2015 The 5th International Workshop on Computer Science and Engineering
doi: 10.18178/wcse.2015.04.121

A Comprehensive Survey on Network Intrusion Detection Techniques

Qazi Emad-ul-Haq[1,a], Hatim Aboalsamh[1], Jalal Al-Muhtadi[1], Muhammad Hussain[1], Wadood Abdul[2], Sanaa Ghouzali[3] and Saeed Bamatraf[1]

[1]Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

[2]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

[3]Department of Information Technology, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

[a]emadhaq@student.ksu.edu.sa

Keywords: Network Intrusion Detection System, Information Security, Signature Patterns, Machine learning, Data mining.

Abstract: In recent years, network attacks have become a problem of utmost importance which greatly affects the availability of information, security and the correctness for the individuals as well as businesses and government corporations. The problem has increased manifold with the widespread use of the internet and due to the need to provide services to the end users at all times. I it is impossible to develop foolproof systems that can stop intrusions but solutions exist to meet the challenges regarding information security and its integrity. One of the solutions is to install a firewall but it is not a fail-safe system since it cannot block the ports which remain open in order to provide standard network services to the users. Therefore, it cannot completely guard the network against the intrusion attempts. Another solution is to employ a network intrusion detection system (NIDS) along with a firewall, which can successfully detect and stop intrusion detection attempts. In this paper, a survey of network intrusion detection systems is presented. Various aspects regarding the deployment of these systems in real life scenarios are analyzed and the main factors which are responsible for the non-deployment of some of these innovative methods for practical use today are identified.

Introduction

A network intrusion detection system is software that automatically detects network intrusion detection attempts. It can be regarded as a data classification problem, i.e., to classify any network traffic event into two classes; either normal or abnormal, i.e., of an intrusive nature. Lately, many techniques have been proposed by the research community but by far the most commonly used approach is based on pattern signatures which define the behaviour of the known attacks [1]. In this approach, a monitored network traffic event is regarded as intrusion if it matches to the existing signature patterns. Although the signature patterns based NIDS are most commonly used and are considered as standard, they have shortcoming too. One of the most prominent is that a new signature pattern is required for every new threat that comes into the light. Another problem is that experts are required to write the signature patterns, therefore, the dependability on the experts and hence their availability is a serious issue. Therefore the main cause of concern in this approach is the absence of required automation level.

In order to get the higher automation level in NIDS, techniques based on machine learning, data mining and statistical methods have been proposed in the research community. These techniques not only provide a higher level of automation but more accuracy as well. Despite the fact that these techniques provide more useful results and help the security personnel in higher level automation of the intrusion detection process, these have not been used commonly in practice. It shows

that accuracy is not the only criteria and there are more factors involved in the successful deployment process of NIDS.

In this paper, a review of the network intrusion detection techniques is presented along with the factors that influence in successful deployment of network intrusion detection techniques. The paper is organized as follows; section 2 describes some background concepts, section 3 explains the most prominent techniques in network intrusion detection, section 4 contains discussions and lastly section 5 concludes the paper.

Background

In this section the basic concepts related to network intrusion detection are discussed.

Network Intrusion. A network intrusion is defined as an unauthorized attempt to gain access into a secured network to gain illegal or unauthorized access to the information. Network intrusion detection system aims to detect unauthorized attempts to access a network by analyzing network traffic. Lately, many attempts have been made to classify the type of network intrusions. The most well-known is proposed by Kendall [2], which classified the intrusion types into the following four categories.

a) Probing: These are types of intrusions which aim at gathering information in order to achieve further intrusion into the system. Network traffic sniffing and port/address scanning are few of the examples of this type.
b) Denial of Service (DoS) Intrusions: These are the types of intrusions which make the system entirely unusable, e.g., servers are bombarded with so much fake requests that it becomes impossible for them to service genuine user requests.
c) User to Root (U2R): These intrusions aim to gain administrator access rights to the system by taking advantage of security loop holes of the operating systems as well as various software, e.g., web browsers. In this type of intrusions, the hackers usually have normal user account in the system.
d) Remote to Local (R2L): These intrusions aim at getting local access but not from inside the network.

Basic NIDS Architecture. Basic intrusion detection system architecture consists of following components.

a) Network Traffic Collector Component: This component is responsible for network data collection, i.e., data frames or information from upper layer protocols.
b) Features Extractor Component: This component performs the feature extraction from the captured traffic by network traffic collector component. This task of feature extraction is of utmost importance. There are two types of features. The low-level features are obtained directly from the captured traffic, e.g., headers etc. The other type of features, i.e., high-level features is obtained from the low-level features after processing. Features are also classified as how those were generated in the first place, e.g., packet features come directly from the packet headers. Payloads are the features which come from the packet payloads.
c) Detection Component: This component is responsible for detection of unauthorized attempts using the features generated by the features extractor component. The detection processes are generally categorized into two main types. First is misuse detection which tries to find correspondence between features of an instance to the existing intrusion definitions. Second is anomaly detection which takes the target system behavior into account. Once an unauthorized intrusion attempt is detected, a system alert is generated to the response management component for appropriate action.
d) Traffic Model Generator Component: This component is responsible for generating the model which is used by the detection component for comparison purposes. This model can be specified by humans or it can automatically be learned from the data.
e) Response Management Component: This component is responsible for taking appropriate actions in response to the detection of unauthorized intrusions in the system.

Various classifications of NIDS have been proposed in the literature [3]. The two main detection techniques proposed are based on anomaly and misuse. The model for traffic is categorized as human knowledge based or based on some automatic model generation process. The detection process is usually carried out in one of the two modes, i.e., real time or in the batch mode. The architecture is categorized as whether the collection of traffic and subsequent processing is done at a single point, i.e., centralized or at multiple points, i.e., distributed. There are four performance metrics [4] that have been defined in the literature for measuring the deployment performance. The first is accuracy of intrusion detection, 2nd is the time required to process the data, third is adaptability, i.e., how NIDS can cope with the new threats and how it adapts itself to the new threats. The system should be adaptable to the new threats. The last is how much system resources are needed by the system to perform its intended functions.

Network Intrusion Detection Techniques

In recent years, a lot of methods have been proposed in the research community for misuse and anomaly based intrusion detection categories. Patcha and Park [5] and Lazarevic et al. [3] proposed the categories of the proposed methods for each of these categories, which are followed here.

Misuse Based Detection Methods. In these methods a model is used for detection purposes, i.e., features which are obtained from the captured network traffic are compared to the model which contains information about the known existing intrusion threats. The model based methods are highly effective at the detection of known threats but are not useful for the new unknown threats. Pattern recognition, data mining based methods, and implication rules are the types of misuse based detection.

Pattern Recognition Methods. These are the most commonly used methods which are successful in practice and are called signature pattern based methods. Each illegal intrusion is defined by a signature pattern and illegal events are detected when the monitored network traffic match with the signature patterns. The signature pattern based methods have drawbacks too, which includes the requirements of new signature pattern for new threats and experts to write the new signature patterns, etc.

Implication Rules-Based Methods. In these methods rule based expert systems are used to implement event detection module in NDS. A set of rules are defined for various network events and occurrence of series of these events give an indication to the system that some illegal intrusion activity has occurred.

Data Mining Techniques. Data mining techniques are based on automatically model learning/training from the sample data, therefore, eliminating the need for making the models manually. Once a model is learned, it can be used for detection purposes not only on the known attacks but the new unknown threats as well. Artificial neural networks (ANN) are one the most well-known data mining techniques for NIDS.

Liu et al. [10] proposed a technique which is based on principal component analysis and ANN for extracting the features. Kumar and Selvakumar [11] used a back propagation algorithm, which is used for learning the ANN model parameters, in order to perform detection of DoS attacks. Another data mining technique that has shown promising results due to its adaptive nature is evolutionary computation (EC). Specifically the genetic algorithms have been used successfully for learning structure of the model and transformations of the model [6,7]. Their main advantage is the simplicity and for learning the rules automatically instead of writing them manually. It has also been proposed in the research community that the efficiency of the EC techniques can be increased substantially by combining them with fuzzy logic and has shown promising results [6,7].

Besides the above mentioned techniques of ANN, EC and fuzzy logic, many researchers have proposed other data mining techniques for NIDS. [7,8] proposed the use of fuzzy association rules in order to avoid using the large time constraints required for fuzzy EC technique. Ye et al. [25] proposed the use of classification trees in order to recognize the signatures. In [9,10] SVMs have been used for classifying patterns by taking them to infinitely high dimensional space and have shown promising results.

It has also been proposed [10] to use ensemble classification techniques using a number of weak classifiers, i.e., each of their individual performances are slightly better than the random classifiers, and then apply them on the data and then perform majority voting scheme for final decision. It has been observed that using multiple weak classifiers instead of a single strong classifier shows more promising results. In [11], weak classifiers based on ANNs are learned on the input features and then fusion is performed using multiple strategies for comparison purposes.

Anomaly Based Methods. These methods are based on using the profiles defining the legitimate network traffic behaviors. In the first step, the profiles which represent the normal behaviors are created. In the second step analysis on the new traffic is performed. If the behaviors of the new traffic are different from the initially created profiles then it is treated as an anomaly. These methods have their advantages as well as disadvantages. The main advantage is their capability to work efficiently on new unseen threats and their biggest disadvantage is to generate more false alarms. The most commonly used anomaly based methods are based on statistical methods, machine learning as well as data mining.

Statistical Methods. These methods are based on having two profiles, i.e., one is the stored statistical profile and the other one is built and updates on current network traffic. Every time a new traffic is observed, the second profile is updated and then compared to the first profile. In [11], a well-known technique based on statistical methods is proposed which continuously monitors the network traffic for any illegitimate activity. In [12], a technique namely stealthy port scan is proposed and it has two major parts. One is the sensor whose responsibility is to monitor the network traffic and to give anomaly score for various activities. The anomaly score is given according to the observation frequency and is directly proportional to that. If observation frequency of a packet is less then it will have higher anomaly score and vice versa. Various other techniques based on statistical methods have been proposed in [12,13].

Machine Learning Techniques. These methods are based on learning from data approach in which some model parameters are learned in the learning process using the training data for anomaly detection. The learned models are then used to predict the output on unseen data. In [14,15], authors have proposed some very impressive work using machine learning techniques for anomaly detection.

Another useful technique namely packet header detector [15] uses the sample information of packet headers from the earlier traffic to carry out anomaly detection on unseen traffic events. In [16], a technique namely application layer anomaly detection is proposed which works on analyzing TCP traffic instead of working on individual packets. Another similar technique is used in [16] in which probability rules are learned using existing network traffic.

Some of the most important machine learning techniques that have performed well in network intrusion detections are described below.

Decision Tree: Decision tree [17] is a known machine learning technique that has shown good results in NIDS. It has a simple structure containing parent nodes and leave nodes. Root node is the first node which performs testing on the first attribute and then decides which internal nodes to move the input data to. As with all the machine learning techniques, first the learning is performed on the training/sample data and once the tree is trained then it is used to predict on the unseen future data.

Back-Propagation Neural Network. Back propagation based neural networks are one of the most important machine learning techniques in practice today and these have been an active area of research these days, especially with respect to deep learning algorithms. Theoretically these are known to learn any function no matter how complex given sufficient layers with input units are used. Many training algorithms like gradient descent, etc., are used to learn the parameters of the model. They have disadvantages too and the most prominent one is about the learning time which is required to learn the parameters of the model. BPNNs are also known to give promising results in NIDS.

Ripper Rule. It is an efficient rule based learning algorithm which has shown good results. It works in two parts. In the first part, the condition are set to some starting values and in the second

part, some optimization algorithms are used reduce the errors. Like other machine learning based techniques, the rules are trained on training dataset and then testing is performed on the test data.

Bayesian Network. These are the types of graphical models and are used quite often in practice to learn joint probability distributions. They contain nodes which represent random variables, and probabilistic models. The conditional independence among the random variables is determined using directed acyclic graph. Conditional independence is represented by edges; therefore, the nodes that are not connected are conditionally independent. The causal relationship between the features and their labels are learned in the learning process and then those are used to predict on future unseen data.

Naive Bayesian Classifier. The Naive Bayes rule is a simple machine learning technique that works using maximum likelihood method for estimating parameter values. It uses Bayes rule for classification. As always, the parameters of the model are learned on training data which is used to predict on unseen future data.

Radial Basis Function Neural Network. Radial basis functions are linked to so many areas of machine learning, e.g., KNNs, regularization, SVM, ANNs. RBF based ANNs have feed forward structure and use RBFs as their activation functions. These are usually used for function approximations and time series predictions.

Data Mining Techniques. These techniques are also commonly used for anomaly detection based methods. Inductive rule generation algorithms are proposed in [17]. Unsupervised learning approach has also been proposed in literature [17] in which clustering is used for NIDS. SVMs [18, 19] are also used for anomaly detection using unsupervised approach. These are well-known for their promising results in the case of anomalies detection.

Like in misuse based detection methods, ensemble of classifiers, i.e., based on weak classifiers have also been applied in literature for anomaly detection. In [19], an unsupervised multiple classifier system is proposed for anomaly detection. Separate group of services or network protocols are modeled separately by unsupervised classifiers. In [20], author proposed another ensemble based approach, which combines five anomaly detection algorithms.

Discussion

Misuse based detection methods have been the most commonly used approach in NIDS and are the most successful. Many products which are quiet well-known are based on these and are in use for many years. The earlier misuse based products used expert systems but there were many disadvantages, e.g., as the number of threats increased, these expert based products were unable to provide real time detection capabilities. Another disadvantage is that these expert based systems were resource hungry and required a lot of memory and processing power with the increase in rule set, therefore, making the system extremely slower. As the weaknesses of expert based system came into notice, some alternate was needed. As an alternative [20] proposed use of signature patterns for misuse detection, and monitors packets using extremely fast algorithms, therefore, the problem of real-time detection is solved but both the above mentioned approaches are resource hungry and use a lot of resource when these have to monitor huge amount of data traffic in fast networks.

Another key problem with the above mentioned approaches is that these are heavily dependent on humans and lack the required automation. Whenever, new threats arrive, signature patterns that define these threats have to be written by human experts. This dependency on the experts and lack of much needed automation is a major drawback of these approaches due to the reason that it is beyond doubt that all new threats cannot come into the knowledge of the experts and, therefore, limit the capability of the system in tacking new threat. Secondly even if some threats are in knowledge of an expert but still the response time required to write the new signature patterns for the threats cannot be fast enough and, therefore, cause delays.

Misuse detection methods [20,21] based on data mining techniques were introduced to meet the above mentioned challenges by reducing the dependability on human experts by automation of model learning, i.e., the models are learns on existing data and then those are used to predict on the unseen future data. These approaches are quiet useful in detection of the known threats as well as

their new variants. However, it should be noted that these approaches are still not able to detect entirely new type of attacks for which existing training data is not available. There are some more issues involved in data mining based methods. One major issue is the availability of enough training data, i.e., enough data should be available for model learning. Another major issue is the labeling of training data, as the data is huge so labeling that for training purposes needs a lot of time and human experts. Lastly but not the least is that model learning processes are time consuming and difficult to do in real- time.

In recent years, NIDS based on anomaly detection based methods have become quite popular in practice. These methods have the tendency to detect new threats/intrusions which are never encountered before. The statistical based methods [21, 22] do not require prior knowledge of the threats for model creation and are quite capable of working in real-time. These statistical approaches make an assumption of quasi-stationary process, which is not always the case and, therefore, is the reason of high false alarms got by these methods. On the contrary, the machine learning and data mining based techniques are more adaptive to the new changes as they don't make any distribution assumption, therefore, these techniques work well with respect to adaptability. But they also have their disadvantages too, e.g., they have resource intensive model generation and therefore these techniques cannot be used in real-time scenarios. Another disadvantage is that they need huge training data of normal/legitimate network traffic.

Unsupervised anomaly detection approaches [22]. [23, 24, 25] are introduced in the literature in order to tackle the shortcomings of supervised anomaly based detections, i.e., the requirement of large labeled dataset. It is mentioned in [25] that unsupervised based approaches work well if the number of attacks are under 1.5% but as it is apparent this assumption cannot be true in practice.

Conclusion

In the last two decades many intrusion detection approaches have been proposed in the research community. The primary focus of all these approaches was in achieving certain level of automation in the intrusion detection process and, therefore, to enhance security. Since the network traffic is non-stationary in nature, therefore, achieving automation is a difficult target to achieve.

The most commonly used and the successful approaches today are based on pattern signatures defining the known intrusion threats. There are several limitations of these approaches and the most important one is the lack of automation since field experts are required to write the new pattern signatures of new threats and it is practically impossible for these experts to know all the new threats present. In order to deal with the short coming of pattern signature based approaches, new techniques have been proposed in the literature. The most noted are the statistical based methods, machine learning and data mining based techniques. Although these techniques achieve higher level of accuracy, there are still some issues with the products based on these techniques. The main issues are that human factor is still involved and resource usage measures along with better feature selection have not been explored in depth yet.

In short, it would be appropriate to say that most of the approaches that have been discussed in the paper have the ability to perform efficiently and achieve high level of accuracy. But the ultimate goal of full automation has not been met yet and there is a need to address the issues like giving labels to the network traffic and the requirement of the system resources should be explored in greater detail.

Acknowledgement

This research work is supported by the Research Centre of College of Computer & Information Sciences and the Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia. The authors are grateful for this support.

References

[1] M. Roesch, SNORT – lightweight intrusion detection for networks, In: Proceedings of the 13th USENIX conference on system administration, LISA 990. Berkeley, CA, USA: USENIX Association. (1999) 229–38.

[2] K. Kendall, A database of computer attacks for the evaluation of intrusion detection systems, Master's thesis, AAI3006082. (1999).

[3] A. Lazarevic, V. Kumar, J. Srivastava, Intrusion detection: a survey. In: V. Kumar, J. Srivastava, A. Lazarevic, editors. Managing cyber threats. Massive computing, US: Springer. 5 (2005) 19–78.

[4] B. Mukherjee, L. Heberlein, K. Levitt, Network intrusion detection. Netw IEEE. 8 (1994) 26–41.

[5] A. Patcha, J-M. Park, Network anomaly detection with incomplete audit data. Comput Netw. 51 (2007) 3935–55.

[6] U. Lindqvist, P. Porras, Detecting computer and network misuse through the production-based expert system toolset (P-BEST), In: Proceedings of the 1999 IEEE symposium on security and privacy. (1999) 146–61.

[7] PA. Porras, PG. Neumann, EMERALD: event monitoring enabling responses to anomalous live disturbances. In: Proceedings of the 20th National Information Systems Security Conference. (1997) 353–65.

[8] W. Lee, SJ. Stolfo, Data mining approaches for intrusion detection. Proceedings of the 7th conference on USENIX security symposium, Berkeley, CA, USA: USENIX Association. 7 (1998) 6.

[9] J. Cannady, Artificial neural networks for misuse detection. In: National information systems security conference, Arlington, VA, USA; (1998) 368–381.

[10] G. Liu, Z. Yi, S. Yang, A hierarchical intrusion detection model based on the PCA neural networks. Neurocomputing. 70 (2007) 1561–8.

[11] P. Kumar, S. Selvakumar, Distributed denial of service attack detection using an ensemble of neural classifier. Comput Commun. 34 (2011) 1328–41.

[12] I. Ahmad, AB. Abdullah, AS. Alghamdi, Artificial neural network approaches to intrusion detection: a review. In: Proceedings of the 8th WSEAS international conference on telecommunications and informatics. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS). (2009) 200–05.

[13] Q. Xu, W. Pei, L. Yang, Q. Zhao, An intrusion detection approach based on understandable neural network trees. Int J Comput Sci Netw Secur. 6 (2006) 229–34.

[14] A. Abraham, C. Grosan, Evolving intrusion detection systems. In: N. Nedjah, L. Mourelle, A. Abraham, editors. Genetic systems programming. Studies in computational intelligence, Berlin/Heidelberg: Springer. 13 (2006) 57–79.

[15] W. Li, Using genetic algorithm for network intrusion detection. In: Proceedings of the United States department of energy cyber security group 2004 training conference, Kansas City, Kansas, Department of Computer Science and Engineering, Mississippi State University, Mississippi State. (2004) 24–7.

[16] RH. Gong, M. Zulkernine, P. Abolmaesumi, A software implementation of a genetic algorithm based approach to network intrusion detection. In International Conference on

Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and International Workshop on Self-Assembling Wireless Networks.(2005) 246–53.

[17] Z. Bankovic, D. Stepanovic, S. Bojanic, O. Nieto-Taladriz, Improving network security using genetic algorithm approach. Comput Electr Eng. 33 (2007) 438–51.

[18] T. Vollmer, J. Alves-Foss, M. Manic, Autonomous rule creation for intrusion detection. In: IEEE Symposium on Computational Intelligence in Cyber Security (CICS). (2011) 1–8.

[19] J. Gomez, D. Dasgupta, Evolving fuzzy classifiers for intrusion detection. In: Proceedings of the 2002 IEEE Workshop on Information Assurance. New York: IEEE Computer Press. (2002).

[20] S. Bridges, R. Vaughn, Fuzzy data mining and genetic algorithms applied to intrusion detection. In: Proceedings of the 23rd National Information Systems Security Conference, Citeseer, held in Baltimore, MA. (2000) 13–31.

[21] C. Chen, S. Mabu, C. Yue, K. Shimada, K. Hirasawa, Analysis of fuzzy class association rule mining based on genetic network programming. In: ICCAS-SICE. (2009) 3480–84.

[22] MS. Abadeh, H. Mohamadi, J. Habibi, Design and analysis of genetic fuzzy systems for intrusion detection in computer networks. Expert Syst Appl. 38 (2011) 7067–75.

[23] J. Luo, Integrating fuzzy logic with data mining methods for intrusion detection, Master's thesis, Department of Computer Science, Mississippi State University. (1999).

[24] G. Florez, S. Bridges, R. Vaughn, An improved algorithm for fuzzy data mining for intrusion detection. In: Fuzzy information processing society. Proceedings. NAFIPS. 2002 Annual Meeting of the North American. (2002) 457–62.

[25] N. Ye, X. Li, S. Emran, Decision tree for signature recognition and state classification. In: Proceedings of IEEE Systems, Man and Cybernetics Information Assurance and Security Workshop. (2000).