Security Algorithms & its Comparative Study

Anushka Vekaria, Nivisha Tated, Shailesh Tanwar, Aashya Khanduja, Prof. Avanish Tiwari
Department of Computer Engineering, Mukesh Patel School of Technology Management and Engineering, NMIMS

Abstract— There has been a substantial increase in the number of attacks like denial of service, buffer overflows, sniffer attacks and application-layer attacks in the last few years. Recent analysis on security incidents has demonstrated that no manual response to such attacks is feasible. Network security attacks aren't theoretical concept that can be ignored and dealt with later. Firewalls do provide some protection but they do not provide full protection and need to be used along with Intrusion detection system (IDS). IDS allow network administrator to detect policy violations hence have become an efficient defense tool against network attacks. Nevertheless, traditional IDS are vulnerable to original and new malicious attacks and it is very inefficient to analyze from large volume of data like logs. Also, there are high false alarm rate, high false positives and negatives for the common data. Data mining is a popular and important method to mine useful information from large data volumes which are noisy, fuzzy, and random. Therefore, we need to integrate the data mining techniques into the Intrusion detection systems. IDS can be developed using various individual algorithms like neural networks, classification, clustering etc. which yields less false alarm rate and good detection rate. Compared to the single algorithm, cascading multiple algorithms gives much better performance and reduces false alarm rates. We are going to use two hybrid algorithms- C4.5 Decision Tree and Support Vector Machine (SVM) to develop IDS to achieve high accuracy and diminish any false alarm rate in this paper.

Index Terms— Anomaly Detection, C4.5, Data mining, Decision Tree, Intrusion Detection, Intrusion Detection Techniques, Security algorithms, Support Vector Machine.

---

1 Introduction

During the past decade, security has become a critical issue for computer systems due to fast development of computers and World Wide Web [1]. An intrusion is as an action or attack in a network or system that hampers or destroys security measures like integrity, confidentiality and authentication [1]. Intrusion Detection Systems (IDS) is designed to detect intrusion or attacks against a computer system due to insecure networks. It detects attacks which violates security policies of a system. It monitors the system and analyzes the gathered intrusion related data to draw a conclusion whether the system is intrusive or not by tracking user activity, system logs, etc. [1]. After detecting any possible intrusion, IDS raises alarm to network administrator and starts protection process.

IDS is a combination of both software and hardware components and is capable of monitoring and analyzing different activities in a network for signs of security threats [3]. There are two major approaches to intrusion detection:

1. Anomaly detection
They consist of building models from observed data that can be used to detect any deviation from the normal model.

2. Misuse detection
It uses patterns of previous well-known intrusions to match and identify intrusions in current [9] unlabeled data sets. Example- commercial and open source intrusion detection systems [3].

1.1 Classification of Intrusion Detection

There are two basic types of intrusion detection each having a distinct approach to secure and monitor data: Host-based and Network-based.

1. Host-based intrusion detection systems (HIDS)
These operate on a single workstation or machine. HIDS uses resources of host machine to monitor traffic on host machine to detect attacks [2]. HIDS also audits data traced by the host operating system.

2. Network-based intrusion detection systems (NIDS)
These operate as stand-alone devices on a network. They monitor network traffic to detect [8] attacks like denial of service attacks. They also scan ports and try to crack into computers by monitoring the system network traffic. NIDS uses this raw content of network packets to analyze the network and identify any suspicious activity [2].

1.2 Networking Attacks

Threats can be categorized into external or internal threats. External threats are the ones originating outside a company or an institution while an internal threat is one originating inside the organization. Security threats and attacks are that involves any layer of OSI stack, from physical to application layer.

1. Denial-of-Service (DOS) attacks
It is an attempt to forbid the authorized users from utilizing the requested service or resources. A more advanced Distributed DOS occurs when in a distributed environment the attacker sends or rather floods the server or a target system with numerous connection requests knocking down the target system leaving them no other option to restart their system [2]. Some known attacks are:

1) SYN Attack
2) Ping of Death

2. Eavesdropping Attacks

It is form of external attack. There is an unauthorized interference of network communication and disclosure of any information exchanged. This can be performed in different layers – for example, by detecting the exchanged packets in network layers or in physical layer by physically wiretapping the access medium [2].

3. Spoofing attack

The attacker impersonates a legitimate user. Example- IP spoofing where the system is convinced that it is communicating with a trusted user and hence provides access to the intruder. The attacker alerts the transport layer buy sending a packet with an IP address of a known host [2].

4. Intrusion attacks or User to Root Attack (U2R)

An unauthorized user tries to gain access to system or root through the network. Buffer overflow attack or User to root attack is an intrusion attack which occurs when a web service receives more data than it has been programmed to handle, i.e. increase in network traffic which leads to data loss [2].

5. Log-on Abuse attacks

A successful abuse attack during log in bypasses the authentication, access control mechanisms and grant a user with more privileges that authorized [2].

6. Application-Level Attacks

The intrude exploits the weakness in the application layer – for example, security weakness in the web server, faulty controls in the filtering of an input on the server side, malicious software attack (viruses, Trojans, etc.), web server attacks, and SQL injection [2].

2 Intrusion Detection Techniques

2.1 Genetic Approach

Genetic Algorithm (GA) technique uses evolution theory-information evolution in order to filter the data from network traffic to reduce complexity. It is a programming technique that mimics biological evolution as a problem-solving strategy. It is based on Darwinian principle of evolution and survival of fittest. It optimizes a population of candidate solutions towards a predefined fitness [6]. The process starts with population of generated chromosomes which represents all the candidate solutions possible for a problem. From each solution or chromosome, different positions referred as genes are encoded as bits, characters or numbers. An evaluation function called Fitness Function is then used to calculate the goodness of each chromosome according to the desired solution [6]. During the evaluation process, simulation of natural reproduction and mutation of species is done using Crossover. The selection of chromosomes is biased towards the fittest chromosomes.

Three factors having vital impact on the effectiveness of the algorithm and applications are used during [6] Genetic Algorithm. They are:

1) the Fitness Function;
2) the representation of individuals
3) the Genetic Algorithm parameters

The determination of these factors often depends on applications and/or implementation.

2.2 Fuzzy Logic

It is an approach based on degrees of truth, as the Boolean works on the concept of '0' or '1', fuzzy logic is in between where the answer can lie between '0' and '1'. To apply this in intrusion detection, the frequently occurring items from the dataset are chosen [6]. These attributes are put into min max deviation to make rules for detection. In the end we get a set of definite and a set of indefinite rules, reducing the total number of rules and decreasing the size of indefinite [6] rules to make it more efficient. Focusing on the definite rules, and by using 'if-then' statement we can use these rules for intrusion detection.

2.3 Statistical Approach

IDS has large amount of raw data that is analyzed from the network traffic to determine whether the traffic is harmful or not. In statistical based analysis, traffic across each active port in every host in a network is analyzed [12]. Each of these ports are treated as a number of sensors in a host running certain applications. The IDS instead of capturing physical parameters mines the whole network traffic data and associates them to the appropriate ports depending on incoming or outgoing traffic at particular time scale [11]. An average and variance is calculated for each port for the traffic at each port in that time scale. The common parameter used in analyzing network traffic data is the number of incoming and outgoing packets from a particular port. The interactions among various ports to determine the deviation of current status from the known patterns is also considered in current approach. Traffic monitoring captures packets at particular time frame, namely every millisecond [12]. The whole data is analyzed carefully to detect any inappropriate access at nearly real time manner. IDS also gets detailed information from the captured incoming data packets like packet size, the origin of IP address, the attacked port number and also its packet type. IDS looks at the content of the packet to determine whether it contains any malicious code. Else it uses deviations from normal behaviors and profiles to detect a threat. Normal behavior is the expected behavior derived from previous regular activities, network connections, hosts or users over certain time period [11]. When a computer network is operated in a safety critical environment, the system should be able to generate detection rate with high accuracy over the correctness of data.

Two methods for traffic monitoring using statistical approach are categorized as-

1. Signature-based IDS
Signature-based IDS has a lower false alarm than the anomaly-based IDS and uses set of rules known as signatures to identify or detect any intrusion packet. However, this method is inefficient [5].

2. Anomaly-based IDS
By identifying deviation from the normal behaviors, this method is able to detect a novel intrusion very efficiently. However, it produces higher rate of false alarm then signature-based IDS [5].

3 Shortcomings Of IDS

Most IDS detects such malicious activity either at the transaction level or at the OS level, or the both of the transaction and OS level. Such IDS has many limitations:

1. Data overload. To detect various attacks, there is a need to collect a large amount of information like system logs, transaction log, etc. to analyze. It is difficult for traditional IDS to comprehend and analyze the collected data due to a large amount of data source of intrusion detection [1].

2. False positives. Another drawback of traditional IDS is the amount of false positives, which occur when normal attack is mistakenly classified [1].

3. False negatives. This is the case where IDS system does not raise an alarm when there is an actual attack to the systems [1].

Obviously, with the increasing of requirement for network security, traditional intrusion detection systems cannot meet the existing requirement due to the fact that the IDS are established inefficient and the cost of research is so much because the IDS can't well mine the user behavior so that do some protecting against various attacks. Recently, the IDS combining with data mining have received much attention recently [1]. In order to overcome the existing current shortcoming of IDS that is inefficient to analyze from a large amount of data, we, in this paper, employ data mining techniques to the intrusion detection system so that achieve a more precise and efficient IDS.

4 Data Mining Techniques And IDS

Data mining is the process to extract useful information from large sets of data that are noisy, fuzzy and dynamic [1]. It analyzes the data sets to discover any unknown relation and sums up the result of analysis to help the owner of data to understand the important patterns in data. Data mining framework detects patterns in the data sets in large amounts automatically and uses these patterns to find any malicious activity [8].

In IDS, information comes from various complex sources like host data, network log data, alarm messages etc., the complexity of the operating system also increases. Also due to huge network traffic, data analysis becomes very hard. The data mining technology is capable of extracting data from large databases obtained from various sources hence it is important to use data mining techniques in [8] intrusion detection. By using data mining techniques, IDS can verify the data to obtain a model and obtain a comparison between the normal behavior and abnormal pattern.

Current problem in IDS is to determine how effectively it separates the attack patterns from normal data patterns over a large network data and how effectively it generates automatic intrusion rules after collected raw network data.

To achieve this, various data mining techniques are used. They are-

1. Classification
It is a form of data analysis which takes each instance of a dataset and assigns it to a particular class [9].

2. Clustering
The network data available is too large, time- consuming and expensive for human labelling. Clustering is the process of labelling such data and assigning it into clusters or groups. It divides data into groups of similar objects [9]. Members of same cluster are quite similar and members of different clusters are different from each other.

3. Association Rule
Here each attribute or value pair is considered as an item. Collection of items is called an item set in a single network request. Association rule algorithm searches for an item set that appears frequently in a network from large number of dataset [9]. The main goal is to derive multi-feature correlations from a database table. Association rule mining finds association rules and/or correlations among large dataset items.

Intrusion detection using Artificial Neural Network (ANN) and fuzzy clustering shows that ANN gives improved performance compared with other traditional models [10]. A new approach for IDS which is based on hierarchical clustering and Support Vector Machine (SVM) provides high qualified training instance to SVM and reduces the training time which increases overall performance of the SVM [10]. Hybrid approaches improves the detection rate and false positives as compared to single approaches as it combines the advantages of two algorithms.

IDS using data mining efficiently identifies the data that user is interested in predicts the results of useful patterns that can be helpful in the future. It captures all the incoming packets transmitted over the network [13]. Data is collected and send for pre-processing to remove the noise and other irrelevant missing attributes are replaced. Pre-processed data is then analyzed and classified into groups depending on their severity measures. If the data is normal based on previous behavior, then it does not need any [10] changes or else it is reported and alarm is raised. The administrator is informed to handle the situation. The above process continues as soon as the transmission of data starts over the network.

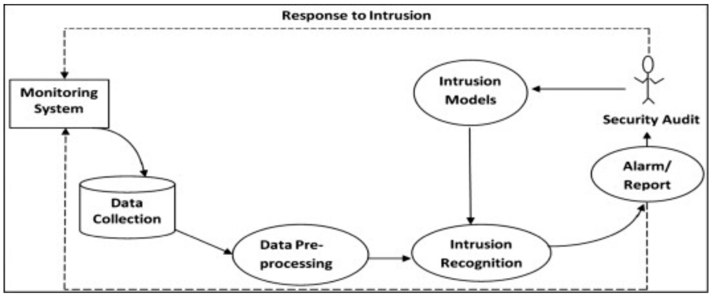

Fig. 1 Architecture of IDS.

Datasets can have mixed, numeric or no attributes and classes. All algorithms cannot perform well for attributes, classes and datasets of different size and types. So, to design a generic tool for classification, behavior of algorithms for various datasets must be studied. Classification algorithms based on data mining techniques like Decision trees and Naïve Bayes are analyzed and applied to KDDCup'99 datasets and then appraised for accuracy by using 10-fold cross validation strategy [4].

4.1 Decision trees

Decision trees integrates the experience gained by people working in pattern recognition and machine learning to form a program called ID3 [3]. ID3 evolved to a new system, named C4.5 and further to J48.

4.1.1 ID3 (Iterative Dichotomiser)

It is an attribute based machine-learning algorithm which adopts greedy approach. ID3 constructs a decision tree based on a training set of data and uses entropy measure to build the leaves of the decision tree [3]. The attribute with highest information gain on the training set is determined first. The attribute is then used as the root of the tree and branches are created using each value that attribute can take [3]. The process is repeated for each branch with the subset of the training set that is classified by this branch.

4.1.2 C4.5

This algorithm is an extension of ID3 algorithm. It builds decision tree from a set of training data in the same as ID3, using the concept of information entropy.

Pseudocode:

```
1. Check base cases
2. For each attribute a,
   Find the normalized information gain ratio from splitting on a
3. Let a_best be the attribute with the highest normalized information gain
4. Create a decision node that splits on a_best
5. Repeat on the sublists obtained by splitting on a_best
   and add those nodes as children of node [5]
```

C4.5 handles both continuous and discrete attributes. To handle continuous attributes, it creates a threshold and splits the list into those whose attribute [9] value is above threshold value and those that are less than or equal to it. It also handles attributes with differing costs and prunes trees after creation.

4.1.3 J48 (enhanced version of C4.5)

It is enhanced version of C4.5 with additional features to overcome limitations of C4.5 like amount of CPU time and memory required. C4.5 uses the successful method to find high accuracy which is done by pruning the rules issued from the tree constructed during the learning phase.

Given a set S of cases, J48 first grows an initial tree using the divide-and-conquer algorithm as follows:

1. If S is a small case or all cases in S belong to the same class, the tree is leaf labelled with the most frequent class in S [3].

2. Else, choose such a test that gives two or more outcomes to a single attribute. Make this test as the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S1, S2, … according to the outcome for each case, and apply the same procedure repeatedly to each subset [3].

J48 uses two heuristic criteria to rank possible tests: information gain, default gain. After the building process, each attribute test along the path from the root to the leaf becomes a rule precondition(antecedent) and the classification at the leaf node becomes the rule post condition(consequence). This rule is pruned by removing any antecedent which does not affect the accuracy by it removal [3]. Since the rules have the "IF...THEN…" format, they can be used as a model for a rule based intrusion detection.

J48 rules generated are concise and intuitive. So, they can be checked and inspected for any further investigation by a security expert. J48 rules generates generalization accuracy.

After building processes similar to already known attacks new intrusions may appear after the building process whose forms are quite similar to known attacks that are considered a priori. Using the generalized accuracy of the rules, new attack variations could then be detected using different rules. Real time IDSs require short rules for efficiency. Post pruning the rules generates accurate conditions and avoids over fitting, thus improves the execution time for real time intrusion detection [3].

4.2 Naïve Bayes Algorithm

It is an extension of Bayes theorem in that it assumes independence of attributes. This assumption is incorrect considering text-extraction based classification from a document. Such problems are called problems of supervised classification [3].

This algorithm is easy to construct and does not need any complex iterative parameter estimation schemes. It can be immediately applied to large data sets and is robust and easy to interpret [3].

4.3 Support Vector Machine

SVM is used to solve Binary classification problems. It maps linear algorithm into non-linear space [7]. For this mapping purpose it uses kernel function. There are various kernel functions like polynomial, radial basis function.

This kernel functions can be used at the time of training of the classifier to selects support vector along the surface of this function [9]. These support vectors are used by SVM to classify data that outline the hyper plane in the feature space. The linear classifier has the form

F(x) = wt+b
Where, w is weight vector and b is bias

The distance between the data and hyper plane is shown in Fig 2. The idea is that SVM maps inputs vectors nonlinearly into the high [9] dimensional feature space and construct the optimum separating hyperplane.

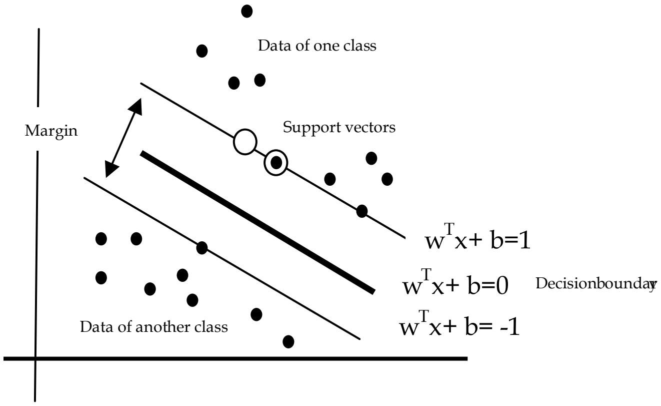

Fig. 2 Hyper Plane of Support Vector System

5 Comparative Study Of Data Mining Algorithms

Here, a comparison of performance of each data mining algorithm discussed above along with their significance is done. The experiments are carried out using WEKA (Waikato Environment for Knowledge Analysis); with full data set for training and 10-fold cross validation for testing purposes. The results presented are obtained on a Pentium4 CPU, 2.8GHz, and 512MB of RAM.

The experimental results using the C4.5 Decision trees along with the results obtained from Naïve Bayes algorithm are shown below. In Table 1, comparison between time taken to build the model and error rate for all the three data mining algorithms is done [3].

TABLE 1
Comparison of data mining algorithms

Experiments	Overall Error Rate	Time taken to build the model in Seconds
Best KDD	7.29%	-
Naïve Bayes	3.56%	0.11
J48	3.47%	0.88
ID3	3.47%	1.8

The kappa statistics for all values, as shown in Fig. 3 suggest that the ability of the various classification methods examined could be considered as good, i.e. the classifier stability is very strong. Root mean square error in the same way suggests that the error rate is very small, which can be considered as a measure of effectiveness of the model [3]. It can be observed from the figure, that the accuracy of overall classification for Naïve Bayes for all classes is better than the overall and ID3 algorithms.

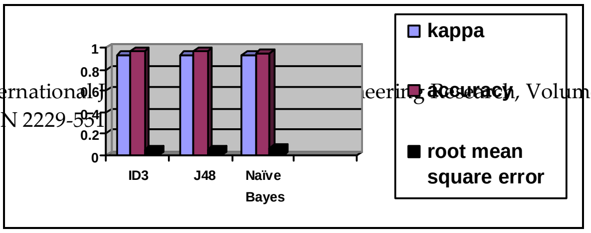

Fig. 3 Comparison of data mining algorithms

6 Proposed Model

IDS allow network administrator to detect policy violations hence have become an efficient defense tool against network attacks. Nevertheless, traditional IDS are vulnerable to original and new malicious attacks and it is very inefficient to analyze from large volume of data like logs. Also there are high false positives and false negatives. Data mining is an important way to mine useful information from large datasets which are noisy, fuzzy, and random.

There has been a substantial increase in the number of attacks like Denial of Service, Buffer overflows, Sniffer attacks and Application-Layer attacks in the last few years. Network security attacks aren't theoretical concept that can be ignored and dealt with later. Firewalls do provide some protection but they do not provide full protection and need to be used along with Intrusion detection system (IDS). IDS allow network administrator to detect policy violations hence have become an efficient defense tool against network attacks. Nevertheless, traditional IDS are vulnerable to original and new malicious attacks and it is very inefficient to analyze from large volume of data like logs. Data mining techniques are being applied in building intrusion detection systems to protect computing resources against unauthorised access.

Therefore, we need to integrate the data mining techniques into the Intrusion detection systems. IDS can be developed using various individual algorithms like neural networks, classification, clustering etc. which yields less false alarm rate and good detection rate. Compared to the single algorithm, cascading multiple algorithms gives much better performance and reduces false alarm rates. In this paper, we are doing to use two hybrid algorithms- C4.5 Decision Tree and Support Vector Machine (SVM) to develop IDS to achieve high accuracy and diminish any false alarm rate.

Advantages our proposed model-

1. It will build models which can be easily interpreted and will be easy to implement.
2. It can use both discrete and continuous values and will deals with noise.
3. It will also have high accuracy rate and will work well even if data is not linearly separable in the base feature space.
4. It will also give better speed as compared to both the algorithms individually and even high complexity and extensive memory for classification will be reduced.

Conclusion

In this paper, we have reviewed effectiveness of the classification algorithm Naïve Bayes with the decision tree algorithms namely, ID3, C4.5 and J48 along with techniques like Genetic Approach, Support Vector Machine, Naïve Bayes Algorithm, Fuzzy Logic and Statistical Approach.

It is observed that the Naïve Bayes model is quite appealing because they are simple, robust, elegant and very effective. On the other hand, decision trees have proven their efficiency in both generalization and detection of new attacks. The results show that there is no single best algorithm to outperform others in all situations. To get better results hybrid algorithms can be used like Intrusion Detection System by combining two data mining algorithms C4.5 Decision Tree and Support Vector Machine. Building an intrusion detection models which is efficient and has good accuracy and real-time performance are very essential. However, other kinds of preprocessing techniques and data mining approach like artificial intelligence neural network models may be tested for a better detection rate in the future research in IDS System.

References

[1] Wang Pu, Wang Jun-qing, "Intrusion Detection System with the Data Mining Technologies", Communication Software and Networks (ICCSN), IEEE 3rd International Conference, 2011

[2] Karthikeyan. K.R, A. Indra, "Intrusion Detection Tools and Techniques –A Survey", International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010

[3] Mrutyunjaya Panda, Manas Ranjan Patra," A Comparative Study of Data Mining Algorithms for Network Intrusion Detection", Emerging Trends in Engineering and Technology, 2008. ICETET '08. First International Conference on July 2008

[4] Santosh Kumar Sahu, Sauravranjan Sarangi, Sanjaya Kumar Jena, "A Detail Analysis on Intrusion Detection Datasets", Advance Computing Conference (IACC), 2014 IEEE International

[5] A.A. Waskita, H. Suhartanto, P.D. Persadha, L.T. Handoko, "A simple statistical analysis approach for Intrusion Detection System", Systems, Process & Control (ICSPC), IEEE Conference, 2013

[6] J. Jabez, Dr. G.S. Anadha Mala, "A Study on Genetic-Fuzzy Based Automatic Intrusion Detection on Network Datasets", Software Engineering and Mobile Application Modelling and Development (ICSEMA 2012), International Conference, 2012

[7] Aleksandar Lazarevic, Aysel Ozgur, Levent Ertoz, Jaideep Srivastava, Vipin Kumar, "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection" SIAM International Conference on Data Mining, 2003

[8] Pragya Diwan, Dr. R.C Jain, "A Combined Approach for Intrusion Detection System Based on the Data Mining Techniques", International Journal of Computational Engineering Research (IJCER), Vol, 04, Issue, 6, June 2014

[9] Vaishali Kosamkar, Sangita S Chaudhari," Improved Intrusion Detection System using C4.5 Decision Tree and Support Vector Machine

“, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014

[10] Peyman Kabiri, Ali A. Ghorbani, “Research on Intrusion Detection and Response: A Survey”, International Journal of Network Security, Vol.1, No.2, PP.84–102, Sep. 2005

[11] Hari Om, Tanmoy Hazra, “Statistical Techniques in Anomaly Intrusion Detection System”, International Journal of Advances in Engineering & Technology, Nov. 2012

[12] Nabil R. Adam, John C. Wortmann, “Security-Control Methods for Statistical Databases: A Comparative Study”, ACM Computing Surveys, Vol. 21, No. 4, December 1989