III Review of research on ML-based IDS

Since in cyberinfrastructure, a huge set of data is included, it is critical for IDS to learn the signatures and behaviours within this large set of data. Given ML and DL's powerful ability to study, they have played a vital role in decision-making and prediction generation in IDS [23]. In this section, several techniques and methods of ML and DL used in IDS are introduced and reviewed.

4.1. Neural network and decision tree used on IDS

Neural networks and decision trees are often used in SIDS previously [24], but this article applied these two methods to both SIDS and AIDS and therefore, they succeed to make an enhancement of IDS. The neural network was used to enhance the detection of known attacks, and a decision tree was used to detect new attacks. Backpropagation was a learning algorithm from a neural network, which was used in the three-layered structure. On different KDD 99 data sets, they performed two models on network-based IDS and AIDS. There were four attack classes in the KDD 99 contest: DoS, probing, U2R and R2L appended with one normal class, and they were represented by five neurons on the output layer respectively.

After the experiment, although the neural network outperformed other works, it failed to distinguish U2R and R2L attacks, while the other two attacks are detected. Therefore, the decision tree is used to enhance it. With the aim of detect new threats, they enhanced the C4.5 program and 60.96% was the percentage of detection rate that U2R class was enhanced. And the ratio of the false negative detections of this class diminished from 82.89% to 21, 93%. But the lack of the percentage of successful prediction rates revealed that the improved C4.5 algorithm is not suitable to processing cases by relying on the proper labels.

DOI: 10.54254/2755-2721/27/20230148

4.2. SVM used on IDS

Proposed a method to apply support vector machine (SVM) into network-based IDS [25]. Numerous Kernel functions deployed with parameter C values of stabilization were used in this experiment under KDD 99 dataset, and several features were analysed by using SVM. Firstly, the training dataset and selecting test dataset were performed to enhance the outcome of SVM IDS. And then, the pre-processing is performed to convert the dataset from only labelled as normal or attack name, to the criteria formula of SVM input. There are 4,898,431 sets under training in total, and one to the tenth of the labelled test set is 311,029 instances.

After the pre-processing and training, the SVM learning and validation experiments can be performed. In terms of kernel functions, functions like linear, 2poly, and RBF were used and for the binary classification, SVMlight was deployed. Through the training process, SVM became a decision model, and several iterations are performed until a giant successful ratio are gained. Abundant kernel functions with C values were used to find the most suitable kernel, such as Radial Basis Functions , linear and 2-poly. During the test processes, the obtained results were compared with the KDD'99 results to verify the effect.

5. Section IV Challenges and predictions

The applications of ML and DL in IDS have been shown in the previous article. Meanwhile, some disadvantages and issues are exposed as well, and the challenges would be discussed in this section.

5.1. Dataset

Several frequently used datasets such as KDD 99, NSL-KDD, ISOT and so on, are actually archaic nowadays. For example, KDD 99 is used for data training in the two pieces of research mentioned in section III, but the results of KDD 99 CUP were published in 2000, which means this dataset is constructed over 20 years ago [26]. And it is possible that this dataset is redundant and not able to cope with the new attacks. On the other hand, the datasets that covers the newest malicious behaviors are private at present but the public datasets often have redundant and anonymous attributes, therefore, various issues exsists [23]. Therefore, valid and open datasets are required.

5.2. Standard metrics

There are no universal standard metrics for evaluations during the experiments, and most researchers are performing evaluations using different parameters. For the experiment of [25], KDD'99 results were compared with the experiment's results to evaluate the effect. Therefore, improvements to meet the demand of an settled standard of evaluation for the model's comparison is promising.

5.3. Real environment problem

During the data training for ML, the datasets used are often labelled with attack types. But in the real world, it is impossible to cover all the types of data with labels to train the ML algorithm, and the training would be insufficient as usual. Unfortunately, ML has a low detection accuracy when training is not enough. So, the gap between testing and the actual environment may cause inferior detection. Besides, if a sufficient dataset is reached during the training, the problem of efficiency may occur. Since real-time IDS is often needed but the complicated models and big sets of data would slow down the speed [27]. However, it can be improved by using a new format to reduce the size [28].

6. Conclusion

In this article, the mechanisms of two types of IDS are introduced in the first section, with the comparison, advantages and disadvantages in detail. And section II focuses on the general concepts of deep learning and machine learning, and three basic approches of machine learning are explained and inspected. Based on the previous two sections, the third section included the applications and innovations from two pieces of research. And finally, the challenges exposed are considered and several urgent requirements are mentioned as well.

DOI: 10.54254/2755-2721/27/20230148

References

[1] ITU. (n.d.). Internet use in urban and rural areas. Retrieved March 2, 2023, from https://www.itu.int/itu-d/reports/statistics/2022/11/24/ff22-internet-use-in-urban-and-rural-areas/

[2] Nguyen, T. (2023, January 6). A review of Cyber Crime. Retrieved March 3, 2023, from https://dzarc.com/social/article/view/244

[3] Rao, U., & Nayak, U. (1970, January 01). Intrusion detection and prevention systems. Retrieved March 3, 2023, from https://link.springer.com/chapter/10.1007/978-1-4302-6383-8_11#Abs1

[4] Dua, S., & Du, X. (2011). Data Mining and machine learning in Cybersecurity. Boca Raton, FL: CRC Press.

[5] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, Perspectives, and prospects. Science, 349(6245), 255-260. doi:10.1126/science.aaa8415

[6] Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., & Ahmad, F. (2020). Network intrusion detection system: A systematic study of machine learning and Deep Learning Approaches. Transactions on Emerging Telecommunications Technologies, 32(1). doi:10.1002/ett.4150

[7] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. doi:10.1038/nature14539

[8] Fraley, J. B., & Cannady, J. (2017). The promise of machine learning in Cybersecurity. SoutheastCon 2017. doi:10.1109/secon.2017.7925283

[9] Prasad, R., & Rohokale, V. (2019). Artificial Intelligence and machine learning in cyber security. Springer Series in Wireless Technology, 231-247. doi:10.1007/978-3-030-31703-4_16

[10] Ioulianou, P., Vassilakis, V., Moscholios, I., & Logothetis, M. (2018, August 31). A signature-based intrusion detection system for the internet of things. Retrieved March 3, 2023, from https://www.ieice.org/publications/proceedings/summary.php?iconf=ICTF&session_num=SESSION02&number=SESSION02_3&year=2018

[11] Folorunso, O., Ayo, F. E., & Babalola, Y. E. (2016). CA-NIDS: A network intrusion detection system using combinatorial algorithm approach. Journal of Information Privacy and Security, 12(4), 181-196. doi:10.1080/15536548.2016.1257680

[12] Hamid, Y., Sugumaran, M., & Journaux, L. (2016). Machine learning techniques for intrusion detection. Proceedings of the International Conference on Informatics and Analytics. doi:10.1145/2980258.2980378

[13] Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153-1176. doi:10.1109/comst.2015.2494502

[14] Purushotham, S., Meng, C., Che, Z., & Liu, Y. (2018). Benchmarking deep learning models on large healthcare datasets. Journal of Biomedical Informatics, 83, 112-134. doi: 10.1016/j.jbi.2018.04.007

[15] Fernandez Maimo, L., Perales Gomez, A. L., Garcia Clemente, F. J., Gil Perez, M., & Martinez Perez, G. (2018). A self-adaptive deep learning-based system for anomaly detection in 5G networks. IEEE Access, 6, 7700-7712. doi:10.1109/access.2018.2803446

[16] Song, Y., & Lu, Y. (2015, April 25). Decision tree methods: Applications for classification and prediction. Retrieved March 3, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/

[17] Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. International Journal of Science and Research (IJSR), 5(4), 2094-2097.

[18] Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in Artificial Neural Network Applications: A survey. Heliyon, 4(11). doi:10.1016/j.heliyon.2018.e00938

[19] Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional

DOI: 10.54254/2755-2721/27/20230148

neural network. 2017 International Conference on Engineering and Technology (ICET). doi:10.1109/icengtechnol.2017.8308186

[20] Meyer, D., & Wien, F. T. (2015). Support vector machines. The Interface to libsvm in package e1071, 28, 20.

[21] Chen, W. H., Hsu, S. H., & Shen, H. P. (2005). Application of SVM and ANN for intrusion detection. Computers & Operations Research, 32(10), 2617-2634.

[22] Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J., & Platt, J. (1999). Support vector method for novelty detection. Advances in neural information processing systems, 12.

[23] Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., & Xu, M. (2020). A survey on machine learning techniques for cyber security in the last decade. IEEE Access, 8, 222310-222354. doi:10.1109/access.2020.3041951

[24] Bouzida, Y., & Cuppens, F. (2006, September). Neural networks vs. decision trees for intrusion detection. In IEEE/IST workshop on monitoring, attack detection and mitigation (MonAM) (Vol. 28, p. 29).

[25] Kim, D. S., & Park, J. S. (2003). Network-based intrusion detection with support Vector Machines. Information Networking, 747-756. doi:10.1007/978-3-540-45235-5_73

[26] Elkan, C. (2000). Results of the KDD'99 classifier learning. Acm Sigkdd Explorations Newsletter, 1(2), 63-64.

[27] Liu, H., & Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. Applied Sciences, 9(20), 4396. doi:10.3390/app9204396

[28] Cococcioni, M., Rossi, F., Ruffaldi, E., & Saponara, S. (2019). Novel arithmetics to accelerate machine learning classifiers in autonomous driving applications. 2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS). doi:10.1109/icecs46596.2019.8965031