A Survey on Intrusion Detection System based, Dataset and Different Approaches

Palak Namdev[1], Prof. Chetan Gupta[2]
M. Tech. Scholar, Department of CSE, SIRTS, Bhopal, India[1], Assistant Professor, Department of CSE, SIRT, Bhopal, India[2]

Abstract — several researchers suggest an intrusion detection system based on data mining technology, focusing on the shortcomings of current intrusion detection models. It addresses the system's lack of self-adaptability, as well as the risks of misreporting or omission. However, when dealing with large amounts of data, an intrusion detection system based on data mining technology must use an increasing amount of resources, slowing down the detection speed. Another issue emerges when we consider the aforementioned approach: when we establish clusters of the same object for future pattern formation, we do not always get a satisfactory outcome. Because each time the support is different, the pruning value varies according to the support. As a consequence, we must process the rule generation dynamic, which overcomes the aforementioned difficulty and allows us to get various results under various scenarios.

Keywords: IDS,NSL-KDD, Dos, Probe, R2L, U2R.

I. Introduction

Intrusion detecting is the process of continuously monitoring and analyzing events in a computer system or network for signs of intrusion. There are two forms of intrusion detection: misuse intrusion detection and anomaly intrusion detection. Nowadays, there are a variety of threats/attacks that affect network security in a variety of ways. An intrusion detection system is used to guard against such attacks. We spoke about Intrusion Detection Systems and detection algorithms like ALAD, PHAD, and SNORT in this paper. We also spoke about the NAL-KDD dataset, which is a more advanced version of the KDD99 dataset. For the processing of correct results, the suggested technique employs a HIDS-based SVM algorithm.

1. Misuse Detection: To detect intrusions, misuse intrusion detection employs well-defined patterns of assault that take use of flaws in system and application software. These patterns are pre-coded and compared to user activity in order to identify infiltration.

2. Anomaly Detection: The usual usage behaviour patterns are used to identify the intrusion in anomaly intrusion detection. The statistical measures of system characteristics, such as CPU and I/O activity by a certain user or application, are used to create the usual usage patterns. Any divergence from the user's constructed usual behaviour is identified as intrusion. To properly protect the system, we have two options: either avoid threats and vulnerabilities resulting from weaknesses in the operating system and application programmers, or identify them and take steps to prevent them in the future while simultaneously repairing the harm. Writing a perfectly secure system is impossible in practice, and even if it were, it would be incredibly complex and expensive. Intrusion Detection (ID) is the process of gathering and analyzing data from multiple important locations on a computer network or system. It may determine whether the network or system has security policy violations or evidence of an attack. Intrusion Detection System (IDS) [2] is a software and hardware intrusion detection system. To assess intrusion detection, current IDS employ misuse detection and anomaly detection approaches. It can be done both during and after testing.

ID has dynamic, active, real-time, and complicated characteristics. How can you get information about a user's interests from a large amount of data? Many novel concepts and algorithms were proposed by many scholars. Wenke Lee [3] of Columbia University in the United States was the first to suggest the use of data mining techniques in intrusion detection systems. Data mining algorithms-based identification has been a popular topic of study, with numerous theoretical findings [4, 5, 6]. To manage enormous volumes of security audit data and extract security related behaviour, build intrusion detection rules, and establish an anomaly detection model, mostly by employing data categorization, association analysis, and sequential pattern mining. The use of data mining techniques in IDS might help to solve the problem of slow response times and inefficiency. It can adapt to the fast advancement of network intrusion detection technologies, increasing detection efficiency while reducing the need for human involvement.

In most cases, an intrusion detection system has three functional components. A data source, often known as the event generator, is the initial component of an intrusion detection system. Host-based monitors, Network-based monitors, Application-based monitors, and Target-based monitors are the four kinds of data sources. The analysis engine is the second component of an intrusion detection system. This component extracts data from the data source and analyses it for signs of attacks or other policy breaches. One or all of the following analysis methodologies can be used by the analysis engine:

Misuse/Signature-Based Detection: Intrusions that follow well-known patterns of assaults (or signatures) that exploit known software vulnerabilities are detected by this sort of detection engine. The fundamental flaw in this method is that it just searches for known flaws and may not be concerned with identifying potential future incursions.

Anomaly/Statistical Detection: An anomaly detection engine will look for something odd or rare. We look at system event streams and use statistical techniques to look for unusual patterns of activity. The main downsides of this system are that it is quite costly and that it might mistake invasive conduct for regular behaviour due to a lack of data. The response manager is the third component of an intrusion detection system. In simple terms, the response manager will only intervene if inaccuracies (potential intrusion assaults) are discovered on the system, telling someone or anything via a response.

II. Literature Survey

In This Paper [1] the suggested work uses a filter and wrapper based technique using the firefly algorithm in the wrapper to choose the features, which affects the speed of the analysis. With the KDD CUP 99 dataset, the generated features are processed to a C4.5 and Bayesian Networks (BN) based classifier. The results of the experiments suggest that ten characteristics are enough to detect the incursion with enhanced accuracy. When compared to prior work, the suggested work shows potential improvements.

An Intrusion Detection based on Crow Search Optimization method with Adaptive Neuro-Fuzzy Inference System is developed in this study [2]. The intrusion detection system (IDS) is used to identify any irregularities in the network or system. The ANFIS is a hybrid of a fuzzy interference system and an artificial neural network, and the crow search optimization technique is used to improve the ANFIS model's performance. The suggested model's intrusion detection ability was validated using the NSLKDD data set. The findings of intrusion detection using the NSL-KDD dataset outperformed earlier methods.

For the intrusion detection system, Liang et al. [9] developed a hybrid positioning technique based on a multi-agent system. A data collecting module, a data management module, an analysis module, and a response module are all included in this system. The analysis module in this study is implemented using an algorithm for a deep neural network for intruder detection. The findings suggest that deep learning methods are adept at identifying transport-layer threats.

Laftah et al. [10] suggested a modified K-mean technique for generating a high-quality training dataset that significantly enhances classifier performance. The modified K-mean is utilized to generate new tiny training datasets that reflect the whole set of original training data, allowing the intrusion detection system to train classifiers in a fraction of the time.

Ali et al. [11] suggested a hybrid machine learning strategy based on a mix of K-medium clusters and Sequential Minimal Optimization (SMO) classification for detecting network intrusions.

Feng et al. [12] used a machine-learning data classification approach to achieve intrusion detection in the network. The main goal is to classify network activities as routine or abnormal connection records in order to reduce classification mistakes inside a network protocol.

Although numerous categorization models have been created for the network intrusion detector, each has its own strengths and weaknesses, including vector machine approaches.

To detect intrusion networks, Ma T et al. [13] introduced a novel approach called KDSVM, which combined k- mean approaches and learning functionality with a deep neural network (DNN) model and a support vector machine (SVM) classifier. There are two phases to KDSVM. The data set is separated into k subsets as a function of each sample distance from the cluster centers of the k-means technique in the first phase, and the test data set is distant from the same cluster centre in the second step, and input in the DNN model using SVM in the third step.

Yuan et al. [14] suggested a deep learning-based DDoS assault detection method (Deep Defense). Deep learning can extract high-level functions from low-level functions automatically, resulting in strong representation and conclusion. To learn patterns from network traffic sequences and track network assault activity, a repeating deep neural network project is presented. The model outperforms typical machine learning models on the bigger dataset, with an error rate of 2.103 percent compared to 7.517 percent for traditional machine learning approaches.

The model's performance in binary classification and the categorization of distinct classes was investigated by Chuanlong Yin et al. [15]. The result analysis is well suited to constructing a classification model with high accuracy, and its performance in binary and multi-class classification is superior to that of standard classification approaches for machine learning.

Zhao et al. [16] suggested a deep belief network (DBN) and probabilistic neural network-based intrusion detection technique (PNN). First, utilizing DBN's nonlinear learning ability, the raw data is transformed into tiny data while the original data's important features are kept. Second, to boost learning speed, a swarm of particle optimization algorithms is employed to optimize the number of nodes with hidden levels per level. PNN is then used to categories low-dimensional data.

Moustafa et al. [17] created a UNSW-NB15 dataset for intrusion detection. This dataset contains nine different types of current attack modes as well as new standard traffic patterns. It has 49 features to discriminate between normal and abnormal observations, including host-flow and network packet control. In this article, we show the UNSW-NB15 dataset's complexity in three ways. It then goes on to explain the mathematical analysis of results and qualities. Second, it does a typical association analysis. Fifth, five existing classifiers are utilized to measure the complexity in terms of accuracy and false alarm rate (FAR). Experiments show that UNSW- NB15 is more complicated than KDD99, and a new dataset for NIDS evaluation is being proposed.

III. Problem Domain

Some intrusion behaviors are comparable to those seen in regular and other incursion situations. Furthermore, several algorithms, like K-Means, fail to effectively discriminate between incursion and typical cases. LI Yin–huan [7] created an intrusion detection model based on the FP-Growth tree in 2012. Their proposed approach yields positive results in the lab. However, the rule is fixed in the data analysis phase, and if the rule generation is fixed, the data clustering that will be formed for pattern detection is fixed as well. However, in today's climate, we see a significant shift in user or customer behaviour on a daily basis. As a result, data analysis is dependent on random selection or dynamic behaviour. The method through which the pattern identifies change is more precise for analysis. This enables us to identify unusual or unexpected trends and take relevant action. If a record matches one in the rule database, it is classified as a known attack. The Administrator Processing Module's (APM) alarm unit will be activated. A new intrusion detection rule is dynamically added to the Rule Database (RD). We also utilise similar patterns to create clusters, which saves time. This also aids in the removal of redundant and noisy data, lowering the likelihood of false positive data being entered into the association analysis module. This allows for a better connection with the homogenous element. When the item set is frequent, our Dynamic method gives you the option to alter the cluster. As a result, the precision of the search has increased. Anomaly

1. Many algorithms have less accuracy and high False Alarm Rate.
2. Less parameter for the comparison of values to get the accurate result.
3. Compare only few types of attack where we can use more detection method for misuse and anomaly types of detection
4. Few parameters to analyze the result. Eg-accuracy, false alarm rate where we can use more parameters like precision, recall, TP (True Positive), TN (True Negative) etc.
5. Some algorithms have low detection rate.

IV. Propose Methodology

We combined data mining with a clustering algorithm to aggregate or cluster the same set of things in our method. When a dataset is clustered, each point is allocated to one of several clusters, each of which may be distinguished by a single reference point, generally an average of the cluster's points. Partitioning is the process of dividing all points in a dataset into clusters. The categorization of plants or animals into discrete groups or species is one of the most well-known uses of clustering. The primary goal of clustering Landsite data is to minimize the dataset's size and complexity. The coordinates of each point in a cluster are replaced with the coordinates of the cluster's reference point to reduce data. Clustered data takes up a lot less storage space and can be manipulated much faster than raw data. The usefulness of a clustering approach is determined by how well the reference points represent the data and how quickly the algorithm executes. We locate the outcomes according to the set following the clustering strategy; the same set of data is described in the same location, while others are placed in separate spots. We then divided the data into categories based on the IDS's behaviour. We use dynamic rule generation to categories data differently each time, resulting in a more accurate and reliable outcome.

V. Conclusion

On the one hand, the integration of broad study topics like as data mining, and on the other, the younger and continuously expanding Intrusion detection system. One major path for future intrusion detection research is the introduction of data mining techniques to IDS. In order to increase system detection performance, using proper data mining methods and building an IDS model are useful strategies.

VI. References

[1] Selvakumar B, Muneeswaran K, “Firefly algorithm based feature selection for network intrusion detection”, Computers & Security, Volume 81, 2019, Pages 148-155, , https://doi.org/10.1016/j.cose.2018.11.005.

[2] S Manimurugan , Al-qdah Majdi , Mustaffa Mohmmed, C Narmatha , R Varatharajan "Intrusion Detection in Networks using Crow Search Optimization algorithm with Adaptive Neuro-Fuzzy Inference System", "Microprocessors and Microsystems" Elsevier 6 September 2020

[2] QING Si-han, JIANG Jian-chun, MA Heng-tai, etc. “Research on Intrusion Detection Techniques: A Survey”. Communications Journal, pp.19-29, July 2004.

[3]Lee W, Stolfo S, Mok k. “Mining Audit Data to build Intrusion Detection Models”. In Proc. of the International Conference on Knowledge and Data Mining, Aug. 1998.

[4] WU Yu-gang, QIN Yong, SONG Ji-guang, etc. “Research Overview of Intrusion Detection Algorithms based on Association Rules”. Computer Engineering and Design, Vol.32.No.3. Mar. 2011.

[5]Eleazar Eskin, Andrew Amold, Michael Prerau, etc. “A Geometric framework for unsupervised Anomaly Detection: Detecting Intrusion in Unlabeled Data”. Kluwer: Data mining for Security Application (DMSA-2002), 2002.

[6]LI Yang. “Application of K-means Clustering Algorithm in Intrusion Detection”. Computer Engineering, Vol.33No. 14, pp.l54-156. 2007.

[7]LI Yin–huan, “Design of Intrusion Detection Model Based on Data Mining Technology”, 2012 International Conference on Industrial Control and Electronics Engineering.

[8]Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir, “Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification”, 2011 7th International Conference on IT in Asia (CITA).

[9] Chao Liang, Bharanidharan Shanmugam, Sami Azam, Mirjam Jonkman, Friso De Boer, Ganthan Narayansamy, “Intrusion Detection System for Internet of Things based on a Machine Learning approach”, International Conference on Vision Towards Emerging Trends in Communication and Networking, IEEE, 2019.

[10] Wathiq Laftah Al-Yaseen , Zulaiha Ali Othman ,Mohd Zakree Ahmad Nazri, “Multi- Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System”, International Journal in Expert Systems With Applications, Elsevier, 2017.

[11] Saad Mohamed Ali Mohamed Gadal and Rania A. Mokhtar, "Anomaly Detection Approach using Hybrid Algorithm of Data Mining Technique", International Conference on Communication, Control, Computing and Electronics Engineering, IEEE, 2017.

[12] Feng, W., "Mining Network data for Intrusion Cetection through Combining SVMs with Ant Colony Networks", Future Gener. Comput. Syst., 2014, 37, 127–140.

[13] Ma T et al, "A Hybrid Methodologies for Intrusion Detection Based Deep Neural Network with Support Vector Machine and Clustering Technique", International conference on frontier computing. Springer, 2016.

[14] Yuan X, Li C, Li X, "Deep Defense: Identifying D-DoS Attack via Deep Learning", IEEE international conference on smart computing (SMARTCOMP), 2017.

[15] Yin C et al, "A Deep Learning Approach for Intrusion Detection using Recurrent Neural Networks", IEEE Access 5:21954–2196, 2017.

[16] Zhao, G.; Zhang, C.; Zheng, L. "Intrusion Detection Using Deep Belief Network and Probabilistic NeuralNetwork", In Proceedings of the 2017 IEEE International Conference on Computational Science andEngineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC),Guangzhou, China, 21– 24 July 2017; Volume 1, pp. 639–642.

[17] Nour Moustafa & Jill Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 Data Set and the comparison with the KDD99 Data Set", Information Security Journal: A Global Perspective, 2015.