Literature Review and Paper Writing

ECL 301

National Institute of Technology
Uttarakhand

Submitted to:
Dr. Hariharan
Muthusamy

Submitted By:
Ashish Dhyani
BT18ECE005

MACHINE LEARNING IN COMPUTER NETWORKS

ABSTRACT :

The Internet and computer networks have become an important part of

our organizations and everyday life. With the increase in our dependence on computers and communication networks, malicious activities have become increasingly prevalent. Network attacks are an important problem in today's communication environments. The network traffic must be monitored and analyzed to detect malicious activities and attacks to ensure reliable functionality of the networks and security of users' information. Recently, machine learning techniques have been applied toward the detection of network attacks. Machine learning models are able to extract similarities and patterns in the network traffic. Unlike signature based methods, there is no need for manual analyses to extract attack patterns. Applying machine learning algorithms can automatically build predictive models for the detection of network attacks. This dissertation reports an empirical analysis of the usage of machine learning methods for the detection of network attacks.

INTRODUCTION

Intrusion Detection

The rapidly growing progress in Internet-based technology has brought tremendous benefits to our society. Communications and other services that the Internet provideshave transformed our lives in many ways. The Internet has opened up a whole new world of possibilities to access the information. Students and researchers do not need to go to the libraries to collect the information they need anymore. If I take an example of myself,due to the availability of Internet I was able to surf through a sea of information in order to write this review. Nowadays, the information is just a few clicks away from one's computer web browser. Social networking sites have eliminated geographic distance and made it easier to be in contact with family and friends. Online services, such as online shopping, online banking, and online learning, have made all these activities more convenient to do.

While the Internet has made our lives much more convenient, its vulnerabilities and the amount of information communicating over it

generate opportunities for adversaries to perform malicious activities within its infrastructure. Any host connected to the public Internet or even a private network is under constant threat from potential attacks. A lot of threats are created every day by individuals and organizations to attack computer networks to steal private information and data. This information can be very critical and sensitive, such as social security numbers or bank account information. This has created the need for security technologies to secure users' information and provide reliable computer network environments. Network security has become a very important factor for the companies and organizations to consider. In that regard, intrusion detection plays an important role in the detection of attacks and with securing computer networks. Intrusion Detection Systems (IDS) monitor and analyze network systems to detect malicious activities. Even though users benefit from the use of IDS technology, more is needed to detect better obfuscated or more complex attack patterns.

Machine Learning for the Detection of Network Attacks :

Machine learning is a subfield of computer science, which uses pattern recognition and artificial intelligence methods to group and extract behaviors and entities from the data. These previously known patterns and relationships trained by machine learning algorithms can be used to do prediction tasks on new data. With today's technology, machine learning algorithms touch our everyday life by being used in a wide range of applications. Examples from common domains, which machine learning algorithms are extensively used, include product recommendations systems, such asmthe ones used by Amazon and Netflix ; natural language processing , ; spam detection , image recognition and fraud detection .

The most common operational network intrusion detection systems are signaturebased systems . These systems consist of a database of attack signatures. Human experts produce the attack signatures by manually analyzing the attack data. The monitored network traffic is matched against this database to detect malicious activities. Producing attack

signatures is a time consuming and manually intensive task.

Recently, machine learning techniques have been applied to build predictive models for the detection of network attacks . Unlike signature based methods, which need manual analysis by human experts to extract attack patterns, machine learning algorithms are able to automatically extract similarities and patterns in the network data. With more data being produced than the human brain has the capacity to monitor, machine learning analysis provides results that even an army of analyst experts would be unable to accomplish.

With machine learning affecting a lot of aspects in our everyday life, it is necessary to study the usage of its interdisciplinary capabilities in the detection of computer network attacks. Machine learning algorithms can be applied on network data to extract patterns and similarities, which distinguish between normal and attack instances. These trained patterns can be used to build intrusion detection systems for the detection of network attacks. With the bulk of the work being carried out by machine learning, the cybersecurity experts can become more productive by focusing on analytical results from machine learning models in order to get more insight about the current and future threats.

Intrusion Detection :

The basic function of an intrusion detection system is to monitor an activity taking place in a system and to generate an alarm report stating whether an attack is happening or whether everything is normal. Figure 1 depicts this aspect of intrusion detection.

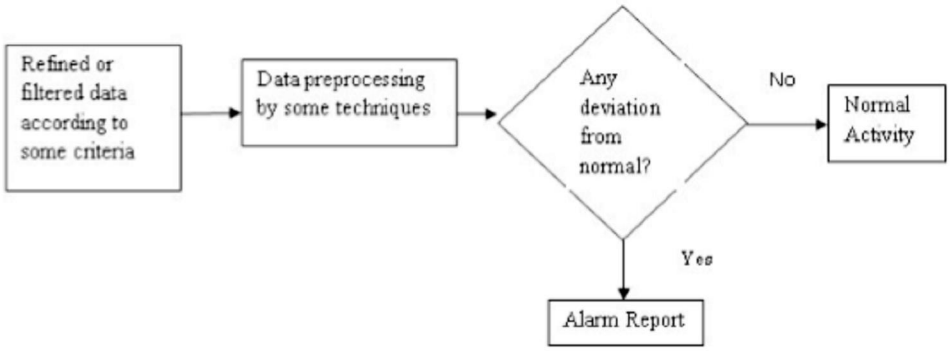

Figure 1. Steps in intrusion detection

Intrusion detection can be classified into two types, namely misuse and anomaly. Misuse detection identifies intruders by comparing them with predefined description of events that denote attack. It matches an event which has already occurred with these predefined events, and any case of deviation is categorized as intrusion. However, this method cannot determine new or unknown attacks. On the other hand, anomaly detection builds models of normal events, and any deviations from these normal events are categorized as attacks.

Technology Types :

Different technologies are applied to detect intruders, but each has its own limitations as summarized in Table 1. The description in the Table 1 reveals that a deeper analysis of the intrusion detection system is necessary. Existing threats and their solutions still need in-depth analysis from the point of view of researchers. A computing system with high computational speed, fault tolerance, and ability to deal with dynamic real time data is necessary. Several works reported successful intrusion detection system based on machine learning techniques which fit perfectly for the aspects like higher computation speed (Rathore et al.,

2016), fault tolerance (Gao et al., 2015), and dealing real time big data (Singh et al., 2014). Different machine learning techniques bear close resemblance to computational statistics methods such as Markov Chain's Monte Carlo method and Kernel density function (Julisch & Dacier, 2002). Due to these similarities of computational statistics, machine learning has attracted researchers' attention to use them in the field of intrusion detection. On the other hand, neural networks become attractive for easy implementation, faster speed of learning, more generalizability and better dealing with nonlinear activation functions and kernels. Due to increasing demand for symbolic representation of problems, symbolic artificial intelligence gains attention of the researchers.

Techniques used to solve intrusion detection	Subsystem	Advantage	Disadvantage
Statistical based system (Julisch & Dacier, 2002)	Univariate	Quick and inexpensive to operate.	Extra attribute handling is difficult.
	Multivariate	Robust	Relationship between variables are complex, hence very difficult in handling.
		Helps establish the relationship between numbers of variables and find out the relationship between them.	For giving the effective result large amount of data is required
Knowledge based or expert based system (Mann & Kaur, 2013; Julisch & Dacier, 2002)	Finite State Machine	It helps in solving the complex system using simple states.	A number of states if large, can be unmanageable.
	Time Series model	It forecasts by converting the nonlinear model to linear model.	Not all packet type especially real-time system packets can be converted to linear model, making the task difficult.
	Expert System	As databases prepared by experts, therefore, all point of view of attributes of dataset are considered.	Human interaction is needed to create the rule.
Data mining technique (Sisodia et al., 2012)	Hierarchical	Less cost to join the clusters	Once the cluster is formed joining two or more clusters, decomposing the cluster is difficult. Performance degrades with noise
	Partitioning	Suitable for dataset where the relationship between attributes are less.	Performance degrades with noise.
		Simple	Reliability on the user to create clusters
	Grid based method	Performance does not degrade with noise.	Clustering is done on the summary of objects and not on an individual object. If any error occurs in an individual object, then the overall result becomes inaccurate.
		Efficient handling of high dimensional data Takes less time	

Table 1. Popular intrusion detection techniques and their advantages and disadvantages

the method of using various techniques in detecting intrusions has been done before(Fossaceca et al., 2015). Mukkamala et al., (2005) combined artificial intelligence-based methods, such as ensemble of Neural Networks, support vector machine, and Multiple Adaptive Regressive Splines has has given good results but to apply it in large data set is still challenging. One can also use combination of multiple classification algorithms that work efficiently against each intrusion type in future

intrusion detection research

Detection Methodologies :

To overcome problems the limitations of various techniques as mentioned in Table 1 in intrusion detection, it is important to identify the exact requirement of intrusion detection in the current environment characterized by rapidly changing data, large volume and variety. I feel that to keep pace with the changing pattern of attack, solving one or two approaches may not be sufficient. Along with a good detection rate and accuracy, the following requirements (presented in the form of objectives – obj1, obj2 …) have to be addressed for a reliable intrusion detection.

Obj1: Efficient clustering and classification, that is, the machine should not be over-trained and should give unbiased result

The articles that I have read classify and cluster the attacks on a particular system. Henceforth, the techniques used by different authors is based on doing feature selection (which select the most relevant attributes contributing to the attack) and then training the system of the possible vulnerabilities. When a new set of attack is fed to the system, they can identify and classify new attack by comparing with the normal data. Any machine learning technique has a tendency of overfitting, which leads to erroneous results on a new data set. Due to overfitting, the technique fails to segregate a normal data from malicious attack and misleads the user by making proper prediction.

Obj2: Less human interaction

Human beings, in the form of system administrator or programmers, are usually involved in intrusion detection for configuring the environmental settings, writing requisite code and analysing the results. Due to the evolution of new ways with every passing days, new ways have to be invented in order to detect these intrusions and anomalies. Machine learning techniques make their task easier because of their self-learning capability, if these techniques are trained with certain set of patterns. The similar kind of patterns can be identified by them automatically without requiring human intervention.

Obj3: Low computational overhead and cost

The data required to be handled for intrusion detection is huge. Handling this huge amount of data is a herculean task for any researcher. Optimal selection of criteria which can identify the majority type of attacks is a challenging task for a researcher. In 2005 Tsang tried to select an optimal feature subset for dimensionality reduction. Aslahi-Shahri et al in 2016 and Chung and Wahid in 2012 did the in order to reduce the execution time to make space management effective. Most studies have strived to achieve this through the feature selection mechanism, otherwise results obtained by analysing un-filtered data by compromising space, time and system performance will have no significance later, especially in case of a real time attack.

Obj4: Identification of the new type of attacks

The mode of attack is constantly changing as new types of virus and worms are invented every day. The software developed today for handling the attack may not be equipped to handle these new types of attack thus resulting in loss of vital data and consequent huge economic loss. To combat this problem the researcher strives to design a system capable of identifying new types of attack by training the system with some predefined pattern. Hu et al. (2014) tried to identify new types of attack by reducing the communication cost. Stopel et al. (2009) proposed a system to detect new types of computer worms which will give good detection rate and accuracy and can be effectively used for identifying newly discovered attacks**.**

Obj5: Robustness- Capability of handling large interrelated datasets and real type packets

Dataset used for intrusion detection is huge with interrelated data, especially when we are dealing with real time data. Different researcher like Bankovic et al (2007), Denning and Neumann (1985), and Creech and (2014) has proposed a system for dealing with real time attacks. To classify a real time attack is a challenging task as they are interrelated with respect to time. For example, it is really difficult to accurately structure the behaviour of the system by extracting on an exact point, when two or more consecutive attacks are executed (Chen et al., 2017). Moreover, the user-content analysis demands accurate result within a

fraction of second. If the system is not robust to handle this situation, it can be catastrophic and can jeopardize the system.

Discussion :

The machine learning techniques and soft computing techniques are special techniques for handling huge data sets. Fuzzy logic techniques effectively cluster overlapping datasets. They are also used extensively in removing attributes which are a misfit for a particular cluster. The fuzzy rule-based system tries to match a pattern with a set of predefined patterns and thereby helps in removing misfit data of the cluster. Genetic Algorithm (GA) is also capable of handling unrelated data in clusters for using the process of selection, crossover and mutation. SOM (Self Organizing Map) reveals the most important relationships between the features and hides unwanted details. Because of abstraction, these techniques are expert in handling complex interrelated data and the suppression of unnecessary details helps in saving time and unwanted processing. SVM (Support Vector Machine) converts the highly complex data, especially the text, into a form suitable for classification. Neural network is effective by being trained on a certain pattern of data. It reports deviation from this pattern, if any, in a new dataset and thus helps in identification of dissimilarity formed by each cluster. Therefore, this technique in stand-alone mode or with other techniques can solve the security threats followed worldwide.

Problems in existing Technologies	Solved effectively with
Dissimilarity of the discovered pattern with existing pattern	Fuzzy clustering, BPNN
Dependencies among data and understanding dependency clustering	Fuzzy logic by the progressive reduction of cognitive dessonance, SOM, GA
Web personalization	GA, Fuzzy rule based system
Data summarization	Fuzzy set theory, SOM, BPNN,
Creation of Association rules	Fuzzy acyclic directed graph, Fuzzy rule based system
Difficulty in processing documents containing images	Fuzzy logic, SOM
Regression	Neural network, neuro fuzzy computation
Extra attribute handling	K-nearest neighbour, GA
Complexity of relationship owing to a large number of variables	SVM, Decision Tree
Time Consuming	SOM
Difficulty in managing huge dataset	Genetic algorithm, SOM, SVM
Reliability on the user to change cluster	Genetic algorithm

Table 3: Machine learning techniques and their solution as identified in Table 1

One can observe that machine learning techniques with their self-learning or supervisory mode are able to detect most of the attacks and have provided very good detection and accuracy rate. However, these metrics alone do not consider the hostility of the environmental condition. For this reason, certain other metrics such as cost and sensitivity need to be taken into consideration. If the environmental condition is not taken into consideration, detection of new types of attack will be a difficult task. Moreover, reliability on KDD dataset is also not a good solution to judge the efficacy of the system regarding detection of attacks. The KDD dataset contains redundant records and has laid less stress on U2R and R2L attacks. In most of the training experiments, these attacks are likely to be missed, as the number of rows is less as compared to other types of attacks. Moreover, 75% to 78% of the records are duplicate. Based on the training of KDD dataset, the new dynamic real time attack may not be handled by the system. There is still requirement for a unified architecture or technique, which will provide a platform to identify real time attack, and also a standardized solution in handling wired and wireless attack, as shown in Table 5. It is evident from

Table 5 that different environmental conditions favour different techniques.

Area of observation	Detail observation	Concluding remark
Percentage of total analyzed literature	SOM and SVM covers 13% and 18% respectively. Fuzzy and GA covers 14% and 19% respectively while Perceptron covers 17%	According to our survey GA,Perceptron and SVM are most popular tool.
Most Common Approach	Fuzzy rule based system with GA covers nearly 11% of analyzed literature	Genetic Algorithm used on knowledge base dataset containing fuzzy rule are popular techniques used for feature selection.
Most common performance metrics	Accuracy and detection rate. Detection rate covers 49% of analyzed data, Accuracy covers 28% of analyzed data	Detection is given more importance than analysis during performance evaluation

Table 5: Observation based on machine learning techniques approach in intrusion detection

Current trends in network involves distributed computing with increasing demand for cloud computing (i.e. more involvement of internet) (Kumar et al., 2010). In addition, ad-hoc and sensor networks have indicated possibility of new type of attacks. The performance of most of the above mentioned techniques in dealing with intrusion in cloud platform or determining black hole attack in sensor networks is questionable. Intrusion detection fails to determine the attacks at different levels of architecture of a cloud. Meanwhile as stated by Modi et al (2013), internal attacks are also increasing. There is a lack of suitable mechanism to handle them. Sensor networks on the other hand are more sensitive to attack. Mobile nodes with poor inbuilt security mechanism are easy to capture via wired networks. An attacker can listen to traffic, modify the traffic or can act as one of the legitimate users. As there is no such central architecture which can help in intrusion detection, proper cryptography via public or private key is difficult to implement in mobile adhoc or sensor network. Our article has identified the impact of machine learning on intruded packets and at the same time has identified the issue of security concern that are left to be handled when dealing with real time data, sensitive data in mobile phones and sensor networks.

Conclusion :

This study provides an insight into the progress of research on intrusion detection based on machine learning techniques. It has discussed the most popular machine learning techniques, and their advantages and disadvantages. As machine learning techniques are extensively used with soft computing techniques that have been analysed . Most of the techniques perform well with KDD. Fuzzy logic techniques perform well with real time datasets. The study also revealed that machine learnings approaches applied to intrusion detection are quite successful except in the matter of fulfilling the objectives of low computational and cost overheads and robustness (capability of handling large interrelated datasets and real type packets). Therefore, the future research may be directed towards those machine learning tools that will achieve both of these objectives (i.e. low computational cost and robustness). More obvious gap is the labelled data application like KDD dataset on which majority of the techniques are applied and it would be more worthy if intrusion data is collected and labelled partially. As the machine learning techniques require training and testing data so they can be trained using partial labelled dataset for the known attack and tested on unknown data for measuring performance. Therefore, the promising techniques may be further tested on these new data set for developing effective and efficient intrusion detection system for breakthrough performance. Moreover since majority of the results are based on KDD dataset, approximation of the actual performance of the intrusion detection system on real time data is difficult to evaluate. The effectiveness of these techniques on real time data and effective performance metrics used for their evaluation is an open area for the future researchers.

References :

a) Aburomman, A. A., & Ibne Reaz, M. Bin. (2016). A novel SVM-kNN-PSO ensemble method for intrusion detection system. Applied Soft Computing Journal, 38, 360–372. https://doi.org/10.1016/j.asoc.2015.10.011

b) Anderson, D., Frivold, T., & Valdes, A. (1995). Next-generation Intrusion Detection Expert System (NIDES): A summary. SRI International, (May 1995), 47. https://doi.org/citeulikearticle-id:7898221

c) Arun Raj Kumar, P., & Selvakumar, S. (2013). Detection of distributed denial of service attacks using an ensemble of adaptive and hybrid neuro-

fuzzy systems. Computer Communications, 36(3), 303–319. https://doi.org/10.1016/j.comcom.2012.09.010

d)Ashfaq, R. A. R., Wang, X. Z., Huang, J. Z., Abbas, H., & He, Y. L. (2017). Fuzziness based semisupervised learning approach for intrusion detection system. Information Sciences, 378, 484–497. https://doi.org/10.1016/j.ins.2016.04.019

e)Aslahi-Shahri, B. M., Rahmani, R., Chizari, M., Maralani, A., Eslami, M., Golkar, M. J., & Ebrahimi, A. (2016). A hybrid method consisting of GA and SVM for intrusion detection system. Neural Computing and Applications, 27(6), 1669–1676. https://doi.org/10.1007/s00521-015-1964-2

f)Banković, Z., Stepanović, D., Bojanić, S., & Nieto-Taladriz, O. (2007). Improving network security using genetic algorithm approach. Computers and Electrical Engineering, 33(5–6), 438–451. https://doi.org/10.1016/j.compeleceng.2007.05.010

g)Chen, C., Ghassami, A., Mohan, S., N. K. (2017), A Reconnaissance Attack Mechanism for FixedPriority Real-Time Systems. Arxiv.Org. Chung, Y. Y., & Wahid, N. (2012). A hybrid network intrusion detection system using simplified swarm optimization (SSO). Applied Soft Computing, 12(9), 3014–3022. https://doi.org/10.1016/j.asoc.2012.04.020

h)Creech, G., & Hu, J. (2014). A semantic approach to host-based intrusion detection systems using contiguous and discontiguous system call patterns. IEEE Transactions on Computers, 63(4), 807–819. https://doi.org/10.1109/TC.2013.13

i)Damopoulos, D., Menesidou, S. A., Kambourakis, G., Papadaki, M., Clarke, N., & Gritzalis, S. (2012). Evaluation of anomaly-based IDS for mobile devices using machine learning classifiers. Security and Communication Networks, 5(1), 3–14. https://doi.org/10.1002/sec.341