Data Mining Techniques Using Various Cyber Security Intrusion Detection

Avula Chitty
Assistant Professor, Dept of CSE
Sri Indu College Of Engineering And Technology, Hyderabad, Telangana, India

Abstract- *Intrusion Detection System (IDS) has started becoming part and parcel of every system considering the growing security breaches in the cyber world. Security threat means violating confidentiality and integrity of system thereby causing potential financial loss to the organizations. Cyber criminals bypass the authentication mechanism and intrude into the system to steal personal and professional information of their victims from database. One of the most common techniques used for intrusion is the SQL Injection attack. This attack is used to either get an unauthorized access to database or retrieve information directly from the database. This attack was ranked third among the top ten database security threats[2].Thus it becomes a serious threat to any database driven website and hence needs to be detected efficiently. IDS combined with data mining technique are one of the way for detecting the intrusions in the system. This paper reviews various types of SQL injection attacks and data mining techniques used for detection.*

Keywords- Database, Data Mining, Intrusion Detection System, Security, SQL Injection attack.

I. INTRODUCTION

With the technological advancement and its ease of availability, a lot of people have started adopting it. Almost every transaction today is done online. This scenario causes cyber criminals to achieve their malicious motive. They compromise the security mechanisms of the system and gain unauthorized access thereby stealing all vital data. One of the most common and oldest techniques for gaining unauthorized access is the SQL Injection attack. SQL injection attacks are initiated by passing some malicious code fragment in web application. The Web application then combines these unsafe SQL fragments with the proper SQL queries generated by the application, thus creating valid SQL requests. These new, malicious requests cause the database to perform the task the attacker intends [1].

So, Implementing an Intrusion detection system helps the database administrator to keep an eye on such kind of intrusions. Whenever there is an intrusion, IDS will detect it and notify it to the database administrator. Administrator can then take the necessary actions on the detected intrusion. The detection mechanisms in IDS can be implemented using data mining techniques. The various algorithms in data mining can be used for detection of intrusions.

II. RELATED WORK

INTRUSION DETECTION SYSTEM

Intrusion Detection System (IDS) is a software application that monitors the system for malicious activities and suspicious transactions. Any such activity that takes place is reported to the database administrator. An IDS works by monitoring system activity through examining vulnerabilities in the system, the integrity of files and conducting an analysis of patterns. It monitors the Internet and searches for any of the new threats which could further result in an attack.

Functions of IDS are as follows:

1. Monitoring and analyzing both user and system activities.
2. Detect abnormal activities.
3. Ability to recognize patterns of attacks.
4. Analyzing system configurations and vulnerabilities.
5. Checking for security policy violations.

Types of IDS:

IDS can be classified in two ways:

i) Based on where the detection takes place.
ii) Based on what detection method is used.

i) Based on where the detection takes place, the intrusion detection systems are classified as follows [4]:

Network Intrusion Detection System:

Network Intrusion Detection System (NIDS) exists at certain points in the network to monitor traffic to and from all the devices in the network. It analyses the network traffic and matches it to the library of known attacks. If an attack is detected, or any abnormal activity is sensed, an alert is sent to the administrator. There are two types of NIDS: On-line and Off-line NIDS. On-line NIDS is used with network in real time whereas off-line NIDS works with stored data.

Host Intrusion Detection System:

Host Intrusion Detection System (HIDS) exists on the individual devices in the network. It tracks the incoming and outgoing packets from the device on which it is installed and notifies the administrator if any suspicious activity is found. A host-based system also has the ability to monitor important system files and any attempt to overwrite these files. HIDS also performs functions like log analysis, integrity checking, event correlation, policy enforcement, and rootkit detection.

ii) Based on detection method, the intrusion detection systems are classified as follows [3]:

Signature based Intrusion Detection System:

It monitors and analyzes network packets and compares them against the signature of the known threats. It is very effective in detecting known attacks or threats that are predefined in the database of IDS. But this system has a disadvantage that a new kind of attack cannot be detected as its signature is not present. In signature based IDS, database of signatures has to be manually updated whenever a new kind of attack is discovered.

Advantages:

It is often considered to be much more accurate at identifying an intrusion attempt of known attack.

Administrators spend less time dealing with false positives and hence time is saved.

Disadvantages:

Signature based systems can only detect an intrusion attempt if it matches a pattern that is in the database, therefore causing databases to constantly be updated Whenever a new virus or attack is identified it can take vendors some time to update their signature databases.

Hosts that are subjected to huge amount of traffic, the IDS can have a difficult time inspecting every single packet that it comes in contact, which then causes some packets to be dropped leaving the potential for hazardous packets getting by without detection.

Anomaly based Intrusion Detection System:

It monitors and analyzes network packets and compares them against some standard baseline. This baseline refers to the normal data transaction. If there is a slight amount of deviation from this normal behavior defined by the baseline then that packet is considered to be malicious. It basically classifies the input request either as normal or anomalous. This system is very much useful in detecting unknown attacks and works at both host level and network level.

Advantages:

New threats can be easily detected even when database of attacks is not up to date.It requires little maintenance after system is installed It continues to learn about network activity and continues to build its profiles.The longer the system is in use the more accurate it can become at identifying threats.

Disadvantages:

If malicious activity appears like normal traffic to the system it will never send an alarm to administrator.False positives can become cumbersome with an anomaly based setup.

III. SQL INJECTION ATTACK

SQL Injection is a kind of attack in which the attacker executes malicious SQL statements that controls the relational database management system. The attacker uses this attack to bypass web application’s authorization and authentication mechanism and retrieve the contents of database. The various types of SQL Injection attacks are as follows [6] [10]:

Tautologies:

The basic aim of tautology based attack is to inject code in one or more conditional statements so that the query is always evaluated to true. The most common use of this attack is to bypass the authentication pages and extract data. In this attack, the attacker targets the WHERE clause where the input is accepted from user.

Logically Incorrect Queries:

In this type of attack, attackers try to gather important information about the type of database and its structure. This technique is normally used in the initial phase before performing the main attack to collect vital knowledge about database. The error pages that are returned after making this attack provide a lot of information. Even if the application sanitizes error messages, the fact that an error is returned or not returned can reveal vulnerable or injectable parameters. This attack helps to retrieve the number of columns, their names and sometimes also the table names.

Union Query:

In this type of attack the malicious query is joined with an authentic query using the keyword UNION to perform union between two or more queries and gets information related to other tables from the database. The attacker can use this attack to insert any malicious query to retrieve information from a table different from the one that is present in the original statement. The database returns a result that is the combination of the results from the original first query and the results from the injected, malicious second query.

Stored Procedure:

Usually the user of the database creates certain procedures and stores it in database for future use. Doing this also saves a lot amount of time. These procedures are known as the stored procedures. This attack attempts to execute database stored procedures. Before doing this attack, the attacker initially determines the database type and then uses that knowledge to determine what stored procedures might exist. Stored procedures are susceptible to privilege escalation, buffer overflows, and even provide access to the operating system.

Blind Injection:

By using this type of attack, the attacker collects information from the replies of the page after querying the server with true/false questions.

Alternate Encodings:

In this type of attack, the injection query is modified by alternate encoding, changing characters to some other characters in the queries. By this way, the attacker evades filters for “wrong characters”. All different kinds of SQL injection attack can be made hidden using this method.

The hexadecimal encoded character are taken as input that is used as char function that returns actual character. This encoded string, shutdowns the database when the command is executed.By making the combination of these attacks any hacker or attacker can get useful information about the database. Such as number of tables, columns, rows, etc. This information about database can be used for different attacks such as Cross site Scripting attacks, Denial of Services attacks, IP Spoofing, etc. to attack the web sites and applications.

IV. DATA MINING TECHNOLOGIES

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

K-means: The k-Means algorithm groups ‘n’ instances into k disjoint clusters, where k is a predefined parameter. Each instance is assigned to its 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 is applied on small dataset. When the data point increases then it takes more execution time [5] [8].

K-Nearest Neighbor (KNN): It is one of the simplest classification techniques. 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 is equal to 1, then the data point 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 [9].

K-Medoids: K-Medoids 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. It minimizes the distance between centroid and data points which means minimizing the squared error. K-Medoids algorithm performs better than K-means algorithm when the number of data points increases. It is robust in presence of noise and outlier because medoid is less influenced by outliers, but processing is more expensive [8] [9].

EM-Clustering: It is the Expectation- Maximization algorithm. 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 represents the probability of membership. In other words there are no strict boundaries in between the clusters [8].

Classification Tree: In machine learning, Classification Tree is also known as Decision tree or predictive model. It is a 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.

C4.5: C4.5 constructs decision trees from a set of available training data using the concept of information entropy. 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. The splitting criterion is the normalized information gain. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists [7].

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. As a result, 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 [7].

Support Vector Machine: Support Vector Machine (SVM) is a supervised machine learning algorithm. It 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. Main significance of the Support Vector Machines is that it is less susceptible for over fitting of the feature input from the input items, this is because SVM is independent of feature space. Here classification accuracy with SVM is quite impressive or high. SVM is fast accurate while training as well as during testing [9].

Naïve Bayes Classifier: Naïve Bayes classifier is a probabilistic classifier. To get the probability, it analyses the relation between the dependent and the independent variables. Bayes Theorem is as follows:

$$P\ (H/X) = \quad (\ /\ ).\ (\ )\ /\ (\ )$$

Where, X is the data record and H is hypothesis which represents data X and belongs to class C. P (H) is the prior probability, P (H/X) is the posterior probability of H conditioned on X and P(X/H) is the posterior probability of X conditioned on H [9].

Neural Networks: An artificial neural network (ANN), also known as neural network (NN), is a mathematical model or computational model which is based on working of human brain. It consists of multiple nodes which resembles biological neurons of human brain. There are two types of ANN: Feed forward and Feedback. In feedforward ANN, the information flow is unidirectional whereas in feedback ANN allows loops. [8]

Genetic Algorithms: Genetic algorithms provide a comprehensive search methodology for machine learning and optimization. Algorithm is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are selected and they are used to form a new population. This is done with the hope, that the new population generated will be better than the old one. Solutions that are selected which will be forming a new solution further are selected based on their fitness. Means, the more appropriate they are the more chances they will have to reproduce. [8]

V. PROPOSAL SYSTEM

This proposed system uses Naïve Bayes classifier algorithm for efficiently detecting all kinds of SQL Injection attacks. Whenever any malicious user tries to bypass the authentication mechanism by using the SQL injection attack at the login screen, the Naïve Bayes classifier algorithm is invoked. This algorithm takes XML file which contains 1024 probable patterns of 10 features that are used in SQL injection. This forms the training data set for the algorithm.

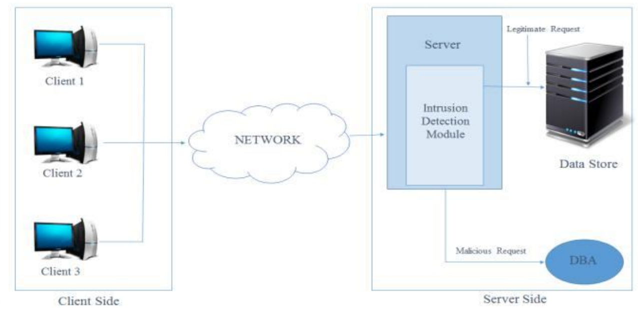

Fig -1: Architecture

The input given by user will be compared with that XML file and will be categorized either as attack or no attack based on probability calculation. If the attack is detected, an alert will be sent to the administrator along with the type of attack after comparing the signatures of attacks.

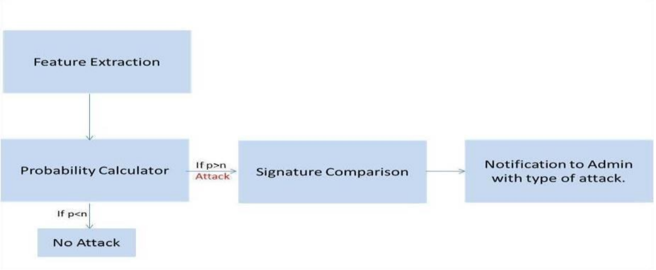

Fig -2: Intrusion Detection Module

Fig.1. shows the architectural diagram of the proposed system and Fig.2. shows the block diagram of the intrusion detection module.

VI. CONCLUSION

Intrusion detection becomes an important component for securing data as the threats in the cyber world continue to increase. This paper discusses the different types of SQL injection attacks and also data mining algorithms that are used in detecting the intrusions and proposes a system for detecting SQL injection attacks using Naïve Bayes Classifieralgorithm.

REFERENCES

[1] Li Qian, Zhenyuan Zhu, lun Hu, Shuying Liu, “Research of SQL Injection Attack and Prevention Technology”,International Conference on Estimation, Detection and Information Fusion , pp 303-306, 2015.

[2] IMPERVA.Top Ten Database Security Threats. 2015.

[3] Archana Thusoo, G.B Jethava, “A Survey : IntrusionDetection System for database using data mining techniques”, International Journal of Engineering Research and General Science ,Volume 3, Issue 2, pp 362-369, March-April, 2015.

[4] Jasmeen Kaur Chahal, Amanjot Kaur, “Use of Data Mining Techniques in Intrusion Detection – A Survey”, Imperial Journal of Interdisciplinary Research (IJIR), Vol-2, Issue-6, pp 452-456, 2016.

[5] Solane Duquea, Dr.Mohd. Nizam bin Omar, “Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)” , Conference Organized by Missouri University of Science and Technology, Procedia Computer Science 61,San Jose, CA, pp 46-51, 2015.

[6] Aniruddh R. Ladole, D. A. Phalke, “A Survey on SQL Injection Attack Countermeasures Techniques”, International Journal of Science and Research (IJSR), Volume 4 ,Issue 11, pp 1556-1561, November 2015 .

[7] Jaina Patel, Mr. Krunal Panchal, “Effective Intrusion Detection System using Data Mining Technique”, Journal of Emerging Technologies and Innovative Research (JETIR), Volume 2, Issue 6, pp 1869- 1878, June 2015.

[8] Shikha Agrawal, Jitendra Agrawal, “Survey on Anomaly Detection using Data Mining Techniques”, 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science 60, pp 708 – 713, 2015.

[9] 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, pp 6938- 6942, June 2014.

[10] Sankaran. S , Sitharthan. S , Ramkumar. M, “Review on SQL Injection Attacks: Detection Techniques and Protection Mechanisms”, International Journal of Computer Science and Information Technologies, Vol. 5 (3) , pp 4019-4022, 2014.