ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

A Survey on Intrusion Detection using Data Mining Technique

D. Shona, A.Shobana

Assistant Professor, Dept. of Computer Science, Sri Krishna Arts & Science College, Coimbatore, India [1]

M.Phil. Scholar, Dept. of Computer Science, Sri Krishna Arts & Science College, Coimbatore, India[2]

ABSTRACT: Security and privacy of a system is compromised, when an intrusion happens. Intrusion Detection System (IDS) plays vital role in network security as it detects various types of attacks in network. Implementation of an IDS is distinguishes between the traffic coming from clients and the traffic originated from the attackers or intruders, in an attempt to simultaneously mitigate the problems of throughput, latency and security of the network. For this reason this survey provides better approaches using data mining techniques. By applying Data Mining techniques on network traffic data is a promising solution that helps develop better intrusion detection systems. Therefore this paper provides Survey of various techniques of Intrusion detection system applied in the data mining. In this paper we present an Intrusion Detection System in data mining with different techniques.

KEYWORDS*: Intrusion Detection System, Security, Privacy, Data Mining and Survey.*

I.INTRODUCTION

Even with today's advanced computer technologies (e. g., machine learning and data mining systems) extracting data is critical one where data set are widely used in machine learning and data mining (DM) tasks. In Data Mining, discovering knowledge from data can still be fiendishly hard due to the characteristics of the computer generated data. Data mining is one of the knowledge extraction process used to discover the knowledge from large datasets and convert it into useful information. In large dataset data are represented in feature value. Size of dataset may be measured in no. of features and no. of instance. As the number and size of the data were increased, these data were transferred through Network [1]. Through this method Internet traffic increase and the need for the intrusion detection grows in step to reduce the overhead required for the intrusion detection and diagnosis, it has made public servers increasingly vulnerable to unauthorized accesses and incursion of intrusions. In addition to maintaining low latency and poor performance for the client, filtering unauthorized accesses has become one of the major concerns of a server administrator. Hence implementation of an Intrusion Detection System (IDS) distinguishes between the traffic coming from clients and the traffic originated from the attackers or intruders, in an attempt to simultaneously mitigate the problems of throughput, latency and security of the network [2]. The exponential growth of computer/network attacks is becoming more and more difficult to identify and the need for better and more efficient intrusion detection systems increases in step. The main problem with current intrusion detection systems is high rate of false alarms. As network-based computer systems have important roles in modern society, they have become the targets of intruders. Therefore, we need to find the best possible ways to protect our systems. The security of a computer system is compromised when an intrusion takes place.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

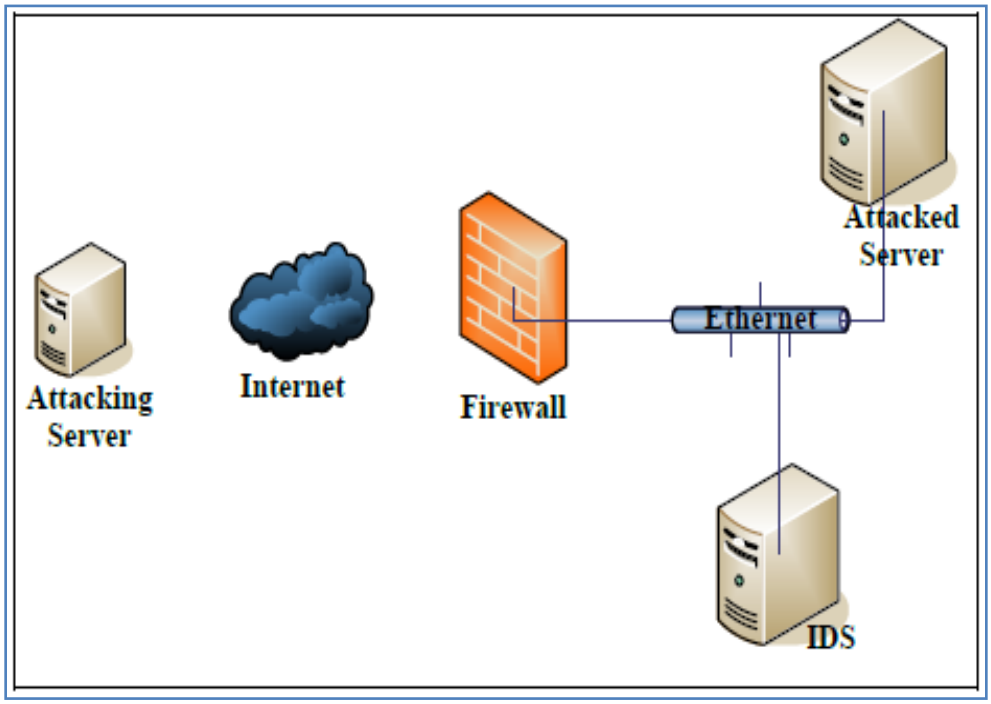

Fig 1: IDS Architecture

An intrusion can be defined as any action done to hamper the integrity, confidentiality or availability of the system. There are some intrusion prevention techniques which can be used to protect computer systems as a first line of defense. But only intrusion prevention is not enough. As systems become more complex, there are always exploitable weaknesses in the systems due to design and programming errors, or various penetration techniques [3]. Data mining techniques can be used for misuse and anomaly intrusion detection. Misuse refers to known attacks and harmful activities that exploit the known sensitivities of the system. In misuse detection, each instance in a data set is labeled as ―normal‖ or‖ intrusion‖ and a learning algorithm is trained over the labeled data. Anomaly means a usual activity in general that could indicate an intrusion. Because of Misuse IDS suffer from a number of major drawbacks; first, known intrusions have to be hand coded by experts. Second, signature library needs to be updated whenever a new signature is discovered, network configuration has been changed, or a new software version has been installed. Third, misuse IDS are unable to detect new (previously unknown) intrusions that do not match signatures; they can only identify cases that match signatures. Thus, the system fails to identify a new event as an intrusion when it is in fact an intrusion, this is called false negative. On the other hand, current anomaly detection systems suffer from high percentage of false positives (i.e., an event incorrectly identified by the IDS as being an intrusion when it is not). An additional drawback is that selecting the right set of system features to be measured is ad hoc and based on experience. Traditional intrusion detection systems are limited and do not provide a complete solution for the problem. In addition, they require exhaustive manual processing and human expert interference. Applying Data Mining (DM) techniques on network traffic data is a promising solution that helps develop better intrusion detection systems. Therefore this paper provides Survey of various techniques of Intrusion detection system applied in the data mining. So here we present an Intrusion Detection System in data mining with different techniques [4].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

II.LITERATURE REVIEW

Huy Anh Nguyen and Deokjai Choi [5] evaluate that a network attacks have increased in number and severity over the past few years, intrusion detection system (IDS) is increasingly becoming a critical component to secure the network. Due to large volumes of security audit data as well as complex and dynamic properties of intrusion behaviors, optimizing performance of IDS becomes an important open problem that is receiving more and more attention from the research community. The uncertainty to explore if certain algorithms perform better for certain attack classes constitutes the motivation for the reported herein. In this paper, they evaluate performance of a comprehensive set of classifier algorithms using KDD99 dataset. Based on evaluation results, best algorithms for each attack category is chosen and two classifier algorithm selection models are proposed. The simulation result comparison indicates that noticeable performance improvement and real-time intrusion detection can be achieved as they apply the proposed models to detect different kinds of network attacks.

Yeung and Chow [6] proposed a novelty detection approach using no-parametic density estimation based on Parzen-window estimators with Gaussian kernels to build an intrusion detection system using normal data only. This novelty detection approach was employed to detect attack categories in the KDD dataset. The technique has surprisingly good reported results: 96.71% of DoS, 99.17% of Probe, 93.57% of U2R and 31.17% of R2L respectively. However, due to the fact that no FP was reported by the authors and a nearly impossible detection rate of 93.57% of U2R category, they really have to question the authentic of the reported numbers.

From Krishna Kant et al[7] In these days an increasing number of public and commercial services are used through the Internet, so that security of information becomes more important issue in the society information Intrusion Detection System (IDS) used against attacks for protected to the Computer net-works. On another way, some data mining techniques also contribute to intrusion detection. Some data mining techniques used for intrusion detection can be classified into two classes: misuse intrusion detection and anomaly intrusion detection. Misuse always refers to known at-tacks and harmful activities that exploit the known sensitivity of the system. Anomaly generally means a generally activity that is able to indicate an intrusion. In this paper, comparison made between 23 related papers of using data mining techniques for intrusion detection. Our work provide an overview on data mining and soft computing techniques such as Artificial Neural Network (ANN), Support Vector Machine (SVM) and Multivari-ate Adaptive Regression Spline (MARS), etc. In this paper comparison shown between IDS data mining techniques and tuples used for intrusion detection. In those 23 related papers, 7 research papers use ANN and 4 ones use SVM, because of ANN and SVM are more reliable than other models and structures. In addition, 8 re-searches use the DARPA1998 tuples and 13 researches use the KDDCup1999, because the standard tuples are much more credible than others. There is no best intrusion detection model in present time. However, future research directions for intrusion detection should be explored in this paper.

From this paper [8], they present an overview of our research in real time data mining-based intrusion detection systems (IDSs). They focus on issues related to deploying a data mining-based IDS in a real time environment. They describe our approaches to address three types of issues: accuracy, efficiency, and usability. To improve accuracy, data mining programs are used to analyze audit data and extract features that can distinguish normal activities from intrusions; they use artificial anomalies along with normal and/or intrusion data to produce more effective misuse and anomaly detection models. To improve efficiency, the computational costs of features are analyzed and a multiple-model costbased approach is used to produce detection models with low cost and high accuracy. They also present a distributed architecture for evaluating cost-sensitive models in realtime. To improve usability, adaptive learning algorithms are used to facilitate model construction and incremental updates; unsupervised anomaly detection algorithms are used to reduce the reliance on labeled data. They also present an architecture consisting of sensors, detectors, a data warehouse, and model generation components. This architecture facilitates the sharing and storage of audit data and the distribution of new or updated models. This architecture also improves the efficiency and scalability of the IDS.

Anazida Zainal et al. [9] in paper has discussed the Efficiency is one of the major issues in intrusion detection. Inefficiency is often attributed to high overhead and this is caused by several reasons. The purpose of the paper is to address the issue of continuous detection by introducing traffic monitoring mechanism. In traffic monitoring, a new recognition paradigm is

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

proposed in which it minimizes unnecessary recognition. Therefore, the purpose of traffic monitoring is two-folds; to reduce amount of data to be recognized and to avoid unnecessary recognition. For this Adaptive Neural Fuzzy Inference System and Linear Genetic Programming to form ensemble classifiers that shows a small improvement using the ensemble approach for DoS and R2L classes (attacks).

S.A.Joshi et al. [10] in paper has presented that with the tremendous growth in information technology, network security is one of the challenging issue and so as Intrusion Detection system (IDS). The traditional IDS are unable to manage various newly arising attacks. To overcome this type of problem Data Mining techniques, Feature Selection, Multiboosting were applied. With data mining, it is easy to identify valid, useful and understandable pattern in large volume of data. Features are selected using binary classifiers for more accuracy in each type of attack. Multiboosting is used to reduce both the variance and bias. Thus the efficiency and accuracy of Intrusion Detection system are increased and security of network so is also enhanced.

III. INTRUSION DETECTION SYSTEM

An IDS is a combination of software and hardware which are used for detecting intrusion. It gathers and analyzes the network traffic & detects the malicious patterns and finally alert to the proper authority. The main function of IDS includes:

- Monitoring and analyzing the information gathered from both user and system activities.
- Analyzing configurations of system and evaluating the file integrity and system integrity.
- For static records, it finds out the abnormal pattern.
- To recognize abnormal pattern, it use static records and alert to system administrator.

According to techniques used for intrusion detection based on whether attack's patterns are known or unknown, IDS classified into two categories

- Misuse detection
- Anomaly detection

Misuse detection: It 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.

Anomaly detection*:* It is different from Misuse detection. Here baseline of normal data in network data in network 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.

IV. IDS USING DATA MINING TECHNIQUE

a. IDS USING CLASSIFICATION ANALYSIS*:* The goal of classification is to assign objects (intrusions) to classes based on the values of the object's features. Classification algorithms can be used for both misuse and anomaly detections [11]. In misuse detection, network traffic data are collected and labelled as "normal" or "intrusion". This labelled dataset is used as a training data to learn classifiers of different types. The SVM is one of the most prominent classification algorithms in the data mining area, but its drawback is its extensive training time. The Support Vector Machine is one of the most successful classification algorithms in the data mining area.SVM uses a high dimension space to find a hyper-plane to perform binary classification.SVM approach is a classification technique based on Statistical Learning Theory (SLT).It is based on the idea of hyper plane classifier [12]. The goal of SVM is to find a linear optimal hyper plane. Here present the SVM model for classification. While intrusion behaviors happen, SVM will detect the intrusion. A classification task involves training set and testing set which consist of instances. Each instance in the training set contains one "target value" (class labels: Normal or Attack) and several "attributes" (features).The goal of SVM is to produce a model which predicts target value of data instance in the testing set which is given only attributes.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

b. IDS USING CLUSTERING ANALYSIS*:* Clustering assign objects (intrusions) to groups (clusters) on the basis of distance measurements made on the objects. As opposed to classification, clustering is an unsupervised learning process since no information is available on the labels of the training data. In anomaly detection, clustering and outlier analysis can be used to drive the ID model. Distance or similarity measure plays an important role in grouping observations in homogeneous clusters. It is important to formulate a metric to determine whether an event is deemed normal or anomalous using measures such as Jaccard similarity measure, Cosine similarity measure, Euclidian distance measure and longest common subsequence (LCS) measure [13]. Jaccard similarity coefficient is a statistical measure of similarity between sample sets and can be defined as the degree of commonality between two sets. Cosine similarity is a common vector based similarity measure and mostly used in text databases and it calculates the angle of difference in direction of two vectors, irrespective of their lengths. Euclidean distance is a widely used distance measure for vector spaces, for two vectors X and Y in a dimensional Euclidean space; Euclidean distance can be defined as the square root of the sum of differences of the corresponding dimensions of the vectors [14].

c. IDS USING ANN

Artificial Neural Network (ANN) is relatively crude electronic models based on the neural structure of the brain. The brain basically learns from his experience. This is natural proof that some problems that are beyond the scope and range of current computers are indeed solvable by small energy efficient packages. The brain modeling of a technical way to develop machine solutions is the new arrival approach to computing also provides a more graceful degradation during system overload than its more habitual counter-parts [15]. A neural network is an interrelated group of artificial neurons that uses a mathematical model or computational model for information processing based on a connection approach to computation. A neural network could not contains do-main knowledge in the beginning, but it can be supervised to make decisions by mapping example pairs of input data into example output vectors, and estimating its weights so that it maps each input example vector into the corresponding out-put example vector approx [16].

TECHNIQUE	ADVANTAGE	DISADVANTAGE
Ids using classification analysis	High dimension space to find a hyper-plane	Slow in classifying and testing tuples
Ids using clustering analysis	Grouping observations in homogeneous clusters	Require lot of training data
Ids using ANN	Interrelated group of artificial neurons, highest detection rate	Cannot detect novel attacks an d variation of known attacks

Table 1 : A COMPARATIVE ANALYSIS OF IDS

V. CONCLUSION

The security of computer networks plays an important role in modern computer system. Detection of intrusion attacks is the most important issue in computer network security. IDS can be divided into two classifies according to the detection approaches: anomaly detection and misuse detection. There are several different methods to anomaly detection and misuse detection and misconfiguration. Approaches to anomaly detection have neural network, Statistics, Predictive pattern generation, and sequence matching and supervising. In misuse detection, there are state transition analysis, pattern matching, model-based, keystroke monitoring and Expert system. This survey presents a various techniques of the data mining approach to solve the intrusion detection problems.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

REFERENCES

[1] Ahmed Youssef and Ahmed Emam, "Network Intrusion Detection Using Data Mining And Network Behaviour Analysis".
[2] Jian Pei, Shambhu J. Upadhyaya Faisal Farooq, Venugopal Govindaraju, "Data Mining for Intrusion Detection– Techniques, Applications and Systems"
[3] Yogita B. Bhavsar1, Kalyani C.Waghmare Intrusion Detection System Using Data Mining Technique: Support Vector Machine
[4] Sahilpreet Singh,Meenakshi Bansal, "A Survey on Intrusion Detection System in Data Mining"
[5] Huy Anh Nguyen and Deokjai Choi, "Application of Data Mining to Network Intrusion Detection: Classifier Selection Model"
[6] Yeung, D.Y., Chow, C.: Prazen-window Network Intrusion Detectors. In: 16th International Conference on Pattern Recognition, Quebec, Canada, pp. 11–15 (August 2002)
[7] Krishna Kant Tiwari, Susheel Tiwari, Sriram Yadav, "Intrusion Detection Using Data Mining Techniques"
[8] Wenke Lee, Salvatore J. Stolfo , Philip K. Chan, "Real Time Data Mining-based Intrusion Detection".
[9] Anazida Zainal, Mohd Aizaini Maarof and Siti Mariyam Shamsuddin "Data Reduction and Ensemble Classifiers in Intrusion Detection" in *2008 IEEE.*
[10] S.A.Joshi, Varsha S.Pimprale "Network Intrusion Detection System (NIDS) based on Data Mining" *International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 1, January 2013*
[11] Abhaya, Kaushal Kumar, Ranjeeta Jha, Sumaiya Afroz, "Data Mining Techniques for Intrusion Detection: A Review".
[12] Chunhua Gu and Xueqin Zhang," A Rough Set and SVM Based Intrusion Detection Classifier", Second International Workshop on Computer Science and Engineering, 2009.
[13] Steven Noel et al, " Modern intrusion detection, data mining, and degrees of attack guilt"
[14] L. Pornoy. Intrusion detection with unlabeled data using clustering. In *Undergraduate Thesis*, Columbia University, Department of Computer Science, 2000.
[15] Joo, D., Hong, T., and Han, I. "The neural network models for IDS based on the asymmetric costs of false negative errors and false positive errors." Expert Sys-tems with Applications 25, 69–75 2000.
[16] Lippmann, R.P., and Cunningham, R.K. "Improving Intrusion Detection Performance Using Keyword Selection and Neural Network."