ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

A Survey on Intrusion Detection System Using Data Mining Techniques

Kavitha.N[1], Blessy Boaz[2]

PG Scholar, Department of Computer Science, Stella Maris College, Chennai, India[1]

Assistant Professor, Department of Computer Science, Stella Maris College, Chennai, India[2]

ABSTRACT: Nowadays, an increasing number of populations are accessing the Internet for commercial services which is the major cause for attack. Threats are created everyday by an individual or by the organization that attacks the network system. Unusual Malicious activities and unauthorized access are identified by observing the network in Intrusion Detection System. IDS is a passive monitoring system, it warns for suspicious activity but does not prevent them. Various techniques are proposed in IDS to identify attacks or threats in the system. An Intrusion Detection System (IDS) monitors the network for any suspicious or malicious activity and alarms the network administrator for any kind of vulnerabilities. Its main goal is to protect the data's integrity, confidentiality, or availability of the resources. IDS use data mining techniques to detect intrusions. This survey studies various techniques like k-means clustering, ANN, fuzzy neural network SVM classifiers, AODV based LDK model with AES , k-nearest neighbor classifier, GNS, Naive Bayes, Support vector machines (SVM), J48(C45) with Random Forest, Random Forest with Random Tree Classifiers. All these techniques have been implemented using serial algorithm. IDS have significant overhead in CPU time and memory used by these techniques and so recently researchers are focusing on multi-core CPU to achieve average speedup compared to serial implementation. GPUs act as co-processors which grabs the attention due to their high performance capabilities.

KEYWORDS: Intrusion Detection System (IDS), GPU, IMS, MANET, AODV.

I. INTRODUCTION

An intrusion detection system (IDS) monitors the network for any suspicious or malicious activity occurs and it will alarm the administrator when something happen and also monitors the internet for any recent attacks which could lead to future attacks. The main purpose of IDS is to alarm the administrator when any suspicious activity is detected in the network. Machine learning techniques plays an important role in identifying network intrusions (or attacks), which helps the network administrator take security actions to prevent intrusions. Most of the techniques used in today's IDS are not able to handle the dynamic and complex nature of cyber-attacks on computer networks. Hence, effective and adaptive techniques are used that results in higher detection rates, lower false alarm rates. Machine learning methods provide reasonable computation and communication costs. Pattern matching is fast enough to check the presence of a signature in the incoming packet sequence and detects the malicious behavior and it has to meet with the increase in both the number of signature and the link speed. There are several approaches on the implementation of IDS.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

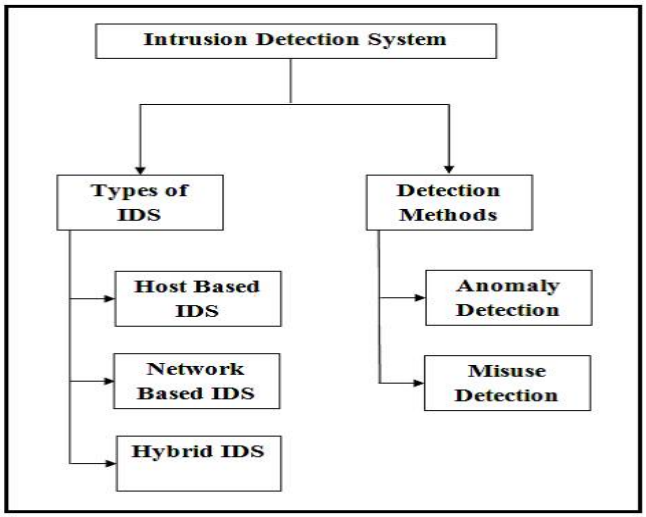

Fig-1: classification of IDS

From the Fig-1 , IDS are classified into three types [2].

1. Host-based IDS (HIDS): Data is collected from single host or different host log records that include audit record of operation system, system logs, application programs information, and so on.

2. Network-based IDS (NIDS): Data collected from the network collective flow to network section, such as: Internet packets.

3. Hybrid IDS(HIDS) : HIDS will monitor the network traffic for a specific host, as well as it will monitors all network traffic like NIDS.

From the Fig-1,Intrusion Detection fall into two category[2]

1. Anomaly-based IDS: It attempts to find malicious activities by measuring the traffic deviation from the expected baseline. It relies on machine learning and artificial intelligence to classify the network traffic into normal or abnormal.

2. Signature-based IDS: Human analysts investigate suspicious traffic, extract features of known intrusions and use the predefined signatures to discover malicious packets.

Anomaly-based IDS relies on machine learning systems which automatically learn from data. Machine learning technique is increasing used since it is attractive and rapidly construct IDS in them. Web search, spam filters, recommender systems, ad placement, credit scoring, fraud detection, stock trading, drug design, and many other applications are using machine learning. Machine learning can be classified as supervised and unsupervised learning methods. There are various kinds of learning methods in the supervised and unsupervised classification of algorithms. Among those the following set of learning methods have been used in this paper from the different IDS papers are K-means Clustering, ANN, Fuzzy neural network, SVM classifiers, AODV based LDK model with AES, K-Nearest Neighbor classifier, GNS, Naive Bayes, Support vector machines (SVM), J48(C45) with Random Forest, Random Forest with Random Tree Classifiers

Signature-based IDS uses pattern matching algorithm, it can be single or multiple pattern matching algorithm. In single pattern matching one pattern is matched against the entire text at a time. In multiple pattern matching algorithm

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

compares the text sequence against all signatures all at once; this consumes more memory and requires a pre-processing phase.

Parallel programming can be used to speed up the process, many computations are performed concurrently. It can utilize multicore processors in a single machine or multi-processors in a cluster of machines. Even GPUs can be used, due to its high performance capabilities. Map Reduce is used to handle the processing of large dataset in parallel.

II. TYPES OF ATTACK

2.1 Denial of service (DOS):
DOS stops the approved user from acquiring the requested services. The distributed denial-of-service attack (DDoS attack), different sources flood the incoming traffic. It is impossible to stop this attack, since it is not possible to block a single source. The types of DOS attacks are:

a) SYN attack: SYN attack is a synchronization attack, to use the resources the attacker will send the request to destination.
b) Ping of death: Sends the request to target system it has 65,536 bytes that may crash the system. The formal size in bytes for ping of death is 56 bytes or 84 bytes.

2.2 *PROBE:* Probe attack removes the data from target machine, it will collect the information from computers to identify the vulnerabilities.
2.3 Remote to local (R2L): In R2L unauthorized users' attempts to get the access from a remote to local machine by guessing password or breaking it.
2.4. User to root (U2R): In U2L rights given is to access the local machine but it gets the access right of administrator. The overflow occurs when web services get more data it may lead to data loss.

III. LITERATURE SURVEY

Intrusion Detection using Data Mining Techniques [2] uses classification tree and support vector machines to detect the intrusion using KDD cup99 dataset. The data is pre-processed before preparation and it is tested. Data includes four main attacks they are Denial-of-service (Dos), Remote-to-local (R2L), User to root (U2R), PROBING. C4.5 algorithm performs better than SVM for both detection and false alarm rate.

To implement a Network Intrusion Detection System [7] K-means Clustering Algorithm is amalgamated with Standard MLP and SVM based Neural Networks that significantly improves the detection of intrusion. From the attacks of DOS, PROBE, R2L, U2L 50% of the record are used for training and 50% for testing. The other set shows that 25% for training and 75% for testing set to check whether they are working properly with less training. In DOS 97.51% and PROBE 98.79% of accuracy for the R2L and U2L gives 98.89% and 98.87% of accuracy. SVM classifier gives more accuracy when compared to other classifier.

Linear regression and K-Means clustering methods are used to identify the network attacks in Network Intrusion detection System using various data mining techniques [11] which generates rules automatically to classify network activities and the accuracy is 80% in detecting the attack using linear regression, 67.5% of accuracy have been noted for K-Means clustering.
Intrusion Detection technique is proposed using K-means, fuzzy neural network and SVM classifiers [3]. Data is clustered, train the neural networks, generate the vectors and then classification is performed to find intrusion. The accuracy achieved is 98.80% for DOS attack, 97.31% for PROBE attack and 97.5% for R2L, U2R attacks.
AODV based LDK model for Intrusion Detection System in MANET (Mobile Adhoc Network) [15] is used for the set of nodes in the network to perform intrusion free and energy efficient routing scheme and sets up data security parameters with the help of Advanced Cryptographic Standard (AES) algorithm. So, that the data transmission can be trustworthy from source to destination in energy efficient manner, it selects path dynamically by using Game Theoretic Modeling. In LDK when the source node starts it will identify the best path and start transmitting data via the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

respective Neighbor Node (NN). If any node interrupts the active nodes it will be considered as a malicious node. Its primary goal is to detect any anomalous behaviour and conserve energy.

Intrusion detection and prevention using light weight virtualization in web applications [8]. Nowadays web services are important in our daily life for communication, because of the increasing use of web services the application have been moved to multitier architecture. In this paper to run applications web server acts as front end and database servers as back end. Attacks in the web services are future session attack, sql injection attack, cross site scripting attack etc. If the attack is certified then the honey pot will block the access privileges to user. To reduce such attacks in web services honey pot is used in anomaly detection to detect the abnormal behaviour and gives the behaviour of legal user. This approach alerts and identifies the attack with minimal false positive rate.

Intrusion Detection System for IP Multimedia Subsystem using K-Nearest Neighbor classifier. In [1] the KNN classifier are based on network anomaly detection mechanism it detects whether the system behaviour is normal or unusual. KNN classifier is used for protecting PCSCF from attacks, when the user communicates to IMS via PCSCF(Proxy Call Session Control Function) although it is expensive it checks each receiving packets and it may lead to jamming problem so care should be taken while implementing it and this approach gives low false positive rate.

Modified AODV Algorithm using Data Mining Process: Classification and Clustering **[**4] verifies for attacks as well as checks node for sending and receiving the data packets. Based on the classification and pattern it will divide the dataset. Although it is very expensive it gives better security and reliability. The detection, prevention and decision making are done in this technique. Normal and intrusion behaviours are formed by K-mediods clustering.

Intelligent Network Intrusion Detection System using Data Mining Techniques [12] in this recent enhancement of naive Bayes algorithm Average one dependence estimators (AODE) is proposed. AODE enhances the attack detection accuracy which is good compared to traditional model and this suits well for incremental learning. Weka tool has been used to show the results using NSLKDD dataset for four attacks. AODE gives high detection rate (DR) and low false alarm rate (FAR) compared to naïve bayes. The accuracy of the proposed algorithm is 7% more for DOS attack, probe 10% more, and R2L, U2R is 8% more than the naïve bayes.

Reducing false positives in intrusion detection systems using data-mining techniques utilizes support vector machines, decision trees, and naïve Bayes for off-line analysis [13]. The proposed algorithm combines SVM, decision tree uses J48 and Naive Bayes, its main concern is to reduce false positive rate which will increase IDS efficiency. The model was tested using Weka open source tool with the split of 66% training testing data. High accuracy is maintained by reducing false positives. SVM is trained using binary classification, decision tree works on unclassified attacks to classify the traffic attack through routing, polling is done by Naive Bayes.

Combination of Data Mining Techniques for Intrusion Detection System [9] is an anomaly based intrusion detection, the major problem is to reduce to high false alarm rate. Extremely effective way of detecting known and unknown attacks is anomaly based intrusion detection system. To measure the performance of the proposed algorithm two parameters: attack detection rate and false alarm rate are used. Combination of classifiers J48 with Random Tree, RandomForest with RandomTree, J48 with RandomForest are used to detect attacks. Weka open source tool is used for experimental results using KDDcup dataset with 41 attributes. It performs better for U2R and R2L type of attack when compare to other classifier and achieves better attack detection rate.

An initial grouping of number of predefined clusters, partitions data points. Partition-based clustering iteratively rearranges data points [10]. Repeated run scan overcome the performance problem that relies on the cluster centroids initial arrangement. Partition points in k-mean clusters tend to near each other, as it has no prior classification and it is unsupervised. Fuzzy clustering is also used by researcher where the predicate logic is precisely deduced than approximate reasoning. Statistics community uses Bayesian clustering. Supervised classification learning technique uses training set with predefined labels to classify data items, it is very difficult to obtain labeled intrusion data like the K-nearest neighbors technique. SVM uses two-class problems, classes cannot be separated in order to solve this problem. Create a hyperplane for linear separation in high-dimensional feature space using a kernel function. CART is widely used as the trees gives the decision rules to update rule-based defences. C4.5 algorithm generates decision trees used for classification which is a statistical classifier. Aprior algorithm is an association rule based data mining technique to mine association rules. Model-based approaches such as neural networks, agent-based and genetic algorithms are used in IDS.

Comparative study of Data Mining Techniques to Enhance Intrusion Detection [5] this study proposes to enhance the intrusion detection by comparing different techniques. The most widely used data mining techniques for IDS are

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

classification, clustering and association rules. Among these preferred technique is clustering and so K-means, Y-means and Fuzzy C-means clustering are analysed. k-means algorithm reduces the speed, ease of use and they are not suited for large dataset. y-mean algorithm when compared to k-mean it eliminate the empty clusters and it overcomes the limitations namely dependency on the number of clusters, degeneracy and dependency on the initial centroids. Fuzzy c-means allows more than one clusters and reduces the objective function that checks the standard of the partition on the fuzzy logic. The performance and efficiency of fuzzy c-means clustering is the best, it is a competent method for the performance measures high detection rate, lower false positive rate and but it is time consuming. Overall performance is based on the initial number of clusters and it works well for small as well as large data sets.

To overcome the traditional problems of FCM, an Improved IDS Model Based on PCA and FCM Algorithms [14] proposes an algorithm PCA-FC, PCA is used to extract features and dimensionality reduction then extract classification using attribute weight fuzzy clustering algorithm. Redundant attributes among the samples are eliminated and extracted using PCA. To build the simulation model machine learning UCI library is adopted. The improved algorithm effectively reduces the dimensionality and speed up the response time. Performance of PCA-FC is analysed with FCM and K-means, PCA-FC has little effect for detection accuracy and false alarm rate, it is suggested that FAR can be the reduced further. Feature extraction process is time consuming which can be ignored. PCAFC has advantage over complete performance.

Intrusion Detection Using Data Mining Techniques [6] survey compares different algorithms, mostly the researches have focused on anomaly detection. ANN is commonly used by researches because it is more reliable than other algorithms. Next to ANN, SVM is preferred by researchers than KNN, decision tree, genetic algorithm. Researches have used DARPA1998 or KDDCup1999 datasets to build IDS technique. K-means clustering gives better accuracy and efficiency. Naïve Bayes is one of the old techniques adopted by researcher. Recently the researches are working with NSL-KDD and KDDCup99 datasets.

Pattern matching of signature-based IDS using Myers algorithm under Map reduce framework [16]. In the existing signature-based IDS the usage of resources such as time and memory is high, to reduce this overhead. Pattern-matching algorithm is parallelized and executed on a multi-core CPU. The parallelism can be bit-level, task-level or data-level, and this approach uses bit-level parallelism. Pattern matching algorithm checks the existence of signature in the arrangement of incoming packets and output the location of the string within the packet, which easily detects the malicious behaviour. This paper proposes different parallel implementation of the Myers algorithm to advance the matching process, this is four times faster than the serial versions and gives better performance than the MPI parallelism. Overall MAPCG and Phoenix++ provide excellent solution for data intensive application. MAPCG and Phoenix++ consumes more memory usage when compared to Myers and MPI.

IV. CONCLUSION

This paper has reviewed and examined the kinds of IDS and classes of techniques used for intrusion detection system, it focuses on machine learning methods which is based on anomaly detection IDS. This survey shows how data mining is used to aid the process of Intrusion Detection and the performance improvement in different methods have been shown by researchers. As future enhancement to this study, in order to improve the speed of processing, these learning methods can be supported with parallel programming environment using coprocessors. Apache introduced Spark as an open-source framework for data processing. Spark processing is faster than MapReduce because of its in-memory sharing. Later, Flink was introduced and it is faster than sparks. IDS uses clustering techniques which performs better than the other techniques. So, clustering techniques can be implemented in parallel processing environment to extend the work in stand-alone computers to speedup the process.

REFERENCES

[1] Ashfaq Hussain Farooqi, Ali Munir, “Intrusion Detection System for IP Multimedia Subsystem Using K-Nearest Neighbor classifier”, IEEE, 2008.
[2] Ektefa M., Memar S., ―Intrusion Detection Using Data Mining Techniques, IEEE Trans., 2010.
[3] A. M. Chandrasekhar, K. Raghuveer, “Intrusion Detection technique by using K-means, fuzzy neural network and SVM classifiers”, IEEE Conference on Computer Communication and Informatics, Jan 04-06, 2013.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

[4] Maheshwari Shashikant, Dr. Sumit Shrivastava, "Modified AODV Algorithm using Data Mining Process: Classification and Clustering", AC IEEE 2013
[5] Mitchell D'silva, Deepali Vora, "Comparative Study of Data Mining Techniques to Enhance Intrusion Detection", *International Journal of Engineering Research and Applications, January -February 2013*
[6] Krishna Kant Tiwari , Susheel Tiwari , Sriram Yadav," Intrusion Detection Using Data Mining Techniques" International Journal of Advanced Computer Technology (IJACT)2013.
[7] A.M. Chandrashekhar K. Raghuveer, "Amalgamation of K-means Clustering Algorithm with Standard MLP and SVM Based Neural Networks to Implement Network Intrusion Detection System", SPRINGER, 2014.
[8] J.Angeline Linora , M.Nava Barathy "Intrusion detection and prevention by using light weight virtualization in web applications",International Journal of Computer Science and Mobile Computing, March- 2014.
[9] Kailas Shivshankar Elekar," Combination of Data Mining Techniques for Intrusion Detection System" IEEE International Conference on Computer, Communication and Control (IC4-2015).
[10] Inadyuti Dutt , Dr. Samarjeet Borah, "Some Studies in Intrusion Detection using Data Mining Techniques", International Journal of Innovative Research in Science, Engineering and Technolog, July 2015 .
[11] DikshantGupta, SuhaniSinghal, Shamita Malik, Archana Singh "Network Intrusion Detection System Using various data mining techniques" International Conference on Research Advances in Integrated Navigation Systems (RAINS - 2016),April 06-07, 2016.
[12] Amreen Sultana , M.A.Jabbar "Intelligent Network Intrusion Detection System using Data Mining Techniques",IEEE 2016.
[13] Kathleen Goeschel," Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naïve bayes for off-line analysis",IEEE 2016.
[14] Gao Yan,"AnImproved IDS Model Based on PCA and FCM Algorithms**",** 2016 International Conference on Smart City and Systems Engineering.
[15] B. Abinaya" An Intrusion Detection System in MANET", International journal of innovative research in science ,engineering and technology ,february 2017.
[16] mother aldwairi ,Ansam M.Abu-Dalo and Moath jarrah," Pattern matching of signature-based IDS using Myers algorithm under Mapreduce framework" Springeropen 2017.