A Review Paper on Intrusion Detection System

M Naga Surya Lakshmi [1], Dr. K V N Sunitha [2]

[1]*Research Scholar, Department of Computer Science, Rayalaseema University, A.P., India*

[2]*Professor, Department of Computer Engineering, BVRITCEW, Telangana, India.*

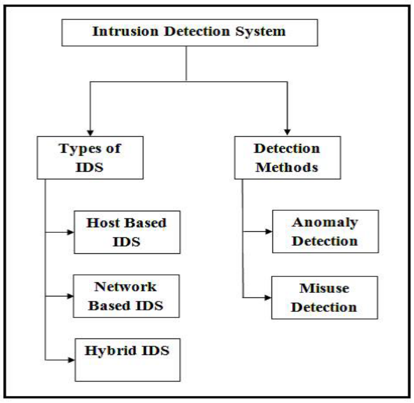
Abstract- Today it is very important to provide a high level security to protect highly sensitive and private information. In Network Security, Intrusion Detection System is an essential technology. In recent days researchers are working on intrusion detection system using Data mining techniques as a skill. IDS is a device that collects information from variety of systems, network sources for dealing attacks. It also analyses symptoms of security problems. This paper focus mainly on complete study and growth of IDS system and various approaches proposed by variant authors. Researchers are focusing on multi-core CPU to achieve average speedup compared to serial implementation because of significant overhead in CPU time and memory when IDS is used. GPUs due to high performance capabilities act as co-processors.

Keywords- IDS, Intrusion Detection System , GPU, component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Computers Are Always At Risk Against Un-Authorization And Intrusion. Everyone Is Using Internet For Commercial Services Which Are The Major Cause For Attack. So Authentication Is Of Prime Concern. To Avoid Threats And Intruders Every Aspect Of Security Has To Be Overlooked In Every Transaction. According To Internet User Growth, Numbers Of Intruders Are Increasing Day By Day. New Techniques Should Be Introduced To Detect These Intrusions. In Information Security, Intrusion Detection Is A Major Technique. It Detects Attacks And Secures The Network System. Intrusion Detection Is The Process Of Identifying Security Problems Through Observation And Analyses Of Events Arising In A Network System. Monitoring, Detecting And Responding To Unauthorized Activities Are The Major Security Functions Provided By Ids.Ease Of Use

FIG.1: Classification Of Ids



II. CLASSIFICATION OF IDS

IDS is Classified into three categories. They are Host-based (HIDS), Network-based (NIDS) and Hybrid IDS. Supervision of individual systems is done in Host-based IDS which has small programs (or agents) installed in the systems. These installed programs monitor and then write data to log files and trigger alarms. NIDS consists of Network application (or sensor) with a Network Interface Card (NIC) working in special mode and a separate management of interface. Hybrid IDS (HIDS) monitors the network traffic like NIDS for a specific host. IDS are placed on a network segment or boundary to monitor all traffic on that segment. Combination of HIDS and NIDS is current trend in Intrusion detection. Hybrid systems are developed in this combination that is more efficient.

III. DETECTION METHODS

They are classified as follows:

- Anomaly-based IDS monitors network traffic and compare it against an established baseline which identifies the normal part of network, sort of bandwidth used, protocols, ports and devices connected to each other and alert the administrator or user. If there is any deviation from baseline, data is notified as intrusion. It is also called behaviour based Intrusion detection system.
- Misuse IDS analyze the gathered information to compare it with large databases of attack signatures. If the attack is already been registered, then IDS look for specific attack. Misuse detection software is only as good as database that it uses to compare packets.

IV. WORKING OF IDS

- Data Acquisition: It is the collection of data using particular software from different sources.
- Feature Selection: After data collection this step takes place. Dataset for IDS is large. So to work on large dataset generate feature vectors which contains only necessary data.
- Analysis: Collected data is analysed to determine whether it is suspicious or not. Various Data mining techniques are used for Intrusion detection.
- Action: After detecting attack, IDS alarms the administrator.

V. PERFORMANCE MEASUREMENT OF IDS

- There are some primary factors which are used during performance measurement of Intrusion detection system.

- True positive (TP): During Intrusion detection process the total number of normal data which are detected as normal data is TP.
- True negative (TN): In Intrusion detection THE total number of detected abnormal data which are actually abnormal data is TN.
- False positive (FP): These are detected as normal data but they are actual attack.
- False negative (FN): These are detected as abnormal instances but are normal data.

IDS performance is measured in terms of detection rate, accuracy and false alarm rate.

- **Detection Rate (DR) =**
- **False Alarm Rate (FAR)=** $\dfrac{FP}{Number\ of\ Attacks}$
- **Accuracy =** $\dfrac{TP+TN}{\mathrm{TP+TN+FP+F}}$

VI. LITERATURE SURVEY

Dalian et al. [1]suggested classification techniques and current literature surrounding and methods of intrusion detection. they using following datasets following datasets; DARPA, KDD 99, NSL, KDD, Kyoto 2006 + for reviewing current IDS approaches and CAIDA. Findings suggest that NSL–KDD performed best overall once trained against specified classifiers. The authors conclude by suggesting consideration must be taken during developing of classification techniques identifying optimal dataset that is a rich the recent attacks and which features are selected without confusion, unnecessary overhead and time-consuming selection. Performance Analysis of Dimension Reduction Techniques with Classifier Combination for Intrusion Detection System was proposed by Chauhan and Bahl [2]. A Review of current dimension reduction techniques, search methods, attribute evaluators and classifiers was conducted. they applied Different combinations feature selection and feature classification algorithms on datasets to detect intrusion. and they observed increase in classification accuracy from 52% to 96% of PCA analysis. authors suggest that classification with a good accuracy results in a reduction in completion time and effective outputs .Genetic Algorithm based Feature Selection Approach for Effective Intrusion Detection System was proposed by Desale and Ade [3]. They are using NSL-KDD dataset in this data set The genetic algorithm is used to search method when selecting features. For selecting features The mathematical intersection principle is used that appear in every experiment. The experiment is carried out on both test and training data set, proposed approach is measured against the popular approaches, namely the Correlation Feature Selection (CFS), Information gain (IG), Correlation Attribute Eval (CAE) and the effect on the performance of the Naive Bayes and J48T algorithm classifiers are measured. Finally they concluded that the proposed model selected the minimum features from the dataset, which improved the classifier, accuracy of the Naive Bayes classifier AND also reducing complexity and time. Manish and Hadi [4] proposed that clustering is the process of splitting data into clusters based upon the features of the data. This clustering partitions data into groups of similar objects. Each member within the cluster is similar to one another Cabrera et al. [5] classified the modus operandi that suspects used in the commission of crimes AND suggested a sequential pattern mining to identify attack patterns that hackers frequently submit. Srinivas Mukkamala [6, 7] has given idea about Support Vector Machines: SVM first maps the input vector into a higher dimensional feature space and then obtain the optimal separating hyper-plane in the higher dimensional feature space. An SVM classifier gives better result for binary classification. Generalized approach depends on geometrical characteristics of given training data. Training data is transformed into feature space of a huge dimension by this procedure which means the training vectors are separated into 2 different classes.

VII. TYPES OF ATTACKS

Denial-of-service (Dos), Remote-to-local (R2L), User to root (U2R) and probing are the 4 main attacks.

Denial of service (DOS): It rejects approved user from acquiring the requested services. In distributed denial-of-service (DDoS) attack, different sources flood the incoming traffic. Since it is not possible to block single source this attack cannot be stopped. 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: It's formal size is 56 or 84 bytes. It has 65,536 bytes and sends request to target system that may crash the system.

Probe: This attack identifies vulnerabilities by collecting information from computers. This attack also removes the data from target machine

.

Remote to local (r2l): in this by breaking passwords unauthorized users attempt access a local machine from a remote.

User to root (u2r): In U2L rights given are to access the local machine but it gets the access right of administrator. When web services get more data overflow occurs which leads to loss of data.

VIII. INTRUSION DETECTION BY USING DATA MINING TECHNIQUES

K-means Clustering Algorithm is amalgamated with SVM and Standard MLP based Neural Networks to implement a [7] network intrusion detection system that significantly improves the detection of intrusion from the attacks of DOS, PROBE, R2L, U2L, here the training data 50% and testing data is 50%. Some other set to check whether they are working properly with less training that is 25% for training and 75% for testing. 97.51% In DOS and 98.79% in PROBE, the R2L and U2L gives 98.89% and 98.87% of accuracy. When compared to other classifier SVM classifiers give more accuracy. Linear regression and K-Means clustering methods are used in Network Intrusion detection for identifying the attacks, the accuracy for linear regression is 80% and K-Means clustering the accuracy is 67.5%

IX. GOALS OF DATA MINING:

Widely Speaking, the goals of Data Mining fall into the following groups: prediction, identification, classification and optimization

a)prediction: Relationship between Dependent and independent variables is discovered by prediction. The future behavior of a particular attribute within a data is showed by DATA mining. In some application, business logic is coupled with data mining.

b)identification: Data patterns are using to identify the existence of an item, an event, or an activity or some new patterns of customer behavior. The area known as authentication is a layout of identification.

c)classification: To display data in better way, data is classified into different classes based on combination of parameters. That is separation or classification of Data is done by Data Mining

.

d)optimization: Time, space, money are limited resources. These can be optimized by Data Mining. under certain constraints it also maximizes output variables.

X. ADVANTAGES OF DATA MINING:

Data mining applications are continuously developing in various industries to provide more hidden knowledge that enable to increase business efficiency and grow businesses. Data mining approaches plays an essential role in various domains. A large amount of historical data has to be analyzed for the classification of security problems. Since the data is big or enormous, is difficult to anyone to find a pattern but Data mining seems to overcome this problem and can therefore be used to discover those patterns.

XI. DATA MINING TECHNIQUES

Data mining, also known as knowledge discovery, is the process of analyzing and summarizing data from different perspectives and converting it into useful information which helps in taking certain decisions. It helps in finding correlations among dozens of fields present in the database. The different data mining techniques that are used for detecting intrusions are:

a)k-means: K-Means algorithm groups ‘n’ instances into k disjoint clusters (k is a predefined parameter). Each instance is assigned to its nearest cluster. Use Euclidean distance to measure the distance between centroid and each instance. According to minimum distance, assign each and every data points into cluster. When applied on small dataset, this algorithm takes less execution time. Similarly it takes more time when data point increases

b)k-nearest neighbor (knn): One of the simplest classification techniques used to calculate the distance between different data points on the input vectors. After the classification the nearest neighbor class is assigned with this unlabeled data point. K is an important parameter. Nearest neighbor class is assigned if k value is 1. When value of K is large, then it takes large time for prediction and influence the accuracy by reduces the effect of noise.

*c)k-medoids:*K-Medoids is clustering by partitioning algorithm just like K-means algorithm. Centroid is considered as reference point which is centrally located instance in a cluster. It minimizes the distance between centroid and data points which means minimizing the squared error. This is better than K-means algorithm when data points are increasing. Medoid is less influenced by outliers so it is robust in presence of noise but more costly in processing.

d)EM-clustering: It is the Expectation- Maximization algorithm. Weight represents the probability of membership so in this iterative approach rather than assigning the object to the dedicated cluster, the object is assigned to a cluster according to a weight which means in between clusters there is no strict boundaries.

e)classification tree: In machine learning, Classification Tree is also known as Decision tree or predictive model. It is tree like structure in which the internal nodes represent the test condition and branch represents the result. The most common algorithm of this kind are C4.5, CART etc.

f)C4.5: C4.5 constructs decision trees using information entropy concept from a set of available training data. At each node of the tree, the algorithm selects the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Normalized information gain is splitting criterion. Decision is made by the attribute which has highest normalized information gain. This algorithm then recurs on the smaller sub lists.

g)cart: Classification and regression trees (CART) are machine-learning methods for constructing prediction models from data. These models are obtained through recursively partitioning the data and fitting a prediction model within each partition due to which the partitioning can be represented graphically as a decision tree. Classification trees are designed for variables that are dependent and that take a finite number of unordered values, with prediction error measured in terms of misclassification cost. Regression trees are for dependent variables that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between the observed and predicted values.

h)support vector machine: Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression analysis. This algorithm plots each data item as a point in n-dimensional space (where n is number of features available) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the two classes clearly. SVM is independent of feature space it is less susceptible for over fitting of the feature input from the input items. This is the main significance of SVM. Classification accuracy with SVM is quite impressive. SVM is fast accurate while training as well as during testing.

XII. CONCLUSION AND FUTURE SCOPE

This survey paper focuses on various approaches defined by variant authors. Which gives a comprehensive survey on the existing techniques for creating an effective intrusion detection system? Due to enormous growth in the field of computing and networking, the complexity in handling security attacks in the form either internal or external attacks due to the intrusions. In real time applications like online banking services, e-commerce applications, clinical data preservation, sensitive customer information preservation, and attacks in telecommunication field etc. In future, we would like to propose an effective intrusion detection system for real time applications.

ACKNOWLEDGMENT

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude to my guide Dr K V N Sunita whose contribution in stimulating suggestions and encouragement, helped me to coordinate my paper. I Would like to express my sincere thanks to Head of the Department, Computer science Department, Rayalaseema University. Furthermore I would also like to acknowledge to Director , Research, Rayalaseema University.

XIII. REFERENCES

[1]. B. Dhafian, I. Ahmad and A. AL-Ghamid, “An Overview of the current Classification Techniques in Intrusion Detection,” in International Conference Security and Management, 2015, pp. 82-88 .

[2]. H. Chauhan, V. Kumar, S. Pundir and E.S. Pilli, “A Comparative Study of Classification Techniques for Intrusion Detection,” in Computational and Business Intelligence, International Symposium, 2013, pp. 40-43. IEEE.

[3]. K.S. Desale and R. Ade “Genetic Algorithm based Feature Selection Approach for Effective Intrusion Detection System,” in Computer Communication and Informatics, 3rd International Conference, 2015, pp. 1-6.

[4]. R.J. Manish, and H.T. Hadi, “A review of network traffic analysis and prediction techniques” unpublished.

[5]. J.B.D.Cabrera,L.Lewis,andR.K.Mehra,“Detectionand Classification of Intrusion and Faults Using Sentences of System Calls,” SIGMOD Record,vol.30,no.4,2001, pp.25-34

[6]. Srinivas Mukkamala, Andrew H. Sung, Ajith Abraham (2004). Intrusion Detection Using an Ensemble of Intelligent Paradigms .

[7]. V. Vapnik (1998). Statistical Learning Theory. New York: John Wiley

AUTHOR PROFILE

M. Naga Surya Lakshmi,Reserch Scholar , Department of Computer Science & Engineering, Rayalaseema University, Kurnool . Her area of specialization Data mining & Network security. She worked projects on Securing User-Controlled Routing Infrastructures, Hospital Management and Warehouse Management . She attended number of workshops, seminars , some are “Active Research (WAR-2010”) “Net work Programming and simulators, “Research perspectives on artificial intelligence- A neural Network approach, Professional certification Program From IBM “ IBM Certified Date base Associate DB2 9 Fundamentals” and “IBM certified Associate Developer Rational Application Developer for web sphere Software V6.0” . She published paper on “Hardware Enhanced association rule mining with hashing and pipelining “presented in International conference on communication and computation control on Nano technology”