International Journal of Research in Engineering, Science and Management
Volume-3, Issue-2, February-2020
www.ijresm.com | ISSN (Online): 2581-5792

Detecting Network Intrusion Based on Machine Learning Techniques

J. Priyanka[1], S. Sajithabanu[2], N. Balasubramanian[3]
[1]*Student, Department of Master of Computer Application, Mohamed Sathak Engineering College, Ramanathapuram, India*
[2,3]*Assistant Professor, Department of Master of Computer Application, Mohamed Sathak Engineering College, Ramanathapuram, India*

***Abstract*: Intrusion Detection (IDS) is one of the obtainable mechanisms that is used to sense and classify any abnormal actions. The internet continues to spread itself over the globe, providing the great opportunity for various threats which are growing on a daily basis. To reduce the work, systems capable of analyzing contents of the network by means of machine learning techniques to analyze and classify the malicious contents. KDDCUP 99 Dataset is used for detecting the intrusion. The KDDCUP 99 data to survey and evaluate a research in IDS by identifying attack or normal and feature selection method to reduce the complexity of these datasets. Currently we proposed machine learning based intrusion detection system in order to assist Network Intrusion Detecting system developers to gain a better result.**

***Keywords*: Intrusion detection, Machine learning.**

1. Introduction

Data mining is the computing process of finding patterns in massive datasets involving methods at the interchange of machine learning, statistics and database system. The overall objective of the data mining process is to extract information from a data set and transform it into an understandable structure for further use transform it into an understandable structure for further use. Data mining is the scrutiny step of the "knowledge discovery in datasets" process, or KDD. Data mining is about discovery new information in a lot of data. The information obtained from data mining hopefully both new and useful.

This problem is customary named with the term "big data," which refers to the difficulties and drawback of processing and analyzing enormous amounts of data. It has fascinated much attention in a great number of areas such as bioinformatics, medicine, marketing, or financial businesses, because of the huge collections of raw data that are stored. Recent advances on Cloud Computing technologies allow for embracing standard data mining techniques in order to apply them successfully over massive amounts of data. The adaptation of data mining tools for big data problems may require the rehabilitate of the algorithms and their inclusion in parallel environments.

Among the different algorithms and also to better their perfection by eliminating noisy and redundant data. The esoteric literature describes two main types of data reduction models. On the one hand, instance selection and instance generation processes are concentrating on the instance level. On the other hand, feature selection and feature extraction models work at the level of characteristics. Among the existing techniques, evolutionary approaches have been successfully used for feature selection techniques. Nevertheless, an immoderate increment of the individual size can control their applicability, being unable to provide a preprocessed dataset in a reasonable time when dealing with huge problems. In the current literature, there are no approaches to tackle the feature space with evolutionary big data models. To do this, a Map Reduce algorithm has been developed, which bisection the data and performs a bunch of EFS processes in parallel in the map phase and then combines the solutions in the reduce phase.

Learning from very large databases is a major issue for most of the current data mining and machine learning algorithms. This problem is commonly named with the term "big data," which refers to the difficulties and disadvantages of processing and analyzing huge amounts of data. It has attracted much attention in a great number of areas. Recent advances on Cloud Computing technologies allow for adapting standard data mining techniques in order to apply them successfully over massive amounts of data.

2. Intrusion detection

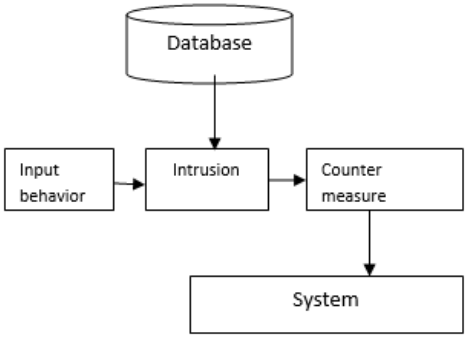

Fig 1. Block Diagram of IDS

An IDS is a hardware or software application that observe

network traffic data on a system or a network. An IDS typically reports any policy violations or security breaches. A block diagram of typical IDS is an shown in figure 1. An Intrusion detection system(IDS) is a device or software application that observe a network or systems for a malicious activity or policy violation. Any intrusion pursuit or policy violation is typically reported either to an administrator or collected using a security details and event management (SIEM) system. A SIEM system combines outputs from multiple sources and uses alarm filtering techniques to distinguish malicious activity from false alarms.

3. Intrusion detection category

IDS can be classified by where detection takes place (network or host) or the detection method that is employed (signature or anomaly based).

A. Network intrusion detection

Network Intrusion Detection systems are set up at an organized point within the network to inspect traffic from all devices on the network. It performs a scrutiny of passing traffic on the whole subnet and contest the traffic and matches the traffic that is passed on the subnets to the collection of known attacks. Once an attack is identified or abnormal behavior is observed, the alert can be sent to the administrator.

4. Machine learning

Machine learning is an application of artificial intelligent (AI) that provides systems the capability to automatically learn and improve from experience without being explicitly programmed. Machine learning pivot on the development of computer programs that can access data and use it learn for themselves Machine learning algorithms are often categorized as supervised or unsupervised.

A. Supervised machine learning algorithms

Algorithm can apply what has been learned in the past to new data using labeled examples to predict future events. starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output and find errors in order to modify the model accordingly.

B. Unsupervised machine learning algorithms

Algorithms are used to train is neither classified nor label, Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data the system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

C. Random forest classification algorithm

It is of the classification tree algorithm, the main goal of this algorithm is to enhance trees classifiers based on the concept of the forest. Random forest classification produced by the referred research, had an accepted accuracy rate and can be implemented to handle noise values of dataset. There is no re modification process during the classification step. To implement this algorithm the number of trees within a forest predicts the expects output and after that have the largest votes number

5. KDD dataset preprocessing and analysis

KDD dataset gave a good understanding of several intrusion behaviors, in the same time it is widely used in several areas for testing and evaluation intrusion detection algorithms. The first published of KDD dataset was 1999 by MIT licoln labs at University of California. It includes instances with 41 attributes. In this work KDD dataset was imported to the SQL server to implement various statistical measurements valuesIt can be concluded that there are 21 type of attacks categorized into four groups with different number of instances and occurrences in the KDD dataset.

The DOS attacks present 79% of KDD dataset while normal packets present 19% and other attack type recorded 2% of existing. Based on these values the KDD dataset appears as an unbalanced dataset but at the same time it included the largest number (41) of packet attributes. The fundamental attribute information for any connection implemented based on TCP/IP connection environment. The main contribution of this dataset is the introduction of 32 expert suggested attributes which help to understand the behavior of different types of attack.

Table 1
Distribution of attacks within KDD dataset

Categories of attack	Attack name	Number of instances
Dos	SMURF	28076
	NEPTUNE	10717
	Back	2203
	POD	264
	Teardrop	979
U2R	Buffer overflow	30
	Load module	9
	PEARL	3
	Root kit	10
R2L	FTP write	8
	Guess passwd	53
	IMAP	12
	MultiHop	7
	PHF	4
	SPY	2
	Warez Mster	20
	Warez client	1020
Normal		2850

6. Relevant works to the KDD dataset

This section presents the related works relevant to using KDD dataset for implementing machine learning algorithms. It also provides a brief overview of the different machine learning algorithms.

From the another perspective, some of the researches focus

on attribute selection algorithm in order to reduce the cost of computation time. We focused on selecting the most significant attributes to design IDS that have a high accuracy rate with low computation time. 10% of KDD was used for training and testing.

In addition, the genetic algorithm was implemented to enhance detection of different types of intrusion within the KDD is proposed. The proposed methodology aims to derive the maximum detection rate for intrusion types, at the same time achieved minimum false positive rate for intrusion types, at the same time achieved the minimum false positive rate.

7. System architecture

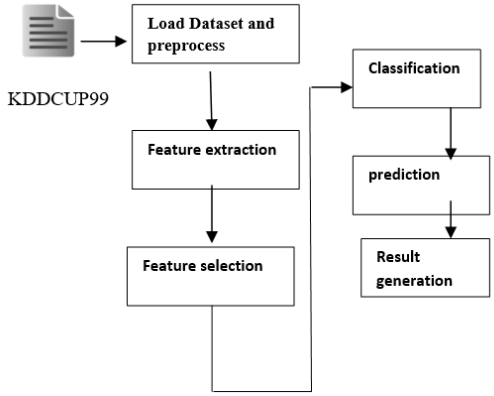

Fig. 2. System architecture

A. KDDCUP99

KDDCUP 99 Data set is a well-known benchmark in the research of intrusion detection techniques. A lot of work is going on for the improvement of intrusion detection strategies while the research on the data used for training and testing the detection model is equality of prime concern because better data quality can improve offline intrusion detection. The 1999 KDD intrusion detection contest uses a version of this dataset. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between bad connection, called intrusion or attacks, and good normal connections. The database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military simulated in military environment.

B. Load dataset and preprocess

Data selection is the process of selecting the appropriate data set for processing. Each of the record consists of 41 features and one marked as either normal or attack. The KDDCUP 99 Dataset is used for detecting the intrusion. All the data's are selected and loaded into the database for detecting the intrusion.

C. Data preprocessing

The data is pre-processed to remove unwanted data that is presented in the dataset. The Incorrect data may provide the incorrect result so that all the data's are cleaned before processing. It is a process of cleaning the data for providing the better result.

D. Feature selection

Feature selection refers to the process of reducing the inputs for processing and analysis, or of finding the most meaningful inputs. The meaningful data's are selected from the extracted features of KDD CUP 99 Dataset.

E. Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. A classification model could be used to identify the normal and attacks from the KDD Dataset. The goal of classification is to accurately predict the target class for each case in the data.

F. Prediction

The goal of classification is to accurately predict the target class for each case in the data. The purpose of this module is to predict the attack from the KDD CUP 99 Dataset.

Prediction in data mining is to identify data points purely on the description of another related data value. It is not necessarily related to future events but the used variables are unknown. It identifies the closely related value. The attacks are predicted from the dataset. It increases the accuracy of the prediction result.

G. Result generation

The overall classification report is generated based on the normal and attack that is presented in the KDDCUP Dataset. The report which contain dynamically distributed reviews about the products. The level that describe the attacks and normal data.

8. Conclusion

In this paper, several experiments were performed and tested to evaluate the efficiency and the performance of the following machine learning classifiers. All the tests were based on the KDD Intrusion detection dataset. The rate of the different types of the attacks the KDD dataset are approximately 79% of DOS attacks, 19% of normal packets and 2% of other types of attacks (R2I, U2R and PORE). In the experiments 148753 instances of records have been extracted as training data to build the training models for selected machine learning classification. The testing phases is implemented based on 60000 random instances of records. Several performance metrics are computed (accuracy rate, precision, false negative, false positive, true negative and true positive.

References

[1] K. Dave, S. Lawrence, and D. M. Pennock, ``Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in Proc. WWW, Budapest, Hungary, 2003.
[2] C. Whitelaw, N. Garg, and S. Argamon, ``Using appraisal groups for sentiment analysis," in Proc. CIKM, Bremen, Germany, 2005.
[3] S. R. Das and M. Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the Web," 2005.

[4] A. Devitt and K. Ahmad, ``Sentiment polarity identification financial, Prague, Czech Republic, 2007, pp. 984-991.
[5] B. Liu, “Sentiment analysis and opinion mining," Synth. Lectures Hman Lang Technol, vol. 5, no. 1, pp. 1-167.
[6] Md Nasimuzzaman Chowdhury & Ken Ferens, Mike Ferens (2016), “Network Intrusion Detection Using Machine Learning”, 2016 Int’l Conf. Security and Management, SAM’16.
[7] Eslamnezhad, Mohsen & Varjani, A. (2014). “Intrusion detection basedon MinMax K-means clustering”, 2014 7th International Symposium on Telecommunications, IST 2014. 804-808.
[8] Li, Jiaqi, Zhifeng Zhao and Rongpeng Li. “A Machine Learning Based Intrusion Detection System for Software Defined 5G Network.” CoRR
[9] Shenfield, Alex & Day, David & Ayesh, Aladdin. (2018). “Intelligent intrusion detection systems using artificial neural networks”, ICT Express.
[10] Jaiganesh, V & Sumathi, P & Mangayarkarasi, S. (2013). “An analysis of intrusion detection system using back propagation neural network”.2013 International Conference on Information Communication and Embedded Systems, ICICES 2013, 232-236.
[11] Ghosal, Amrita & Halder, Subir. (2017). “A survey on energy efficien intrusion detection in wireless sensor networks”. Journal of Ambient Intelligence and Smart Environments, vol. 9, 239-261.
[12] Wu P., Zhao H. (2011) “Some Analysis and Research of the AdaBoost Algorithm”. In: Chen R. (eds) Intelligent Computing and Information Science. ICICIS 2011. Communications in Computer and Information Science, vol. 134. Springer, Berlin, Heidelberg.
[13] Moustafa, Nour & Slay, Jill. (2015), “UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)”.
[14] http://www.statsoft.com/textbook/support-vector-machines/.
[15] https://www.datacamp.com/community/tutorials/support-vevtormachines-r/.
[16] https://www.ritchieng.com/one-variable-linear-regression/.
[17] https://machinelearningmastery.com/what-is-machine-learning/
[18] https://www.techopedia.com/definition/12941/network-based-intrusiondetection-system-nids.