An Integrated Approach of Junction Tree & Naive Bayes for Network Intrusion Detection

Jyoti Gupta
Rungta College of Engineering and Technology, Bhilai, Chhattisgarh, 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. In this paper we have proposed an integrated approach of Junction Tree and Naïve Bayes machine learning algorithm for detection of network intrusion, dataset used for experimental evaluation is KDD dataset.

Key words: IDS

I. INTRODUCTION

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 1. It is frequently the case that we can control the system performance by modifying specific IDS parameters, as snippet recommended in Figure 1 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 favours 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.

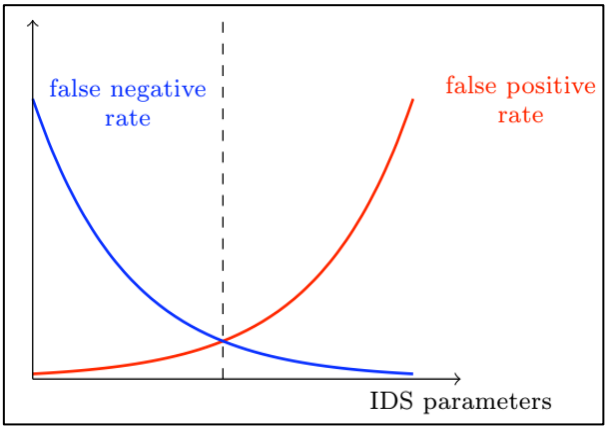

Fig. 1: Trade-Off between False Positive and False Negative Rates

False Positive Rate (FPR)	Normal traffic identified as an attack. FPR = FP/(FP+TN)
False Negative Rate (FNR)	Attack traffic identified as normal. FNR = FN/(TP+FN)
True Positive Rate (TPR)	An attack correctly identified. TPR = TP/(TP+FN)
True Negative Rate (TNR)	Normal traffic correctly identified. TNR = TN/(TN+FP)
Accuracy	TP+TN)/(TP+TN+FN+FP)
Precision	TP/(TP+FP)

Table 1: Accuracy Calculation

Further in this paper in next section we will brief how we motivated towards this research, in next section we provide different literature survey, followed by tabular comparison among literatures, in next section we will provide details of our proposed methodology at last we will conclude our research.

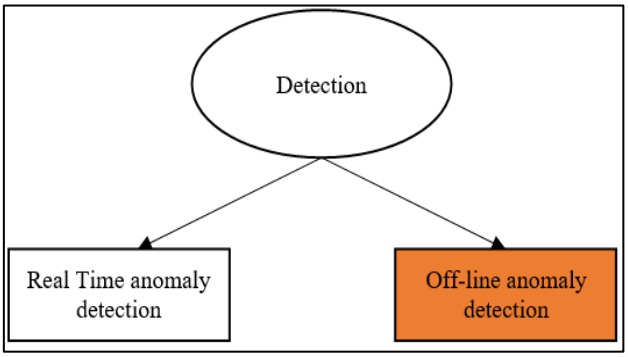

Fig. 2: Detection Classification

II. MOTIVATION

Network security is of vital significance in the present data communication. Programmers and interlopers can make numerous effective endeavours to cause the crash of the systems and web benefits by unapproved interruption. New dangers and related answers for avert these dangers are developing together with the secured framework advancement. Intrusion Detection Systems (IDS) are one of

these arrangements. The principle capacity of Intrusion Detection System is to ensure the assets from dangers. It dissects and predicts the practices of clients, and after that these practices will be considered an assault or an ordinary conduct. Intrusion identification enables association to shield their frameworks from the dangers that accompany expanding system network and dependence on data frameworks. Ought, to, not be regardless of whether to utilize intrusion location yet rather which intrusion discovery highlights and abilities can be utilized.

Machine learning touches our everyday lives from multiple points of view. When you transfer a photo via web-based networking media, for instance, you may be incited to label other individuals in the photograph. That is called picture acknowledgment, a machine learning ability by which the PC figures out how to distinguish facial highlights. Different cases incorporate number and voice recognition applications. From an intrusion discovery point of view, examiners can apply machine learning, data mining and pattern recognition algorithms to recognize typical and noxious network activity. There is lot more has been carried out in this field still with less accuracy of detection henceforth we require an intelligent network intrusion detection system with higher accuracy.

III. LITERATURE SURVEY

Abhinav Kumraet. al. said that proposed method was triumphantly tested on the data log files 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. Buczaket. 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 Raghunathet. 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 Vinchurkaret. 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].

DikshantGuptaet. 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].

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 efficient
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.	Abhinavkumra 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

Table 1:

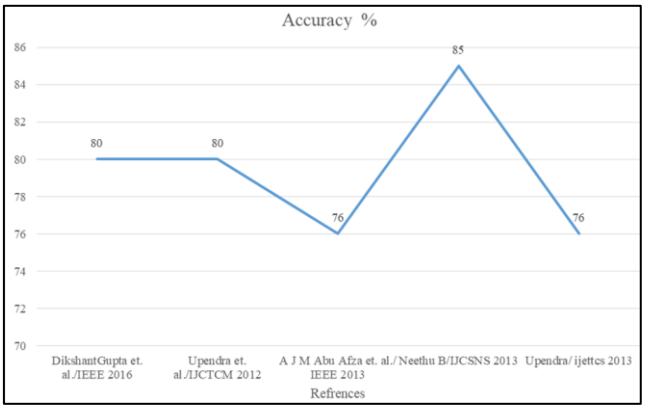

Fig. 3: Accuracy Comparison Graphical Representation

IV. PROBLEM IDENTIFICATION

"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 labelled as 'normal' or 'intrusive' and a learning algorithm is trained over the labelled data.

Figure- 2 shows the different approach for network intrusion detection (NIDS), in this research Off-line anomaly detection using machine learning will be taken into account because from literature review section it can be concluded that still there is need if intelligent NIDS system will detect intruder efficiently with high precision value. There is some bottleneck identifier in earlier algorithm which are as follows:

- Due to very large dataset algorithm needed which should be time efficient.
- Earlier system having FNR (False Negative Rate) is high.
- Accuracy of anomaly detection is less.
- Can classify anomaly.

V. PROPOSED METHODOLOGY

After gone through numerous literature, come across conclusion that there is need of an efficient Network intrusion detection algorithm which should have higher value of precision i.e. algorithm should have high accuracy.

Earlier algorithm uses different data mining or machine learning algorithms such as naïve base classifier, J48 classifier, decision tree classifier etc., in our proposed algorithm we will integrate Junction Tree (J48) and Naïve Bayes algorithm for detection of anomaly.

In our methodology also, we divide the KDD training dataset in different sizes such as 17.8mb, 9mb and 3mb. Then after that we apply integrate Junction tree (J48) and Naive Bayes algorithm on each different data sets. So as to calculate and compare the accuracy of dataset in different sizes.

A. Proposed Algorithm

1) Step-1 Read Training Data.
2) Step-2 Read Test Data.
   //Remove Missing and Repeated Attribute
3) Step-3 Prune Train and Test Data.
4) Step-4 Count number of instances of test dataset.
5) Step-5 Set i-> 0
6) Step-6 Loop i< number of instances
   - Classify each instance of test dataset through naive base classifier
   - Check for anomaly using J48
   - Compare with actual value to calculate accuracy
7) Step-7 End Loop

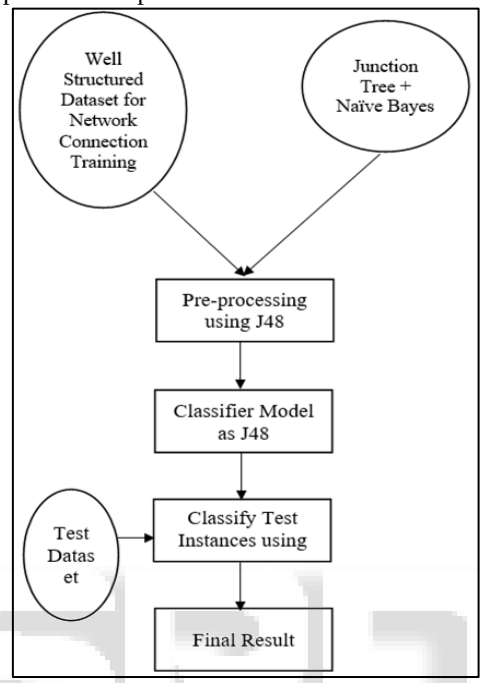

Fig. 4: Proposed Approach

VI. RESULT AND DISCUSSION

For implementation of our proposed algorithm we have opted JDK 1.8 further in this section, we will evaluate the effectiveness of our proposed algorithm. Here we have presented an experimental valuation using data. As a source dataset for experimental evaluation we have used KDD dataset. Furthermore we have compared the performance of Junction Tree and Naïve Bayes algorithm with our proposed algorithm.

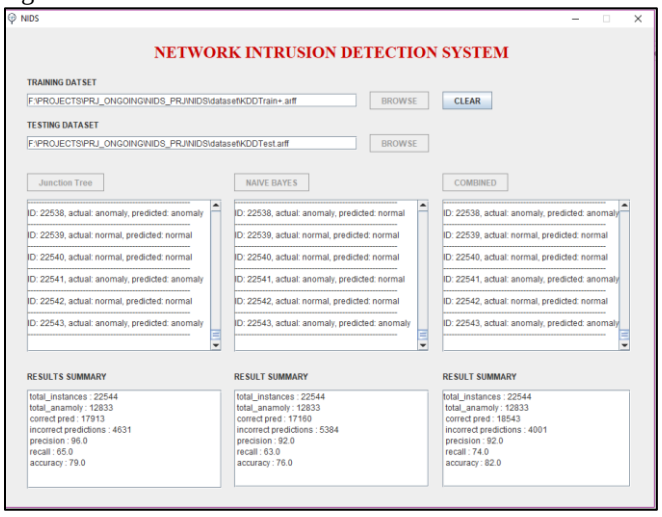

Fig. 5: Main GUI

Algorithm	Accuracy
J48	79
Naïve Bayes	76
Proposed	82

Table 2:

DATASETS IN DIFFERENT SIZES	J48	NAIVE BAYES	PROPOSED METHODOLOGY
17.8MB	79	76	82
9MB	76	70	81
3MB	75	70	80

Table 3:

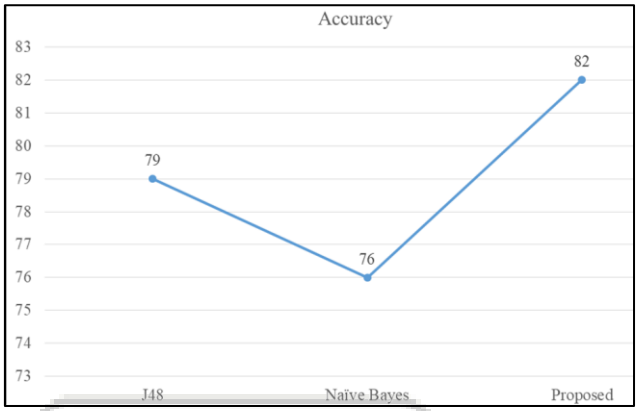

Fig. 6: Graphical Accuracy Comparison

Accuracy Comparison table in different sizes of KDD Dataset

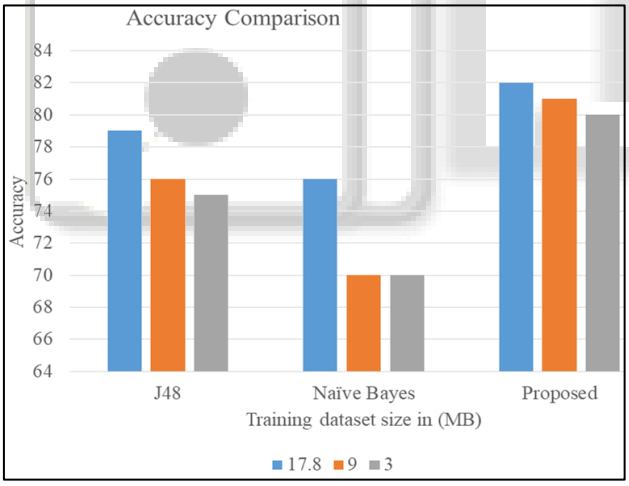

Fig. 7:

VII. Conclusion

An intrusion detection framework expectations to separate between interruption action and ordinary activities. In doing as such, then again, an IDS can acquaint order blunders. A false positive is a delicate contribution for which the framework probably raises a notice. A false negative, interestingly, is a malicious information that the IDS prematurely delivers to report. The suitably grouped info information are ordinarily specified to as evident positives (suspicious assaults) and genuine negatives (ordinary activity).

After going through result and discussion section and comparative graph among NB (Naïve Bayes Classifier), J48 (Junction Tree Classifier) and proposed classifier, accuracy of anomaly detection is more in our proposed algorithm than that of given problem. We also come in conclusion that when we divide the datasets in different sizes then result of accuracy of anomaly detection is more.

References

[1] DikshantGuptaet. al./Network Intrusion Detection System Using various data mining techniques/IEEE 2016.
[2] Upendraet. al./An Empirical Comparison and Feature Reduction Performance Analysis of Intrusion Detection/IJCTCM 2012
[3] A J M Abu Afzaet. 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] Abhinavkumraet. al./Intrusion Detection System Based on Data Mining Techniques/OJCST 2017
[7] Bane Raman Raghunathet. 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.

