A Novel Approach To Detect Hacking Activity Using Data Mining Techniques

Gayathri N[1] and Kalaimathi B[2]

[1]*Assistant Professor, Department of CSE , MVJ College of Engineering*
Channasandra, Near ITPL, Bangalore-67, India

[2]*Assistant Professor, Department of CSE, MVJ College of Engineering*
Channasandra, Near ITPL, Bangalore-67, India

Abstract

Security in sensitive data and information becomes extremely important in today's world. Many organizations face huge loss because of the hackers who intrude in their systems and steal the important data. Though intrusion detection system is adopted by many organizations, still many issues are faced by them because of the hackers who even crack their servers. Intrusion detection system faces with the problem of false positives and false negatives many times. The motivation behind this paper is to combine the data mining techniques with the intrusion detection system so that the detection of the hackers becomes easy and effective. This paper presents an overview about the data mining techniques and their algorithms which are suitable for embedding with IDS and how they are applied to detect the hackers efficiently. The proposed approach proves that this technique reduces the chances of false negatives and false positives and hence provides an optimized way for identifying the hackers.

The first section in this paper gives an overview about intrusion detection system. It explains the possible hacking techniques and the way how intrusion detection system identifies the hackers. The second section in this paper gives an overview about the data mining techniques and the way how they are used in detection purposes. The third section in this paper presents the proposed approach of integrating the data mining technique to intrusion detection system and the way how this approach is going to be effective compared to the previous approaches.

IJERT

Keywords— Intrusion detection system, Data mining, classification, clustering, k-

means clustering neural networks, association rules, hackers.

1. INTRODUCTION.

Network attacks can be categorized as access attacks, modification attacks, denial of service attacks and repudiation attacks. Access attack is an attempt to gain information that an attacker is not authorized to see. It can occur wherever the information resides or may exist during transmission. It is an attack against the confidentiality of information. Modification attack is an attempt to modify information that an attacker is not authorized to modify. This is an attack against the integrity of information. Denials of service attacks are attacks that deny the use of resources to legitimate users of the system, information or capabilities. It might cause the information or application or systems or communications to become unavailable. Repudiation attack is an attempt to give false information or to deny that a real event or transaction should have occurred. It is an attack against the accountability of the information. All these attacks cause serious problems and heavy loss for the organization.

An organization should consist of policies to identify the attacks and implement technical tools for security purposes. Security reporting systems such as monitoring and scanning, authentication systems such as biometrics and firewalls, intrusion detection systems such as anti-virus software and automated log examination can be used by the organizations.

Intrusion detection system (IDS) is a reactive concept that tries to identify a hacker when a penetration is attempted. There are two types of intrusion detection systems: A host-based and a network-based. A host-based intrusion detection system (HIDS) resides on a particular host and looks for indication of attacks on that host. A network-based intrusion detection system (NIDS) resides on a separate system that watches the network traffic, looking for indications of attacks that traverse that portion of the network. IDS help to identify reconnaissance events, attacks, policy violations, and suspicious events. These can be achieved using stealthy scans, port scans, Trojan scans, vulnerability scans etc. Suspicious events can be investigated using the traffic to and from the source. Then it can be classified as a hacking activity or not. If it is classified as a hacking activity, the hacker can also be identified using intrusion detection system.

IJERT

Data mining is the technique of extracting knowledge from large sets of data. It includes steps like classification, association analysis, clustering, and anomaly detection. Classification is a process to classify the input data set into any one of the output categories. Association analysis is a process to associate the data sets that go together. Clustering is the process of forming clusters of data sets that have similarities in deriving the output. Anomaly detection is the process of detecting the outliers and identifying the data sets that have different behaviours to eliminate. Then finally the data is fed to the training model for classification or prediction purposes. This data mining is widely used in many applications and domains for prediction purposes.

The idea behind this paper is to use data mining steps to the intrusion detection system so that the classification of an activity as a hacking activity or the classification of a source as a hacker can be made very accurate. IDS though is good in detecting a suspicious event, it often suffers from the problem of false positives and false negatives. False positives mean that a suspicious event is classified as a hacking activity but in real it might be a normal activity. The same way false negative mean that a suspicious event may be classified as a normal activity but in reality it might be a hacking event. To avoid these kinds of errors in IDS, and to make the classification of a suspicious event more accurate, data mining is used along with IDS. This paper presents an approach of detecting the hacking activity and hacker effectively using data mining techniques in IDS.

2. Identifying hacking activity and hacker using Intrusion detection system.

An intrusion detection system (IDS) is a device or software application that monitors network or system activities for malicious activities or policy violations and produces reports to a Management Station. Network Intrusion Detection System is an independent platform that identifies intrusions by examining network traffic and monitors multiple hosts. Sensors capture all network traffic and analyze the content of individual packets for malicious traffic. Host based intrusion detection system consists of an agent on a host that identifies intrusions by analyzing system calls, application logs, file-system modifications etc.

An activity is considered as a suspicious activity if it differs from the pattern of normal activity. For example, missing log files in a system, empty contents in log file are all examples of suspicious activity. A suspicious activity can be classified as a normal activity or hacking activity by the intrusion detection system using the following steps.

2.1 Identify the systems.

All the systems involved in that suspicious activity are identified. Then the IP addresses are resolved to their host names using DNS entry. Some cases encounter failure in DNS lookups. If identifying the host name fails by all means for source or destination involved in that activity, it might possibly be an attack.

2.2 Log additional traffic between source and destination.

Signatures can be configured in IDS to identify the traffic between the source and destination. For example, WIZ command in network is a command used by administrators to get some confidential information. If our traffic contains only WIZ command in the mails from source to destination, possibly it might be an attack. If it contains other traffic also, then possibly it might not be an attack. The same way the IDS can be configured for

IJERT IJERT

several other signatures and attack can be identified.

2.3 Log all traffic from the source.

All traffic from source to destination is collected for identification. The IDS detector is configured to collect all the information from the suspicious source. After investigating the traffic, the information such as source system's name, type and frequency of traffic exchanged between source and destination, type and frequency of traffic exchanged between source and any other systems, if it is a web traffic or mail traffic are all used to identify the suspicious activity.

2.4 Log the contents of packets from the source.

The final step is to log the contents of the packets from the source. Logging the contents is used to gather a complete record of the session and what commands are actually being sent to the destination. After examining the data, the legitimate activity and suspicious activity can be easily detected.

3. Data mining techniques.

Data mining includes lot of techniques like classification, clustering, association analysis, decision trees, neural networks etc. This section gives an overview about the following major data mining techniques that are used in this paper.

3.1 Clustering.

It is a statistical technique to group large data into smaller subsets called clusters. A cluster is a logical group or collection of objects, attributes or properties of entities such that the elements within each collection are more alike than elements in different collections. Clustering algorithm is used to extract distinct clusters, if any, in the input data. There are many algorithms available for clustering. One such types of algorithms include hierarchical clustering algorithms (HCA). HCA partitions available data into sub-clusters iteratively using dendrogram. A dendrogram represents a nested grouping of clusters from coarser to finer details. There are two popular strategies for HCA – agglomerative (bottom-up) and divisive (top-down). The algorithm that we are going to use in this paper is divisive algorithm.

Divisive algorithm:
Initially we assume all data points belong to a single cluster. At each subsequent iteration, the data are appropriately partitioned into sub-clusters using a similarity metric. This process is continued until further splits become meaningless.

3.2 Association analysis.

Attributes that often go together are predicted by association rule mining (ARM) algorithms considering a user specified minimum support and confidence. An association in data mining indicates a logical dependency between various attributes of an entity or various properties of an event. Every association rule has two parts called the antecedent (left hand side) and the consequent (right hand side).

IJERT IJERT

Ex:- Bread=>Milk.

Several algorithms like AIS algorithm and its variants, divide-and conquer algorithm, apriori algorithm, SETM algorithms are all used for predicting association rules. In our paper, we are going to use the apriori algorithm.

Apriori algorithm:

The algorithm attempts to find subsets which are common to at least a minimum number C of the item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

3.3 Classification.

Classification process in data mining is used to predict categorical class labels. It classifies data based on the training set and the values in the classifying attribute. Classification can be supervised or unsupervised. There are many techniques for classification in data mining which includes decision trees, bayesian classifier, neural networks etc. In this paper, we use neural networks for classification.

Neural networks:

Neural networks are non-linear statistical data modeling tools. Back propogation algorithm is used for classification via neural network in this paper. The input nodes are labeled and the data are fed into layer in the correct order which get propagated through the intermediate layers. The output produced at the output layer by the forward pass is noted down and the error E is found at each of the nodes. In backward-pass, starting with output layer, the errors are propagated, and the synaptic strengths are modified to minimise expected error.

4. The proposed approach.

The following section discusses the proposed technique for identifying the hacker and hacking activity using data mining techniques. The first step includes clustering.

4.1 Clustering.

As discussed in the earlier section, there are many clustering algorithms, which clusters the data sets into groups. The algorithm that we are using is divisive algorithm that initially considers all the datasets as a group and then divides into clusters depending on some metric. Initially all the training set of data here is considered as a single cluster. Then the divisive algorithm is executed in 4 phases with a metric upon which they are clustered in each phase. Figure 1 depicts the approach.

Step 1: - Initially all the datasets are assumed as a single cluster.

Step 2: - A metric – 'Missing log files or Changed log files' is chosen in phase1 which divides the cluster into two groups. Cluster 1 form the group of data sets that has empty contents in their log files or some changes in the log files. Cluster 2 forms the

IJERT

group of data sets, which doesn't have any change in their log files.

Step 3: - A metric – 'Without host name or with host name' is chosen in phase 2 which divides the cluster into two groups based on that metric. Cluster 1A forms the group of data sets that do not have a corresponding host name in the DNS entry. Cluster 1B forms the group of data sets that have a host name in the DNS entry. The same way for the cluster 2 also it is classified.

Step 4:- A metric – 'More kind of WIZ packets' is chosen in phase 3 that divides the cluster into two groups based on the traffic from the source to the destination. If the traffic contains more number of WIZ packets from source to destination, it forms cluster 11A and those with less number of WIZ packets from source to destination will form cluster 11B. The same way clustering is also done for Cluster 1B, 2A, 2B.

Step 5:- A metric – 'Frequency and type of traffic' is chosen in phase4 that divides the cluster into two groups. If the traffic is more frequent and mail traffic, it forms cluster 111A and the traffic that are less frequent and web traffic forms cluster 111B. The same way clustering is also done for Cluster 12A, 11B, 12B, 21A, 22A, 21B, 22B.

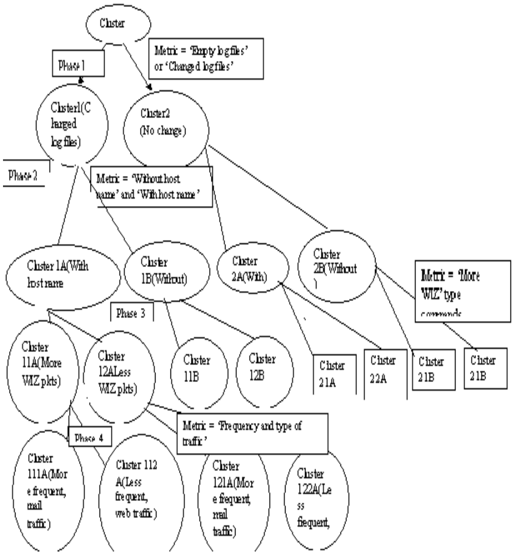

Cluster
Phase 1
Metric = 'Empty log files' or 'Changed log files'
Cluster1(Changed log files)
Cluster2 (No change)
Phase 2
Metric = 'Without host name' and 'With host name'
Cluster 1A(With host name
Cluster 1B(Without)
Cluster 2A(With)
Cluster 2B(Without)
Metric = 'More WIZ' type commands
Phase 3
Cluster 11A(More WIZ pkts)
Cluster 12ALess WIZ pkts)
Cluster 11B
Cluster 12B
Cluster 21A
Cluster 22A
Cluster 21B
Cluster 21B
Phase 4
Metric = 'Frequency and type of traffic'
Cluster 111A(More frequent, mail traffic)
Cluster 112 A(Less frequent, web traffic)
Cluster 121A(More frequent, mail traffic)
Cluster 122A(Less frequent,

Fig1:-Clustering using divisive approach

The output of clustering is a group of clusters with each cluster having its own characteristics. In the above case phase 1 starts with an input of 1 cluster and it becomes 2 clusters as output depending on the log file. Phase 2 starts with an input of 2 clusters and it becomes 4 clusters based on two characteristics log file + host name. Phase 3 starts with 4 clusters and becomes 8 clusters based on one more characteristic frequent WIZ kind of traffic. Phase 4 takes 8 clusters as input and becomes 16 clusters with one more characteristic added that is the frequency and type of traffic. Hence a single cluster as input to the divisive algorithm became 16 clusters each with its own characteristic.

IJERT

Examples for output cluster.

1. Empty content in log file + without host name + more frequent WIZ packets + more frequent mail traffic.
2. Empty content in log file + with host name + more frequent WIZ packets + more frequent mail traffic.

Totally there are now 16 clusters with a combination of all 4 different characteristics.

4.2 Association rule mining.

The next step in this approach is association rule mining. As discussed in the previous section association rule mining can be implemented using several algorithms. In our approach, we are going to use Apriori algorithm for finding the association between the parameters that determine the hacking activity and the hacker.
The confidence and support are measured for the associated parameters.

Ex:-

The data set, which has empty log contents or changed log file contents are getting traffic from some source that doesn't have host name in the DNS lookup.

In this case, the association is as follows:

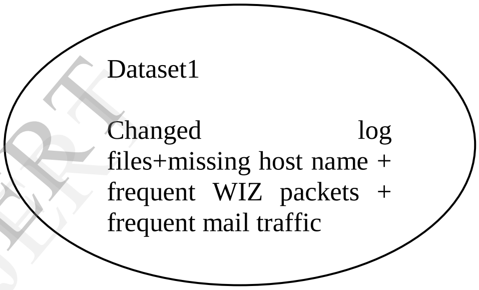
Changed log files => source without host name.

Support = the data sets in the database for which both the antecedent and consequent are satisfied.

Confidence = the satisfied data sets/the datasets which covers only LHS(changed log files.)

The same way the support and confidence are measured for each of the association between the parameters and the associations with high support can be considered for input to the training model.
As in APRIORI algorithm, the previous hacking activities are examined and the data sets of the previous hacking activities and normal activities are analysed.

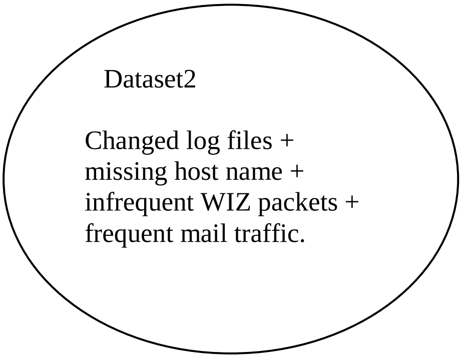
Ex:-The following data sets represent hacking activity.



Dataset1

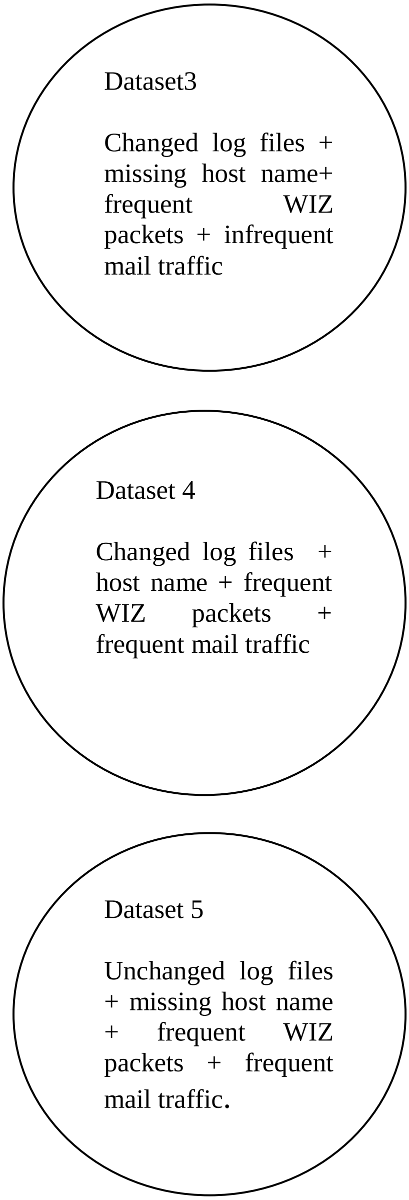
Changed log files+missing host name + frequent WIZ packets + frequent mail traffic

IJERT



Dataset2

Changed log files + missing host name + infrequent WIZ packets + frequent mail traffic.



Dataset3

Changed log files + missing host name+ frequent WIZ packets + infrequent mail traffic

Dataset 4

Changed log files + host name + frequent WIZ packets + frequent mail traffic

Dataset 5

Unchanged log files + missing host name + frequent WIZ packets + frequent mail traffic.

Fig 2: Associated data sets

In the above datasets, we can predict that
Changed log files => missing DNS name - association has good support and confidence level that can be considered for input to the training model.
Missing host name => frequent WIZ packets – association has good support and confidence level that can be considered as input for the training model.
The same way several datasets are examined and several associations are predicted using APRIORI algorithm. The input to the APRIORI algorithm is the clusters obtained from divisive method and the data sets in the database that can be used for training purposes.
The output of the APRIORI algorithm is the association rules that can be predicted. This has to be applied to those 16 clusters and thus the decision can be taken for the clusters and association rules that have to be fed to the training model for the accurate classification.

IJERT

Example for Output :-

Cluster 1 – possibly hacking activity
Cluster 2 – possibly normal activity
………….

4.3 Classification.

This is the final step where the input data set is classified as either of the category whether it is a hacking activity or not. As discussed earlier, classification technique uses several algorithms and models. In our approach we are using neural network training model and backpropogation algorithm.

The input nodes are labelled. The activation function is used in the hidden layer and using the activation function, the output nodes are obtained. The error rate is checked for the training dataset and the corrections are made. Then the actual dataset is fed to the training model and the output is obtained which maps the activity into either of the classification-hacking or normal.

The activation function is obtained as follows based on the previous two steps.

$Weight = w1+w2+w3+w4/4$ where

W1 = changed log files or unchanged log files which takes the values 0.8 and 0.2 respectively.
W2 = without host name and with host name that takes the values 0.7 and 0.3 respectively.
W3 = Frequent WIZ kind of packets and infrequent WIZ packets which takes the values 0.6 and 0.4 respectively.
W4 = Frequent mail traffic and infrequent web traffic which takes the values 0.7 and 0.3 respectively.

If the output weight is equal to or greater than 0.5, the activity is considered as a hacking activity else as a normal activity.

The input nodes are the training data sets. The hidden layer contains the above activation function. The output nodes contain the actual classified node that determines if the activity is hacking or not.

IJERT

Example:-

Consider the cluster that is input to the neural networks with the following characteristics:-

Changed log files	– 0.8
With DNS	- 0.3
Frequent WIZ packets	- 0.6
Frequent mail traffic	- 0.7

The weight of the cluster is,

$Weight = 0.8+0.3+0.6+0.7/4 = 0.6$

Since the weight is above 0.5, this activity is considered as a hacking activity and the source is hacker.

If it is an error, the backpropogation algorithm is followed which considers the error rate and changes the weight accordingly so that the correct activation function can be obtained. Hence the output of the training model yields the correct classification.

5. Conclusion and future enhancement.

The above model clearly depicts how data mining techniques are used with intrusion detection system so that a hacking activity and a hacker can be predicted very accurately. The above model proves that false positives and false negatives of the intrusion detection system can be reduced to a greater extent. This model can be enhanced in future by optimizing the algorithms that are used. Few other parameters of the network traffic can also be considered in this model to make this approach still efficient and accurate.

6. References.

[1] Luca Becchetti, Carlos Castillo, Debora Donato, Stefano Leonardi, and Ricardo Baeza Yates.
Link-based characterization and detection of web spam.
In Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2006

[2] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri.
Know your neighbors: web spam detection using the web topology.
Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 423–430,2007.

[3] B. Zhou, J. Pei, and Z. Tang.
A spamicity approach to web spam detection. In Proceedings of the 2008 SIAM
International Conference on Data Mining (SDM'08),
pages 277{288, Atlanta, GA, USA, 2008. SIAM.

[4] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly.
Detecting spam web pages through content analysis. InProceedings of the 15th International World Wide Web Conference (WWW'06), pages 83{92, New York, NY, USA, 2006. ACM Press

[5] L. Liu, M. Kantarcioglu, and B. Thuraisingham, "Privacy Preserving
Decision Tree Mining from Perturbed Data," Proc. 42nd Hawaii Int'l Conf. System Sciences (HICSS '09), 2009.

[6] M. Shaneck and Y. Kim,
"Efficient Cryptographic Primitives for Private Data Mining,"
Proc. 43rd Hawaii Int'l Conf. System Sciences (HICSS), pp. 1-9, 2010.

[7] K. Chen and L. Liu,
"Privacy Preserving Data Classification with Rotation Perturbation,"
Proc. Fifth IEEE Int'l Conf. Data Mining (ICDM), 2005.

IJERT
IJERT