Machine Learning for Network Intrusion Detection: A Survey

Jyoti Gupta[1] Anshul Khurana[2]
[1,2]Shri Ram Institute of Technology, Madhya Pradesh, India

Abstract— "Network intrusion detection system (NIDS)" monitors traffic on a network looking for doubtful activity, which could be an attack or illegal activity. The intrusion detection techniques based upon data mining are generally plummet into one of two categories: misuse detection and anomaly detection. In misuse detection, each instance in a data set is labeled as 'normal' or 'intrusive' and a learning algorithm is trained over the labeled data. In this paper we will discuss about the steps involved in NIDS, further we will compare different techniques of NIDS based on accuracy parameter i.e. precision and recall.

Key words: IDS

I. INTRODUCTION

Malevolent users and crackers pursue weedy targets such as unpatched systems, systems infested with Trojans, and networks running anxious services. The guarantee of reliability and safety should be functional to computer systems and data. Internet has made possible the data flow to the huge extent. Also all together it has to face many threats and assaults. Thus the safety alert is essential to control the attacks and threats. A report must be sent to the administrators and security team members about the numerous threats and attacks which has transpired so that they can react in real-time to the intimidation.

Intrusion detection is the process of classifying and responding to malevolent activities targeted at computing and network resources". An intrusion endeavor, also known as attack, mentions to a sequence of actions by use of which an intruder endeavors to gain control over a system.

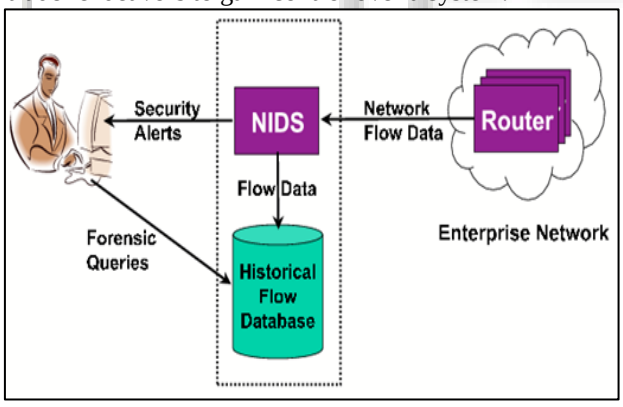

Fig. 1: Working of NIDS

The intrusion detection techniques based upon data mining are generally fall into one of two categories: misuse detection and anomaly detection.

An intrusion detection system intentions to differentiate between intrusion activity and normal actions. In doing so, conversely, an IDS can familiarize classification errors. A false positive is a gentle input for which the system speciously raises a notification. A false negative, in contrast, is a malevolent input that the IDS miscarries to report. The appropriately classified input data are typically mentioned to as true positives (suspicious attacks) and true negatives (normal traffic). There is a natural trade-off between distinguishing all malevolent events (at the outlay of floating alarms too over and over again, i.e., having high false positives), and missing anomalies (i.e., having high false negatives, but not give out many false alarms). We graphically show this trade-off in Figure 2. It is frequently the case that we can control the system performance by modifying specific IDS parameters, as snippet recommended in Figure 2 by the dashed line: the area to the left of the line results in low false positive rate, but an increasing false negative rate; similarly, the region to the right favors a low false negative rate, but it has a higher false positive rate. Which factor of the trade-off is more significant is a case-specific decision, and preferably, we would want to improve both factors. We might want to classify all malevolent attempts, because this would make our network harmless. However, this would be of no use if the number of alerts would overload the IT specialist accountable for handling them.

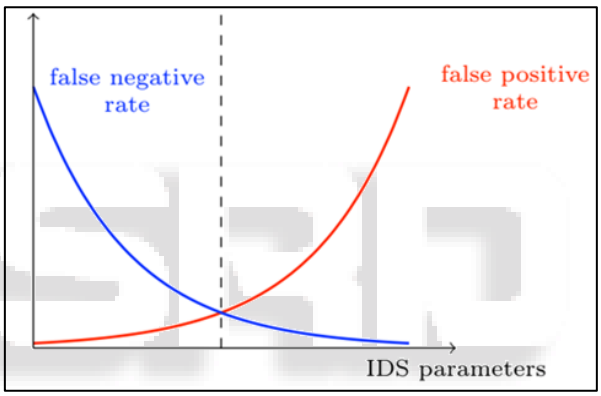

Fig. 2: Trade-off between false positive and false negative rates

Further in this paper in section II we will explain the different literatures and conclusion of literature, in section III we will provide the tabular comparison among different literature, in section IV we will give brief introduction about Weka api, at last we will conclude our survey.

II. LITERATURE SURVEY

Abhinav Kumra et. al. said that proposed method was triumphantly tested on the data log fles and the database. The results of the proposed testimony are produce more accurate and irrelevant sets of patterns and the discovery time is less than other approach. As a naïve Bayesian network is a restricted network that has only two layers and assumes complete independence between the information nodes. This poses a limitation to this research work. In order to alleviate this problem so as to reduce the false positives, active platform or event based classification may be thought of using Bayesian network [OJCST 2017].

Anna L. Buczak et. al. 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 used 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 some recommendations on when to use a given method are provided [IEEE 2016].

Kathleen Goeschel has shown that high accuracy may be maintained while reducing false positives using the proposed model composed of SVMs, decision trees, and Naïve Bayes. First, the SVM is trained based upon a new binary classification added to the dataset to specify if the instance is an attack or normal traffic. Second, attack traffic is routed through a decision tree for classification. Third, Naïve Bayes and the decision tree will then vote on any unclassified attacks. Future work is to write this model as a Java class such that it may be applied in other systems or applications. Further future work is to test this model on other network traffic data sets for more in-depth analysis [IEEE 2016].

Bane Raman Raghunath et. al. focuses on two specific contributions: (i) an unsupervised anomaly detection technique that assigns a score to each network connection that reflects how anomalous the connection is, and (ii) an association pattern analysis based module that summarizes those network connections that are ranked highly anomalous by the anomaly detection module.

Deepika P Vinchurkar et. al. said that in the recent years, Intrusion Detection materializes the high network security. Thus tries to be the most perfect system to deal with the network security and the intrusions attacks. Monitoring activity of the network and that of threats is the feature of the ideal Intrusion Detection System. Intrusion Detection System is classified on the basis of the source of Data and Model of Intrusion. There are some challenges faced by the Intrusion Detection System. Neural Network amd Machine Learning are the approaches through which the challenges can be overwhelmed. Anomaly in the Anomaly based Intrusion Detection System can be detected using various Anomaly detection techniques. Dimension Reduction can be done using Principle Component Analysis. Support Vector Machine can be used to specify the classifier construction problem. Author describes the various approaches of Intrusion detection system in briefly [IJESIT 2012].

DikshantGupta et. al. said that there are many risk of network attacks in the Internet environment. Nowdays, Security on the internet is a vital issue and therefore, the intrusion detection is one of the major research problem for business and personal networks which resist external attacks. A Network Intrusion Detection System (NIDS) is a software application that monitors the network or system activities for malicious activities and unauthorized access to devices. The goal of designing NIDS is to protect the data's confidentiality and integrity. Author focuses on these issues with the help of Data Mining. This research paper includes the implementation of different data mining algorithms including Linear regression and K-Means Clustering to automatically generate the rules for classify network activities. A comparative analysis of these techniques to detect intrusions has also been made. To learn the patterns of the attacks, NSL-KDD dataset has been used [IEEE 2016].

Jayveer Singh et. al. said that the rapid development of computer networks in the past decades has created many security problems related to intrusions on computer and network systems. Intrusion Detection Systems IDSs incorporate methods that help to detect and identify intrusive and non-intrusive network packets. Most of the existing intrusion detection systems rely heavily on human analysts to analyze system logs or network traffic to differentiate between intrusive and non-intrusive network traffic. With the increase in data of network traffic, involvement of human in the detection system is a non-trivial problem. IDS's ability to perform based on human expertise brings limitations to the system's capability to perform autonomously over exponentially increasing data in the network. However, human expertise and their ability to analyze the system can be efficiently modeled using soft-computing techniques. Intrusion detection techniques based on machine learning and softcomputing techniques enable autonomous packet detections. They have the potential to analyze the data packets, autonomously. These techniques are heavily based on statistical analysis of data. The ability of the algorithms that handle these data-sets can use patterns found in previous data to make decisions for the new evolving data-patterns in the network traffic. In this paper, we present a rigorous survey study that envisages various soft-computing and machine learning techniques used to build autonomous IDSs. A robust IDSs system lays a foundation to build an efficient Intrusion Detection and Prevention System IDPS [IJARCET 2013].

Salima Omar et. al. said that Intrusion detection has gain a broad attention and become a fertile field for several researches, and still being the subject of widespread interest by researchers. The intrusion detection community still confronts difficult problems even after many years of research. Reducing the large number of false alerts during the process of detecting unknown attack patterns remains unresolved problem. However, several research results recently have shown that there are potential solutions to this problem. Anomaly detection is a key issue of intrusion detection in which perturbations of normal behavior indicates a presence of intended or unintended induced attacks, faults, defects and others. This paper presents an overview of research directions for applying supervised and unsupervised methods for managing the problem of anomaly detection. The references cited will cover the major theoretical issues, guiding the researcher in interesting research directions.

III. MOTIVATION

Network based Intrusion Detection System (NIDS) monitors the traffic as it flows to other host. Monitoring criteria for a specific host in the network can be increased or decreased with relative ease. NIDS should be capable of standing against large amount of network traffic to remain effective. As network traffic increases exponentially NIDS must clutch all the traffic and analyze in a timely manner.

There are some challenges in making accurate and efficient NIDS, which motivated towards this research, are as follows:

- The size of input dataset is very large.
- Value of FNR (False Negative rate) i.e. accuracy is less.

- Misuse detection (signature detection) method is that it cannot detect novel attacks and variation of known attacks.
- Another challenge for the IDS is to generalize from the previously observed behavior to recognize similar future behavior.

IV. COMPARISON

S. No.	Author/Title/Year Publication	Description	Algorithm Used and Performance
1.	DikshantGupta et. al./Network Intrusion Detection System Using various data mining techniques/IEEE 2016	Paper includes the implementation of different data mining algorithms including Linear regression and K-Means Clustering to automatically generate the rules for classify network activities.	Linear regression and K-Means Clustering Linear regression-80% Accuracy K-Means Clustering-67% Accuracy
2.	Upendra et. al./An Empirical Comparison and Feature Reduction Performance Analysis of Intrusion Detection/IJCTCM 2012	Compared the performance measure of five machine learning classifiers such as Decision tree J48,BayesNet,OneR,Naive Bayes and ZeroR.The results are compared and found That J48 is excellent in performance than other classifiers with respect to accuracy.	J48, BayesNet, ZeroR Approximate all algorithm gave 80% accuracy
3.	A J M Abu Afza et. al./Intrusion Detection Learning Algorithm through Network Mining/ IEEE 2013	Author present a Dependable Network Intrusion Detection System (DNIDS) by integrating detection method with an intrusion-tolerant mechanism. To address the detection issue, we propose a Combined Strangeness and Isolation measure K-Nearest Neighbor (CSI-KNN) algorithm. The algorithm employs a combined model that uses two different measures to improve its detection ability.	CSI-KNN Algorithm 76% Accuracy
4.	Neethu B/Adaptive Intrusion Detection Using Machine Learning/IJCSNS 2013	Paper applies PCA for feature selection with Naïve Bayes for classification in order to build a network intrusion detection system. For experimental analysis, KDDCup 1999 intrusion detection benchmark dataset have been used. The 2 class classification is performed. The experimental results show that the proposed approach is very accurate with low false positive rate and takes less time in comparison to other existing approaches while building an efficient network intrusion detection system.	PCA 85% Accuracy but not time efficent
5.	Upendra/An Efficient Feature Reduction Comparison of Machine Learning Algorithms for Intrusion Detection System/ijettcs 2013	Intrusion detection present an important line of defend against all variety of attacks that can compromise the security and proper functioning of information system initiative. In this paper author compared the performance of intrusion detection. The evaluation of the Intrusion Detection System (IDS) execution analysis for any given security system configuration improvement is necessary to achieve real time capability. Author analyse two learning algorithms (NB and C4.5) for the task of detecting intrusions and compare their relative performances.	NB , C4.5 Accuracy 76%
6.	Abhinav kumra et. al./Intrusion Detection System Based on Data Mining Techniques/OJCST 2017	Author has proposed method was triumphantly tested on the data log fles and the database. The results of the proposed testimony are produce more accurate and irrelevant sets of patterns and the discovery time is less than other approach. As a naïve Bayesian network is a restricted network that has only two layers and assumes complete independence between the information nodes.	Naïve Bayesian network
7.	Bane Raman Raghunath et. al./Network Intrusion Detection System (NIDS)/IEEE 2008	This paper introduces the Network Intrusion Detection System (NIDS), which uses a suite of data mining techniques to automatically detect attacks against computer networks and systems. This paper focuses on two specific contributions: (i) an unsupervised anomaly detection technique that assigns a	Association rule mining and k-nearest neighborhood

		score to each network connection that reflects how anomalous the connection is, and (ii) an association pattern analysis based module that summarizes those network connections that are ranked highly anomalous by the anomaly detection module.	

Table 1: Comparison

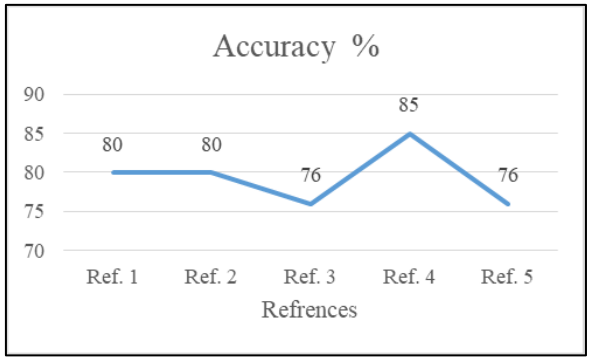

Fig. 3: Accuracy percentage as per Table-1

V. WEKA

Weka is a gathering of machine learning algorithms for data mining tasks. The algorithms can either be useful unswervingly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Weka is open source software issued under the GNU General Public License.

For NIDS implementation KDD dataset is used. This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad" connections, called intrusions or attacks, and ``good" normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

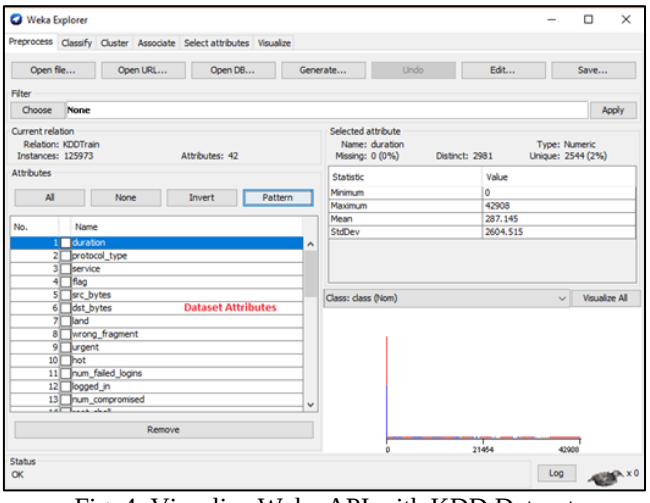

Fig. 4: Visualize Weka API with KDD Dataset

VI. CONCLUSION

Achieving network security through examining the behavior of network and network flow, became a vital research area. There are several research has been carried out is this field several data mining, machine learning algorithm has been applied over input dataset, from fig.-3 we can conclude that still need of improvement in accuracy of existing algorithms and for comparison table we can say algorithms which are more accuracy those are not time efficient.

REFERENCES

[1] DikshantGupta et. al./Network Intrusion Detection System Using various data mining techniques/IEEE 2016.
[2] Upendra et. al./An Empirical Comparison and Feature Reduction Performance Analysis of Intrusion Detection/IJCTCM 2012
[3] A J M Abu Afza et. al./Intrusion Detection Learning Algorithm through Network Mining/ IEEE 2013
[4] Neethu B/Adaptive Intrusion Detection Using Machine Learning/IJCSNS 2013
[5] Upendra/An Efficient Feature Reduction Comparison of Machine Learning Algorithms for Intrusion Detection System/ijettcs 2013
[6] Abhinav kumra et. al./Intrusion Detection System Based on Data Mining Techniques/OJCST 2017
[7] Bane Raman Raghunath et. al./Network Intrusion Detection System (NIDS)/IEEE 2008.
[8] Annie George, ‗Anomaly Detection based on Machine Learning: Dimensionality Reduction using PCA and Classification using SVM', International Journal of Computer Applications (0975 – 8887) Volume 47– No.21, June 2012.
[9] W.K. Lee, S.J.Stolfo. ―A data mining framework for building intrusion detection model‖, In: Gong L., Reiter M.K. (eds.): Proceedings of the IEEE Symposium on Security and Privacy. Oakland, CA: IEEE Computer Society Press, pp.120~132, 1999.
[10] V. Jyothsna, V. V. Rama Prasad, K. Munivara Prasad, ‗A Review of Anomaly based Intrusion Detection Systems' International Journal of Computer Applications (0975 – 8887) Volume 28– No.7, August 2011.
[11] Neethu B, ‗Classification of Intrusion Detection Dataset using machine learning Approaches' International Journal of Electronics and Computer Science Engineering 1044 ISSN- 2277-1956. Available Online at www.ijecse.org.
[12] Lindsay I Smith, ―A tutorial on Principal Components Analysis‖.
[13] CHEN Bo, Ma Wu, ―Research of Intrusion Detection based on Principal Components Analysis‖, Information Engineering Institute, Dalian University, China, Second

International Conference on Information and Computing Science, 2009.
[14] T. J.Hastie, R. J.Tibshirani, and J. H.Friedman. The elements of statistical learning: Data mining, inference, and prediction, Springer-Verlag, 2001.

