Intrusion Detection System: Classification, Algorithms and Datasets to apply

K.Gayathri[1], Srikanth Yadav.M[2], P.Sravani[3], R.Bhavya Deepthi[4]

[1]*Assistant Professor, Dept. of. CSE, Tirumala Engineering College, Jonnalagadda, NRT, AP, India*
[2]*Associate Professor, Dept. of. CSE, Tirumala Engineering College, Jonnalagadda, NRT, AP, India*
[3, 4]*U.G. Students, Dept. of. CSE, Tirumala Engineering College, Jonnalagadda, NRT, AP, India*

Abstract—

With the escalation of the internet, Security of network traffic is becoming a major problem of computer network system. As time is passing the number of attacks on the network are increasing. Such attacks on network are nothing but the Intrusions. Intrusion detection system has been used for detecting intrusion and to protect the data and network from attacks. Data mining techniques are used to monitor and analyze large amount of network data & classify these network data into anomalous and normal data. Since data comes from various sources, network traffic is large. Data mining techniques such as classification and clustering are applied to build Intrusion detection system. This paper presents the classification of IDS, different Data mining techniques and datasets for the effective detection of pattern for both malicious and normal activities in network, which helps to develop secure information system. Also it provides a brief study of various datasets that are useful for an intrusion detection system.

Keywords— *Data mining; Intrusion Detection System; Anomaly Detection; Misuse Detection; Clustering; Classifications, KDD99, GureKDD, NSL-KDD.*

I. INTRODUCTION

The importance of security problem for the data has been increasing day by day along with the rapid development of the computer network. Security means degree of protection given to the network or system. The main goals of security are confidentiality, Integrity and availability of data [1]. Attacks on network can be referred as Intrusion. Intrusion means any set of malicious activities that attempt to compromise the security goals of the information. Intrusion detection is one of the enormous information security problems. IDS (Intrusion Detection System) assist the system in resisting external attacks.

In early days, only conventional approaches were used for network such as encryption, firewalls, virtual private network etc. but they were not enough to secure network completely. It is difficult to depend completely on static defense techniques. This increases the need for dynamic technique, which can be monitors system and identify illegal activities. Thus to enhance the network security dynamic approach is introduced and known as Intrusion Detection System. Intrusion detection system collects online information from the network after that monitors and analyzes this information and partitions it into normal & malicious activities, provide the result to system administrator [2].

IDS is the area, where Data mining is used extensively, this is due to limited scalability, adaptability and validity. In IDS data is collected from various sources like network log data, host data etc. Since the network traffic is large, the analysis of data is too hard. This give rise to the need of using IDS along with different Data mining techniques for intrusion detection.
This paper is organized as follows. Section 1 gives Introduction. Section 2 discusses about the literature survey. Section 3 overviews the intrusion detection system and its classification. Section 4 gives various data mining techniques for IDS. Section 5 discusses about the various datasets that are useful to build an IDS and the next section is of conclusion.

II. LITERATURE REVIEW

M.Govindarajan et.al.(2009)[7], proposed new K-nearest neighbour classifier applied on Intrusion detection system and evaluate performance in term of Run time and Error rate on normal and malicious dataset. This new classifier is more accurate than existing K-nearest neighbour classifier.

Mohammadreza Ektela et.al.(2010)[8], used Support Vector Machine and classification tree Data mining technique for intrusion detection in network. They compared C4.5 and Support Vector Machine by experimental result and found that C4.5 algorithm has better performance in term of detection rate and false alarm rate than SVM, but for U2R attack SVM performs better.

Deepthy k Denatious et.al.(2012)[1], describe different data mining techniques applied for detecting intrusions. Also describe the classification of Intrusion detection system and its working. For large amount of network traffics, clustering is more suitable than classification in the domain of intrusion detection because enormous amount of data needed to collect to use classification.

P. Amudha et.al.(2011)[11], observed that Random forest gives better detection rate, accuracy and false alarm rate for Probe and DOS attack & Naive Bayes Tree gives better performance in case of U2R and R2L attack. Also the

execution time of Naive Bayes Tree is more as compared to other classifier.

R. China Appala Naidu et.al.(2012)[12], used three Data mining techniques SVM, Ripper rule and C5.0 tree for Intrusion detection and also compared the efficiency. By experimental result, C5.0 decision tree is efficient than other. All the three Data mining technique gives higher than 96% detection rate.

Roshan Chitrakar et.al.(2012)[13], proposed a hybrid approach to intrusion detection by using k-Medoids clustering with Naïve Bayes classification and observed that it gives better performance than K-Means clustering technique followed by Naïve Bayes classification but also time complexity increases when increase the number of data points.

Roshan Chitrakar et.al.(2012)[14], proposed a hybrid approach of combining k-Medoids clustering with Support Vector Machine classification technique and produced better performance compared to k-Medoids with Naïve Bayes classification. The approach shows improvement in both Accuracy and Detection Rate while reducing False Alarm Rate as compared to the k- Medoids clustering approach followed by Naïve bayes classification technique.

III. PROPOSED SYSTEM

First we will see what is intrusion. Intrusion can be defined as any set of actions that threatens the integrity, availability, or confidentiality of a network resource. So on the basis of this Intrusion detection can be defined as the act of detecting actions that attempt to compromise the confidentiality, integrity or availability of a resource (1). The goal of intrusion detection is to identify entities attempting to subvert in-place security controls.

An intrusion detection system (IDS) is a device or software application that monitors a network or systems for malicious activity or policy violations. Any detected activity or violation is typically reported either to an administrator or collected centrally using a security information 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.

A. Common types of Intrusion Detection:

There is a wide spectrum of IDS, varying from antivirus software to hierarchical systems that monitor the traffic of an entire backbone network. The most common classifications are network intrusion detection systems (NIDS) and host-based intrusion detection systems (HIDS).

B. Network Based (Network IDS):

Network based intrusion detection attempts to identify unauthorized, illicit, and anomalous behavior based solely on network traffic. A network IDS, using either a network tap, span port, or hub collects packets that traverse a given network. Using the captured data, the IDS system processes and flags any suspicious traffic. Unlike an intrusion prevention system, an intrusion detection system does not actively block network traffic. The role of a network IDS is passive, only gathering, identifying, logging and alerting. Examples of Network IDS: SNORT

C. Host Based (HIDS):

Often referred to as HIDS, host based intrusion detection attempts to identify unauthorized, illicit, and anomalous behavior on a specific device. HIDS generally involves an agent installed on each system, monitoring and alerting on local OS and application activity. The installed agent uses a combination of signatures, rules, and heuristics to identify unauthorized activity. The role of a host IDS is passive, only gathering, identifying, logging, and alerting. Examples of HIDS: OSSEC (Open Source Host-based Intrusion Detection System), Tripwire, AIDE (Advanced Intrusion Detection Environment), Prelude Hybrid IDS.

D. Physical (Physical IDS):

Physical intrusion detection is the act of identifying threats to physical systems. Physical intrusion detection is most often seen as physical controls put in place to ensure CIA. In many cases physical intrusion detection systems act as prevention systems as well. Examples of Physical intrusion detections are: Security Guards, Security Cameras, Access Control Systems (Card, Biometric), Firewalls, Man Traps, and Motion Sensors.

IV. CLASSIFICATION OF INTRUSION DETECTION SYSTEMS

It is also possible to classify IDS by detection approach:

A. Signature-based detection:

It is also known as misuse detection. So misuse detection is Signature based IDS where detection of intrusion is based on the behaviors of known attacks like antivirus software. Antivirus software compares the data with known code of virus. In Misuse detection, pattern of known malicious activity is stored in the dataset and identify suspicious data by comparing new instances with the stored pattern of attacks.

B. Anomaly-based detection:

It is different from Misuse detection. Here baseline of normal data in network data in network for example, load on network traffic, protocol and packet size etc is defined by system administrator and according to this baseline, Anomaly detector monitors new instances. The new instances are compared with the baseline, if there is any deviation from baseline, data is notified as intrusion. For this reason, it is also called behavior based Intrusion detection system.

V. DATA MINING TECHNIQUES FOR INTRUSION DETECTION

There are many data mining techniques for intrusion detection such as, frequent pattern mining, classification,

clustering, mining data streams, etc. Let us see some of them here. Classification is the task of taking each and every instances of dataset under consideration and assigning it to a particular class normal and abnormal means known structure is used for new instances. It can be effective for both misuse detection and anomaly detection, but more frequently used for misuse detection. Classification categorized the datasets into predetermined sets. It is less efficient in intrusion detection as compared to clustering. Different classification techniques such as decision tree, naive bayes classifier, K-nearest neighbour classifier, Support vector machine etc. are used in IDS.

A. Decision Tree:

Decision tree [21] is a recursive and tree like structure for expressing classification rules. It uses divide and conquer method for splitting according to attribute values. Classification of the data proceeds from root node to leaf node, where each node represents the attribute and its value & each leaf node represent class label of data. Tree based classifier have highest performance in case of large dataset. Different decision tree algorithms are described below [6]:

B. ID3 algorithm:

It is famous decision tree algorithm developed by Quinlan. ID3 algorithm basically attribute based algorithm that constructs decision tree according to training dataset. The attribute which has highest information gain is used as a root of the tree.

C. J48 algorithm:

It is based on ID3 algorithm and developed by Ross Quinlan. In WEKA, C4.5 decision tree algorithm is known as J48 algorithm. It constructs decision tree using information gain, attribute which have highest information gain is selected to make decision. The main disadvantage of this algorithm is that it takes more CPU time and memory in execution. Another different tree based classifier [15].

D. AD Tree:

Alternating decision tree is used for classification. AD Tree have prediction node as both leaf node and root node.

E. Random Forest:

Random Forest [22] is first introduced by Lepetit et.al. and it is ensemble classification technique which consists of two or more decision trees. In Random Forest, every tree is prepared by randomly select the data from dataset. By using Random Forest improve the accuracy and prediction power because it is less sensitive to outlier data. It can easily deal with high dimensional data.

F. K-Nearest Neighbor:

It is one of the simplest classification technique. It calculates the distance between different data points on the input vectors and assigns the unlabeled data point to its nearest neighbor class. K is an important parameter. If k=1, then the object is assigned to the class of its nearest neighbor. When value of K is large, then it takes large time for prediction and influence the accuracy by reduces the effect of noise.

VI. CLUSTERING

Since the network data is too huge, labelling of each and every instances or data points in classification is expensive and time consuming. Clustering is the technique of labelling data and assign into groups of similar objects without using known structure of data points. Members of same cluster are similar and instances of different clusters are different from each other. Clustering technique can be classified into four groups: Hierarchical algorithm, partitioning algorithm, Grid based algorithm and Density based algorithm. Some clustering algorithms are explained here.

A. K-Means Clustering algorithm:

K-Means clustering algorithm [23][13] is simplest and widely used clustering technique proposed by James Macqueen. In this algorithm, number of clusters K is specified by user means classifies instances into predefined number of cluster. The first step of K-Means clustering is to choose k instances as a center of clusters. Next assign each instances of dataset to nearest cluster. For instance assignment, measure the distance between centroid and each instances using Euclidean distance and according to minimum distance assign each and every data points into cluster. K –Means algorithm takes less execution time, when it applied on small dataset. When the data point increases to maximum then it takes maximum execution time. It is fast iterative algorithm but it is sensitive to outlier and noise.

B. K-Medoids clustering algorithm:

K-Medoids [13] is clustering by partitioning algorithm as like as K-means algorithm. The most centrally situated instance in a cluster is considered as centroid in place of taking mean value of the objects in K-Means clustering. This centrally located object is called reference point and medoid. It minimizes the distance between centroid and data spoints means minimize the squared error. KMedoids algorithm performs better than K-Means algorithm when the number of data points increases to maximum. It is robust in presence of noise and outlier because medoid is less influenced by outliers, but processing is more expensive.

C. Data Set for Experimental Study

The objective of 1999 KDD intrusion detection contest is to create a standard dataset for survey and evaluate research in intrusion detection which is prepared and managed by MIT, Lincoln Labs by DARPA Intrusion Detection Evaluation Program. After capturing nine weeks of raw TCP dump data for LAN simulating a typical U.S. Air Force LAB. They operated the LAN as if it were a true Air Force environment, but peppered it with multiple attacks.

The raw training data contains four gigabytes of compressed binary TCP dump data from seven weeks of network traffic by processed into about five million connection records. Similarly, the test data yielded around two million connection records which are captured last two weeks of the experiment. The KDD dataset was used in the UCI KDD1999 competition. The objective of the competition is to develop intrusion

detection system models to detect attack categories i.e. DOS, PROBE, R2L and U2R.

VII. CONCLUSION

On the basis of detection rate, accuracy, execution time and false alarm rate, the paper has analysed different classification and clustering data mining techniques for intrusion detection. According to given necessary parameter, execution time of Support vector machine is less and produces high accuracy with smaller dataset, while construction of Naive Bayes classifier is easy. Also decision tree has high detection rate in case of large dataset. In clustering techniques, execution time of KMeans clustering algorithm is less in case of small dataset, but when number of data point increases, K-Medoids performs better.

The NSL-KDD data set is the refined version of the KDD cup99 data set. Many types of analysis have been carried out by many researchers on the NSL-KDD dataset employing different techniques and tools with a universal objective to develop an effective intrusion detection system. An exhaustive analysis on various data sets like KDD99, GureKDD and NSL-KDD are made in using various data mining based machine learning algorithms like Support Vector Machine (SVM), Decision Tree, K-nearest neighbour, K-Means and Fuzzy C-Mean clustering algorithms.

VIII. REFERENCES

[1] Deepthy K Denatious & Anita John, "Survey on Data Mining Techniques to Enhance Intrusion Detection", International Conference on Computer Communication and Informatics (ICCCI 2012), Jan. 10 – 12, 2012, Coimbatore, INDIA

[2] Rung-Ching Chen, Kai-Fan Cheng and Chia-Fen Hsieh, "Using Rough Set And Support Vector Machine For Network Intrusion Detection", International Journal of Network Security & Its Applications (IJNSA), Vol 1, No 1, April 2009

[3] Deepak Upadhyaya and Shubha Jain, "Hybrid Approach for Network Intrusion Detection System Using K-Medoid Clustering and Naïve Bayes Classification", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 1, pp 231-236, May 2013

[4] Xiang, M.Y. Chong and H. L. Zhu, "Design of Multiple-level Tree classifiers for intrusion detection system", IEEE conference on Cybernetics and Intelligent system, 2004

[5] Peddabachigiri S., A. Abraham., C. Grosan and J. Thomas, "Modeling of Intrusion Detection System Using Hybrid intelligent systems" , Journals of network computer application, 2007

[6] Mrutyunjaya Panda and Manas Ranjan Patra, "A Comparative Study Of Data Mining Algorithms For Network Intrusion Detection", First International Conference on Emerging Trends in Engineering and Technology, pp 504-507, IEEE, 2008

[7] M.Govindarajan and Rlvl.Chandrasekaran, "Intrusion Detection Using k-Nearest Neighbor" pp 13-20, ICAC, IEEE, 2009

[8] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi and Lilly Suriani Affendey, "Intrusion Detection Using Data Mining Techniques", pp 200-203, IEEE, 2010

[9] Song Naiping and Zhou Genyuan, "A study on Intrusion Detection Based on Data Mining", International Conference of Information Science and Management Engineering , Pp 135- 138, IEEE,2010

[10] T. Velmurugan and T. Santhanam, "Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points", Journal of Computer Science 6 (3): 363-368, 2010

[11] P Amudha and H Abdul Rauf, "Performance Analysis of Data Mining Approaches in Intrusion Detection", IEEE, 2011. [12] R.China Appala Naidu and P.S.Avadhani, "A Comparison of Data Mining Techniques for Intrusion Detection", International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), pp-41-44, IEEE, 2012

[13] Roshan Chitrakar and Huang Chuanhe, "Anomaly based Intrusion Detection using Hybrid Learning Approach of combining kMedoids Clustering and Naïve Bayes Classification", IEEE,2012

[14] Roshan Chitrakar and Huang Chuanhe, "Anomaly Detection using Support Vector Machine Classification with k-Medoids Clustering", IEEE, 2012

[15] Sumaiya Thaseen and Ch. Aswani Kumar, "An Analysis of Supervised Tree Based Classifiers for Intrusion Detection System", International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), IEEE, February 21-22 2013

[16] David Ndumiyana, Richard Gotora and Hilton Chikwiriro, "Data Mining Techniques in Intrusion Detection: Tightening Network Security", International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 5, May – 2013

[17] Muhammad K. Asif, Talha A. Khan,Talha A. Taj, Umar Naeem and Sufyan Yakoob, " Network Intrusion Detection and its Strategic Importance", Business Engineering and Industrial Applications Colloquium(BEIAC), IEEE, 2013

[18] Kapil Wankhade, Sadia Patka and Ravindra Thools, "An Efficient Approach for Intrusion Detection Using Data Mining Methods", IEEE 2013

[19] Fatin Norsyafawati Mohd Sabri, Norita Md Norwawi and Kamaruzzaman Seman, "Hybrid of Rough Set Theory and Artificial Immune Recognition System as a Solution to Decrease False Alarm Rate in Intrusion Detection System", IEEE 2011

[20] Vaishali B Kosamkar and Sangita S Chaudhari, "Data Mining Algorithms for Intrusion Detection System: An Overview", International Conference in Recent Trends in Information Technology and Computer Science, 2012

[21] Hind Tribak , Blanca L. Delgado-Marquez, P.Rojas, O.Valenzuela, H. Pomares and I. Rojas, " Statistical Analysis of Different Artificial Intelligent Techniques applied to Intrusion Detection System", IEEE, 2012

[22] S. Revathi and A. Malathi, "Data Preprocessing for Intrusion Detection System using Swarm Intelligence Techniques", International Journal of Computer Applications , Volume 75– No.6, August 2013

[23] Iwan Syarif, Adam Pruge Bennett and Gary Wills, "Unsupervised clustering approach for network anomaly detection", IEEE.

[24] Wikipedia: http://en.wikipedia.org/wiki/Intrusion_detection

[25] J.S. shanthini, Dr. S. Rajalakshmi, "Data Mining Techniques For Efficient Intrusion Detection System: A Survey", International Journal On Engineering Technology and Sciences – IJETS™ ISSN(P): 2349-3968, ISSN (O): 2349-3976 Volume II, Issue XI, November – 2015

[26] Abhaya, Kaushal Kumar, Ranjeeta Jha, Sumaiya Afroz, "Data Mining Techniques for Intrusion Detection: A Review", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 6, June 2014.

[27] Matthew V. Mahoney, Philip K. Chan, "An Analysis of the 1999 DARPA/Lincoln Laboratory Evaluation Data for Network Anomaly Detection", 6th International Symposium, RAID 2003, Pittsburgh, PA, USA, September 8-10, 2003. Proceedings, Copyright Holder Springer-Verlag Berlin Heidelberg.