To study of Intrusion Detection System to prevent from R2L, U2R attacks and Improve False Alarm rate in Cyber Infrastructure Using ML and DM

Mr. Ankit Chakrawarti
Research Scholar
Computer Science and Engineering
Rabindranath Tagore University, Bhopal

Dr. Shiv Shakti Shrivastava
Professor
Computer Science and Engineering
Rabindranath Tagore University,Bhopal

Abstract

With the limitation of traditional technologies like firewalls and thanks to the advancement within the era of technologies the network security are on high risk, which further emerges the necessity of latest technologies and more advanced solutions for cyber security. Many Intrusion detection systems which aren't very capable of identifying and classifying the attacks present within the network like DoS(Denial of Service), Probe, U2R(User to Root) and R2L(Remote to Local).This survey paper describes a focused literature survey of machine learning (ML) and data mining (DM) methods for cyber analytics in support of intrusion detection. Short tutorial descriptions of each ML/DM method are provided. Based on the number of citations or the relevance of an emerging method, papers representing each method were identified, read, and summarized. Because data are so important in ML/DM approaches, some well-known cyber data sets utilized in ML/DM are described. The complexity of ML/DM algorithms is addressed, discussion of challenges for using ML/DM for cyber security is presented, and a few recommendations on when to use a given method are provided.

Key Words: IDS, Machine Learning, DM, R2L, U2R, KDD datasets.

1. INTRODUCTION

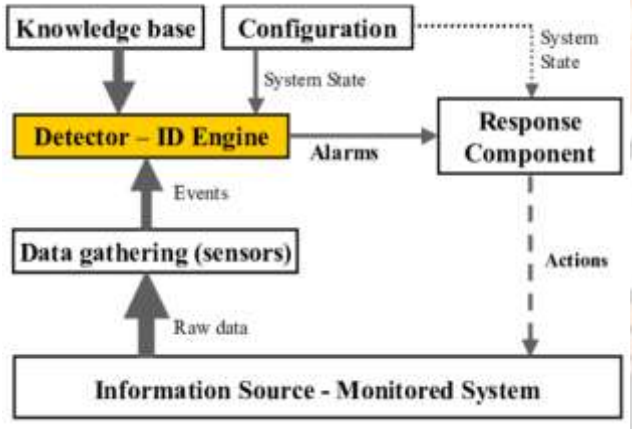
In latest years, cyber safety has been acquired hobby from numerous studies groups with recognize to Intrusion Detection System (IDS). Cyber safety is "a fast-developing subject worrying a first rate deal of interest because of first-rate progresses in social networks, cloud and net technologies, on line banking, cell environment, clever grid, etc." An IDS is a software program that video display units or a community of computer systems from malicious sports (assaults). Detecting an intrusion or prevention (because of enlarge the use of internet), is turning into a vital issue. In past, numerous strategies are proposed to overcome or stumble on intrusion at some point of a community, But maximum of the strategies (used now days in detecting IDS) are not prepared to triumph over this problem (in green manner).Together this, Machine Learning (ML) additionally has been followed in numerous applications, because of supplying suitable accuracy outcomes (in respective domain). Hence, this paintings discusses "How device studying and facts mining are regularly wont to stumble on IDS at some point of a community" in close to future.ML use green strategies like classification, regression, etc., with green outcomes like excessive detection prices, decrease fake alarm prices and much less conversation costs.

Due to the impact of terrific processes in data era and large-scale utilization of conversation and Internet, humans are prompted to switch data the usage of IT-primarily based totally environment. This has benefits and benefits, like trimming the large graphical distance and trade of data with ease. On the opposite side, such form of data switch creates issues like intrusion and malicious sports which could disturb the conversation. The safety control turns into greater hard due to clever hacking strategies. The in advance protection developments of pc networks rely upon desk bound strategies wherein the OS want to be up to date regularly for prevention of safety dumps, and firewalls also are deployed on the extreme community place to enhance the safety. The

dreams of firewalls are to modify and manage the waft of data outside and inside of a community instead of to stumble on whether or not or now no longer the community is beneath neath assault. For balancing easy firewall, the intrusion detection device became used to acquire and examine community logs to be expecting feasible safety threats together with intrusion-like outer corporation assault and misuse detection-like assault from the corporation. IDS has an crucial position in community safety infrastructure to offer crucial safety line, however unfortunately, maximum of those met some of demanding situations like low detection fee and excessive fake alarm fee; those issues are because of the complexity of the threats and feature similarities to the everyday behavior. IDS may be hardware or software program or a mixture of both, accountable to show the intrusion from community log facts. A try of the intruder may be an advanced step to interrupt the device safety and initiated through gathering the data approximately the device just like the used protocol and device to be had on community. Hackers begin to probe each device to investigate unique vulnerabilities; after the vulnerabilities are focused, hackers attempt to get number one manage through remote-to-local (R2L) assault. After consumer get entry to, attacker attempts to get possibility through consumer-to-root (U2R) assault, if attacker receives grasp consumer get entry to it is able to have privilege of stealing or modification, and if the focused device is negotiated/compromised then the attackers have authority to head in addition at this step.

2. Intrusion detection system (IDS)

An intrusion detection device (IDS) constantly video display units the community for any suspicious or malicious pastime and alarms the community administrator, if it detects the sort of form of malicious pastime. Intrusion detection device (IDS) are categorized to host primarily based totally and community primarily based totally. Basic structure of intrusion detection device (IDS). The intrusion detection device (IDS) structure consists of 5 additives 1. Data accumulating tool (sensor) 2. Detector (Intrusion detection (ID) evaluation engine) 3. Knowledge base (database) 4. Configuration tool 5. Response component.



Intrusion detection system (IDS) should have the following characteristic: l. Predictive performance 2. Time performance and 3. Fault tolerance. Attacks in intrusion detection system (IDS) are classified into four types: –

- **DoS:** Denial of service (DoS) is a kind of attack in which a legitimate user does not have access to the system and network resources. Online banking services, email may be affected. DoS attacks comprise of the SYN flood attack and the Smurf attack.
- **R2L**: Remote to Local (R2L) is an attack where an attacker tries to gain access to the victim machine without having an account in it.
- **U2R**: User to Root (U2R) is an attack where an attacker tries to gain privileges having local access in the victim machine.
- **PROBE**: In Probe, the attacker targets the host and tries to get information about the host

3. PREVIOUS WORK

In **[1]**, this study demonstrates an efficient approach to detect and identify network intrusion using machine learning algorithms. Standard data set KDD data set has been used for building the model. Many other algorithms has been used which are effective in the identification and classification of the attack types like denial of service, user to remote etc. Features and values of features play an important role in designing the model as it directly affects the performance and complexity of the model. That reduction in the number of the features increases the accuracy of the model.

In **[2]**,this paper, The authors proposed the exist datasets for the test and assessments of IDSs, and displayed another system to assess datasets with the following attributes: Attack Diversity, Anonymity, Available Protocols, Complete Capture, Complete Interaction, Complete Network Configuration, Complete Traffic, Feature Set, Heterogeneity, Labeled Dataset, and Metadata. Later on, author plan to produce and make new

dataset that will be available to upkeep all the above standards. Different Intrusion Detection Schemes are reviewed in this paper. All the approaches conferred here try to detect intrusion in one way or another. In any case, attackers are equipped for finding new methods and approaches to break security policies. From the literature, it is apparent that various IDS practices rely upon high time, memory and cost prerequisites separated from focal points. Therefore any Intrusion Detection System must have high accuracy, low false positive and false negative rates with low computational, time and cost overheads.

I This paper **[9]** author describe an ML method for R2L to detect attacks. The accuracy of detecting R2L attacks using SVM based on 8 features with payload is 95,99%. The accuracy of detecting R2L attacks using SVM based on 24 Features without payload is 95,73% and the accuracy of detecting R2L attacks using SVM based on 24 Features with payload is 95.91 % better than 8. The accuracy of detecting R2L attacks using SVM based on 28 Features without payload is 95,91% and the accuracy of detecting R2L attacks using SVM based on 28 Features with payload is 96,08% better than all. From the result we can conclude the best feature of 8, 24, and 28 is 28 Feature.

In this paper, explain an adaptive false alarm filter to help reduce the number of false alarms in real deployment of intrusion detection systems. In particular, first compare with the performance of six specific machine learning schemes in a unified platform aiming to point out that false alarm filter should adaptively choose different algorithms based on distinct network contexts. Then propose the architecture of adaptive false alarm filter that has the ability to adapt itself by selecting the best single-performance machine learning algorithm according to the specific contexts. In the evaluation, adaptive false alarm filter keeps up a stable reduction rate of false alarms and filter out more than 80% false alarms. The results show that our approach is effective and encouraging in network settings [7].

In this paper [11], compare in view of the shortcomings of k-means algorithm, an implemented improved k-means algorithm. The k value does not need to be determined in advance, and the data set automatically produces an optimal value through clustering. The clustering center does not need to be changed after determination and the whole data set only needs to be scanned once. The efficiency of both the improved k-means algorithm and the clustering effect has improved greatly. The improved k-means algorithm is verified through experiments. The experimental data show that the improved k-means algorithm has detection efficiency and detection ability. Compared with the previous algorithm, the algorithm improves the detection efficiency and detection ability of the system. A maximum improvement in the detection accuracy of 90.57% can be seen for new attack type intrusion using the proposed algorithm. Apply the improved algorithm to the IDS and design an IDS based on data mining, which mainly includes pre-detection modules, feature extraction modules, etc. and conducts intrusion detection experiments on the new IDS. The experimental results show that the proposed system improves the detection efficiency and reduces the false detection rate.

4. Role of Machine Learning and Data Mining

The differences between the info mining, machine earning, and KDD is that KDD is that the method to extract knowledge useful information from the record. Data processing contains algorithms to acknowledge pattern from data. Availability of the many research studies and results informed us that each one KDD processes of DM defines KDD steps (preparation, selection, cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of DM) which used specific algorithms to extract patterns from record. ML and DM are mostly confusing due to similar meanings, in order that they have a meaningful similarity. The founding father of machine learning, Arthur Samuel, describes it as a neighborhood of research which provides the power to find out without being explicitly programmed. ML interacts with learning the pattern recognition and computational learning theory in AI. ML is older than DM. within the recent days, the term data processing is extra popular than its sibling machine learning which may be the rationale for a few scholars to truly highlight their study for data processing than machine learning, so during this study, machine learning and data processing are discussed together**.**

Training, Validation, and Testing Set. In machine learning, the subsequent three steps are involved: training, validation, and testing. A training phase is a component of knowledge for exploration of possible analytical relationships, and therefore the test part is that the portion of knowledge use to work out the strength and efficiency of relationship. just in case of most appropriate classifier, the needed training set is employed to coach the algorithms, and therefore the validation set has the power to match their efficiency and obtain the choice, which one to use, and then, finally the test set gets the performance like accuracy, specificity, and sensitivity. There are not any specific criteria to work out how one could split the dataset; for instance, it might be divided as 50 percent training, 25 percent validation, and 25 percent test. Data processing and machine learning are basically classified into supervised and unsupervised approaches. Figure 1 shows the supervised and unsupervised methods for ML. In unsupervised learning work with no labeled data, the goal for unsupervised learning is to seek out knowledge from unlabeled data. If a neighborhood of knowledge is labeled then the matter is understood as semi supervised learning, and if all data are labeled then the matter is understood as supervised learning

PMML (Predictive Model Markup Language) for DM/ML. PMML gives an approach to research the appliance for description and discussion about the predictive model prepared by data processing and machine learning algorithms. It's supported XML and capabilities for logistic regression, and therefore the feed forward neural network advance version supports NB, KNN, and SVM classifiers.

CISP Data Mining Model. This model provides a glance of life sequence of knowledge supported data processing as shown in Figure 2. This model is predicated on six basic phases. Business understanding: during this earlier step, understanding the plan objectives and its needs from the business viewpoint is developed; after this, reshape the knowledge into data processing problem solution to urge an initial plan scheme. Data understanding: It starts with collecting data processes with technique for assessing the simplification about presentation of the analytical model. Here, data are split once or repeatedly for approximating the danger of every method. Some of the info training set is employed to coach the algorithm, and therefore the remaining data validation set is employed to approximate the danger of the tactic. K-fold cross validation is liable for dividing into k divisions, whenever one among the k subsets is employed because the test set, the remaining k-1 subset is placed together to form a training set, and then, the regular errors across all k trials are calculated

5. THE PROPOSED APPROACH

The aim of this project is to improve the to minimize the false alarms, will make an IDS capable of online learning, handling concept drift and have ability to be customized to suit any environment as well as Reduced the multiple number of alarms. Improve the detection rate for R2L and U2R attacks in any network. To make IDS should be capable of handling skewed class distribution and measure performance and efficiency of implemented intrusion detection system.

6. Data Mining and Machine Learning Techniques for Intrusion Detection

- **Decision Tree**

A decision tree is a tree in which each branch node will represent a choice between several alternatives and each leaf node will represents a decision. A decision tree is commonly used for obtaining information so as to fulfill the purpose of decision making. The decision tree starts from a root node which is there for users to take action. From root node users split each and every node recursively into different nodes according to the decision tree learning algorithm. The final result is a decision tree where each branch represents a possible context of the decision and its outcomes.

- **Naive Bayes**

The Naive Bayes algorithm is actually based on the probability theory, i.e. the Bayesian theorem, and is a simple classification method. It is named naive because it solves problems based on two critical assumptions: it assumes that there are zero hidden components that will affect the process of analyzing and it supposes that the prognostic component is conditionally independent with similar classification. This classifier provides an efficient algorithm for data classification and it represents the promising approaches to the discovery of knowledge.

- **Support Vector Machine**

Support Vector Machine is used for classification which is also a supervised learning method. There are three research papers that have used the Support Vector Machine algorithm as their technique to analyze student's performance to review it thoroughly. Support Vector Machine as their analyzing method because it suited well in small datasets.

Support Vector Machine algorithm has a good ability to perform generalization and is actually found faster than other algorithms. At the same time, the study done by Gray et al (2014) explained that the Support Vector Machine algorithm acquires the highest analyzing accuracy in identifying student's performance (Failing Risk).

- **K-Nearest Neighbors**

K-Nearest Neighbor is one of the simple Machine Learning algorithms based on the Supervised Learning technique. It is also called a lazy learner algorithm because it doesn't learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. K-NN algorithm stores all the available data and classifies a new data points based on the similarity. This means when new data appear then it can be easily classified into a good suite category by using K- NN algorithm.

- **Clustering: -** The clustering approach is used to group the specific set of items which depends on the upload of their characteristics, collecting them with regard to their resemblances. Basic benefit of the clustering technique in intrusion detection is that it has the ability to analyze from audit data without the need of the admin to have obvious detail of different attack classes. Clustering can be categorized into many forms in terms of input data: Hierarchical clustering for the connectivity model, K-mean for the centroid method, distribution environment for the expectation maximization method, DBSCAN (Density-based Spatial Clustering of Application with Noise), and clique for the graph type model. Hendry and Yang worked with a modified density-based method for clustering where its skills and weaknesses were discovered. (The SLCT

(Simple Log file Clustering Tool) is used for the purposed approach, and it is an application used for the offline data mining tool. First, they drive that density-based clustering made the skills to minimize wide dataset nicely as well as clustering could efficiently mark suspicious activity from the normal activity by using the KDD dataset. (Those properties which showed the potential of clustering can make adaptive signatures as the occurrence of attack. For the anomaly and hybrid detection area, Blowers and Williams implemented the DBSCAN technique to collect normal versus anomalous attacks. (The KDD dataset used for implementation and performance gave 98% for attack or non-attacked detection, and moreover the study made a good example for briefing the methods of ML for cyber operations.

- **Random Forest:** A random-forest classifier is created of numerous classification trees. The k th classification tree is a classifier indicated by an unlabeled information vector and an arbitrarily produced vector by variety of arbitrary highlights of the training statistics for each node. The erratically produced vector of various classification trees in the forest are not allied to each other but then are produced by the similar scattering algorithm. For unlabeled information, each tree will offer a prophecy or vote thus naming is finished. There are a lot more algorithms that can be used such as the J48 technique and others.

7. **Analysis for Datasets**

Mostly, studies used DARPA 1998, DARPA 1999, DARPA 2000, or KDD 1999 dataset, and only few used Net flow, DNS, and TCPdump data for the intrusion detection cyber security purpose. The reality is that many papers use the DARPA and KDD dataset to get data that are much time consuming although if we have already had dataset and reuse which can gives use easy comparison of accuracy of different techniques. Limited use of the Net flow shows that it has no rich features like Tcpdump, KDD, and DARPA. Another issue connected to the performance of IDS is the kind of ML/DM method applied and overall system layout. The study gets concerns with many DARPA and KDD dataset and applied them with many kinds of ML algorithms. These researches did not make an IDS but inspected the performance of the ML/DM method in Internet security.

8. CONCLUSION

I This paper, an effort is made to find Our expected outcome will be false alarm filter that has the ability to adapt itself by selecting the best single-performance, machine learning algorithm according to the specific contexts. IDS should capable of online learning, handling any environment. Using machine learning schemes first we compare with the performance of some specific in a unified platform aiming to point out that false alarm filter should adaptively choose different algorithms based on distinct network contexts and it will also helpful for reduced the multiple number of alarms. Applied machine learning technique on IDS with proper mix of feature extraction, feature selection, data transformation, clustering, classification it will help us to recognize and improve R2L and U2R attacks detection.it will also helpful for improving performance of intrusion detection system and give proper response at any network.

REFERENCES

1. Namrata Pandey, Dr. Pawan Kumar Patnaik, Mr. Sargam Gupta**. (2020,** JULY**).** An Implementation of Model using Machine Learning Algorithm for Intrusion Detection System**.** In INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING AND TECHNOLOGY (IRJET). E-ISSN: 2395-0056.

2. Hassan Azwar, Muhammad Murtaz, Mehwish Siddique, Saad Rehman (2011, September). Intrusion Detection in secure network for Cybersecurity systems using Machine Learning and Data Mining. In 2018 IEEE 5th International Conference on Engineering Technologies & Applied Sciences, 22- 23 Nov 2018, Bangkok Thailand.

3. Anna L. Buczak, and Erhan Guven on a Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. In IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 18, NO. 2, SECOND QUARTER 2016.

4. Suad Mohammed Othman1*, Fadl Mutaher Ba-Alwi1 , Nabeel T. Alsohybe1 and Amal Y. Al-Hashida Intrusion detection model using machine learning algorithm on Big Data environmentt. In Othman et al. J Big Data (2018) 5:34, springer Open, 2018.

5. Hamed Alqahtani, Iqbal H. Sarker2 , Asra Kalim, Syed Md. Minhaz Hossain, Sheikh Ikhlaq, and Sohrab Hossain, on Cyber Intrusion Detection Using Machine Learning Classification Techniques. In © Springer Nature Singapore Pte Ltd. 2020 N. Chaubey et al. (Eds.): COMS2 2020, CCIS 1235, pp. 121–131, 2020.

6. Ansam Khraisat, Iqbal Gondal, Peter Vamplew and Joarder Kamruzzaman, on Survey of intrusion detection systems: techniques, datasets and challenges In springer Open,2018.springer Open,2019.

7. Yuxin Meng and Lam-for Kwok on Adaptive False Alarm Filter Using Machine Learning in Intrusion Detection In. Practical Applications of Intelligent Systems, Springer-Verlag Berlin Heidelberg 2011.

8. Jugnesh Kumar1 , S B Goyal2 , Pradeep Bedi3 , Sunil Kumar4 and Ashish Shrivastava on Analysis of Machine Learning Techniques for Detection System for Web Applications Using Data Mining, in ASCI-2020 IOP Conf. Series: Materials Science and Engineering 1099 (2021).

9. Bisyron Wahyudi Masduki, Kalamullah Ramli, Ferry Astika Saputra, Dedy Sugiarto, on Study on Implementation of Machine Learning Methods Combination for Improving Attacks Detection Accuracy on Intrusion Detection System (IDS), in 2015 International Conference on Quality in Research 978-1-4799-6551-9115 ©2015 IEEE.
10. Iqbal H. Sarker, Yoosef B. Abushark , Fawaz Alsolami and Asif Irshad Khan, on 'IntruDTree: A Machine Learning Based Cyber Security Intrusion Detection Model', in www.mdpi.com/journal/symmetry 2020.
11. Yongkuan Zhu, Gurjot Singh Gaba, Fahad M. Almansour, Roobaea Alroobaea, and Mehedi Masud on Application of data mining technology in detecting network intrusion and security maintenance in Journal of Intelligent Systems 2021; 30: 664–6762021 edn., : De Gruyter. (2021).
12. Yajing Wang , Juan Ma , Ashutosh Sharma , Pradeep Kumar Singh , Gurjot Singh Gaba , Mehedi Masud , and Mohammed Baz on An Exhaustive Research on the Application of Intrusion Detection Technology in Computer Network Security in Sensor Networks, : Hindawi Journal of Sensors Volume 2021, Article ID 5558860, 11 pages.
13. Gillala Rekha, Shaveta Malik, Amit Kumar Tyagi, Meghna Manoj Nair on Intrusion Detection in Cyber Security: Role of Machine Learning and Data Mining in Cyber Security in Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 72-81 (2020).
14. M. Nikhil Kumar, K.V.S. Koushik, K. John Sundar on Data Mining and Machine Learning Techniques for Cyber Security Intrusion Detection in International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307. (2018).
15. Sara Mohammadi a , Hamid Mirvaziri a , Mostafa Ghazizadeh-Ahsaeea , Hadis Karimipour on Cyber intrusion detection by combined feature selection algorithm ,in Journal of Information Security and Applications. (2019).
16. ELKHADIR Zyad, ARCHI Taha, BENATTOU Mohammed on Improve R2L Attack Detection using Trimmed PCA in IEEE 978-1-5386-8317-0/19 2019.
17. Bilal Ahmad , Wang Jian, and Zain Anwar Ali, on Role of Machine Learning and Data Mining in Internet Security: Standing State with Future Directions in, Hindawi Journal of Computer Networks and Communications Volume , Article ID 6383145, 10 pages edn.(2018) .
18. Yirui Wu , Dabao Wei, and Jun Feng on Network Attacks Detection Methods Based on Deep Learning Techniques in Hindawi Security and Communication Networks Volume 2020, Article ID 8872923, 17 pages. (2020).