International Journal of Research Studies in Science, Engineering and Technology [IJRSSET]
Volume 1, Issue 1, April 2014, PP 93-97

Survey on Data Mining Techniques for Intrusion Detection System

Miss.Kavita Patond
CSE Deept,PR
pote(patil)COET,
Name &SGBAU, Amravati, India

Prof. Pranjali Deshmukh
CSE Deept,
PR pote(patil)COET,
Name &SGBAU, Amravati, India

Abstract: *Today, Intrusion Detection Systems have been employed by majority of the organizations to safeguard the security of information systems. Firewalls that are used for intrusion detection possess certain drawbacks which are overcome by various data mining approaches. Data mining techniques play a vital role in intrusion detection by analyzing the large volumes of network data and classifying it as normal or anomalous. Several data mining techniques like Classification, Clustering and Association rules are widely used to enhance intrusion detection. Among them clustering is preferred over classification since it does not require manual labelling of the training data and the system need not be aware of the new attacks. This paper discusses three different clustering algorithms namely K-Means Clustering, Y-Means Clustering and Fuzzy C-Means Clustering. K-Means clustering results in degeneracy and is not suitable for large databases. Y-Means is an improvement over K-means that eliminates empty clusters.*

Four issues such as Classification of Data, High Level of Human Interaction, Lack of Labelled Data, and Effectiveness of Distributed Denial of Service Attack are being solved using the algorithms like EDADT algorithm, Hybrid IDS model, Semi-Supervised Approach and Varying HOPERAA Algorithm respectively.

Keywords: *HybridIDS, AnomolyDetecton, Clustering, K-means*

1. INTRODUCTION

In this modern world intrusion occurs in a fraction of seconds. Intruders cleverly use the modified version of command and thereby erasing their footprints in audit and log files. Successful IDS intellectually differentiate both intrusive and non intrusive records. Data mining based IDS can efficiently identify these data of user interest and also predicts the results that can be utilized in the future. Data mining or knowledge discovery in databases has gained a great deal of attention in IT industry as well as in the society. Data mining has been involved to analyze the useful information from large volumes of data that are noisy, fuzzy and dynamic. Figure 1 illustrates the overall architecture of IDS. It has been placed centrally to capture all the incoming packets that are transmitted over the network. Data are collected and send for pre-processing to remove the noise; irrele-vant and missing attributes are replaced. Then the pre-processed data are analyzed and classified according to their severity measures. If the record is normal, then it does not re-quire any more change or else it send for report generation to raise alarms. Based on the state of the data, alarms are raised to make the administrator to handle the situation in advance. The attack is modeled so as to enable the classification of net-work data. All the above process continues as soon as the transmission starts.

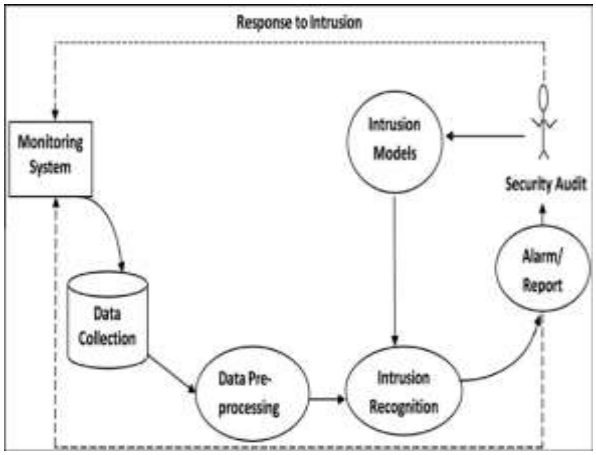

Figure1: *Overall structure of Intrusion Detection System*

2. LITERATURE SURVEY

G.V. Nadiammai, M. Hemalatha [1], With the help of varying clock drift, the client can easily communicate with the server with minimum contact initiation trails and the improved max-imum delivery latency has been achieved. Varying HOPERAA algorithm shows 99% as receiving capacity at 100 ms and slightly deviates

for 40 ms rate with better throughput capac-ity. Our experimental results proved that the proposed algo-rithms solves the above mentioned issues and detects the attacks in an effective manner compared with other existing works. Thus, it will pave the way for an effective means of intrusion detection with better accuracy and reduced false alarm rates.

Shakiba Khademolqorani, Ali Zeinal Hamadani [2] , Nowadays, decision makers invariably need to use decision support technology in order to tackle complex decision making problems. In this area, data mining has an important roleto extract valuable information. Also, the successful application of data mining technology requires that one possess specific DM decision-making skills. For instance, the effective application of a data mining process is littered with many difficult and technical decisions (i.e. data cleansing, feature transformations, algorithms, parameters, evaluation, etc.). Consequently, the use of data mining and decision support methods, including novel visualization methods, can lead to better performance in decision making, improve the effectiveness of developed solutions and enables tackling of new types of problems thathave not been addressed before. On the other hand, the MCDM method deals with the vast area of decision making; choosing the best option among various alternatives, and optimization of goal among multi-objective situations. Therefore, the decision support systems (DSS), data mining (DM) and multiple criteria decision making (MCDM) are complementary methods for decision making process, and they have also a strong link with expert systems.

Zubair Md. Fadlullah,etl [3], In this article, they highlighted the importance on designing appropriate intrusion detection systems to combat attacks against cognitive radio networks. Also, we proposed a simple yet effective IDS, which can be easily implemented in the secondary users' cognitive radio software. Our proposed IDS uses non-parametric cusum algorithm, which offers anomaly detection. By learning the normal mode of operations and sys-tem parameters of a CRN, the proposed IDS is able to detect suspicious (i.e., anomalous or abnormal) behavior arising from an attack. In particular, we presented an example of a jamming attack against a CRN secondary user, and demonstrated how our proposed IDS is able to detect the attack with low detection latency. In future, our work will perform further investigations on how to enhance the detection sensitivity of the IDS.

Mueen Uddin,etl [4], This paper has focused on the efficiency and performance of the new IDS: called signature-based multi- layer IDS using mobile agents. It then discusses the development of a new signature- based ID using mobile agents. The proposed system uses mobile agents to transfer rule-based signatures from large complementary database to small signature database and then regularly update those databases with new signatures detected.

Sahilpreet Singh, Meenakshi Bansa [5], This study is approached to discover the best classification for the application machine learning to intrusion detection. For this, we have presented different neural based data mining classifier algorithms to classify attacks in an efficient manner. After doing experimental work, It is clear that Multilayer Perceptron feed forward neural network has highest classification accuracy and lowest error rate as compared to other neural classifier algorithm network. To enhance the results the feature reduction techniques is applied. The neural algorithms are applied to NSL KDD dataset by reducing its attributes and implemented using WEKA machine learning tool.They showed that machine learning is an effective methodology which can be used in the field of intrusion detection. In future, They will propose a new algorithm which will integrate Multilayer Perception Network with fuzzy inference rules to improve the performance

Paul Dokas,etl[6] Several intrusion detection schemes for detecting net-work intrusions are proposed in this paper. When applied toKDDCup'99 dataset, developed algorithms for learn-ing from rare class were more successful in detecting network attacks than standard data mining techniques. Experimental results performed on DARPA 98 and real network data indicate that the LOF approach was the most promising technique for detecting novel intrusions When performing experiments on DARPA'98data, the unsupervised SVMswere very promising in detecting new intrusions but they had very high false alarm rate. Therefore, future work is need edin order to keep high detection rate while lowering the false alarm rate. In addition, for the Mahal anobis based approach, we are currently investigating the ideao defining several types of "normal" behavior and measuring the distance to each of them in order to identify the anomalies.

3. TECHNIQUES FOR IDS

Each malicious activity or attack has a specific pattern. The patterns of only some of the

attacks are known whereas the other attacks only show some deviation from the normal patterns. Therefore, the techniques used for detecting intrusions are based on whether the patterns of the attacks are known or unknown. The two main techniques used are:

3.1. Anomaly Detection

It is based on the assumption that intrusions always reflect some deviations from normal patterns. The nor mal state of the network, traffic load, breakdown, protocol and packet size are defined by the system administrator in advance. Thus, anomaly detector compares the current state of the network k to the nor mal behaviour and looks for malicious behaviour. It can detect both known and unknown attacks.

3.2. Misuse Detection

It is based on the knowledge of known patterns of previous attacks and system vulnerabilities. Misuse detection continuously compares current activity to known intrusion patterns to ensure that any attacker is not attempting to exploit known vulnerabilities. To accomplish this task, it is requir ed to describe each intrusion pattern in detail It cannot detect unknown attacks.

4. REVIEW OF DATA MINING TECHNIQUE IN INTRUSION DETECTION

Data Mining refers to the process of extracting hidden, previously unknown and useful information from lar ge databases. It extracts patterns and concentrates on issues relating to their It is a convenient way of extracting patterns and focuses on issues relating to their feasibility, utility, efficiency and scalability . Thus data mining techniques help to detect patterns in the data set an use these patterns to detect future intrusions in similar data. The following are a few specific things that make the use of data mining important in an intrusion detection system:

1. Manage firewall rules for anomaly detection.
2. Analy ze large volumes of network data.
3. Same data mining tool can be applied to different data sources.
4. Perfor ms data summarization and visualization.
5. Differentiates data that can be used for deviation analysis.
6. Cluster s the data into groups such that it possess high intra-class similarity and low inter-class similarity .

Data mining techniques play an important role in intrusion detection systems. Different data mining techniques like classification, clustering, association rule mining are used frequently to acquire information about intrusions by observing and analy zing the networ k data[1]. The following describes the different data mining techniques.

A. Classification

It is a supervised learning technique. A classificat ion based IDS will classify all the network traffic into either normal or malicious. Classification technique is mostly used for anomaly detection. The classification process is as follows:

1. It accepts collection of items as input.
2. Maps the items into predefined groups or classes defined by some attributes.
3. After mapping, it outputs a classifier that can accurately predict the class to which a new item belongs.

B. Association Rule

This technique searches a frequently occurring item set from a large dataset. Association rule mining deter mines association[6] rules and/or correlation relationships among large set of data items. The mining process of association rule can be divided into two steps as follows:

1. Frequent Item set Generation Generates all set of items whose support is greater than the specified threshold called as min support.

2. Association Rule Generation From the previously generated frequent item sets, it generates the association rules in the form of ―if then statements that have confidence

greater than the specified threshold called as min confidence.

3. The network data is arranged into a database table where each row represents an audit record and each column is a field of the audit records.

4. The intrusions and user activities shows frequent correlations among the network data. Consistent behaviors in the networ k data can be captured in association rules.

5. Rules based on networ k data can continuously merge the rules from a new r un to aggregate rule set of all previous runs.

6. Thus with the association rule, we get the capability to capture behavior for correctly detectin intrusions and hence lowering the false alarm rate.

C. Clustering

It is an unsuper vised machine learning mechanism for discovering patterns in unlabeled data. It is used to label data and assign it into clusters where each cluster consists of members that are quite similar. Members from differ entclusters are different from each other. Hence clustering methods can be useful for classifying networ k data for detecting intrusions. Cluste ring can be applied on both Anomaly detection and Misuse detection. The basic steps involved in identifying intrusion are follows [3]:

1. Find the largest cluster, which consists of maximum number of instances and label it as normal.
2. Sort the remaining clusters in an ascending order of their distances to the largest cluster.
3. Select the fir st K1 clusters so that the number of data instances in these clusters sum up to ¼`N and label them as nor mal, where ` is the percentage of normal instances.
4. Label all other clusters as malicious.
5. After clustering, heuristics are used to automatically label each cluster as either nor mal or malicious. The self-labelled cluster s are then used to detect attacks in a separate test dataset.

From the three data mining techniques discussed above clustering is widely used for intrusion detection because of the following advantages over the other techniques:

1. Does not require the use of a labeled data set for training.
2. No manual classification of training data needs to be done.
3. Need not have to be aware of new types of intrusions in order for the system to be able to detect them.

5. CLUSTERING TECHNIQUES USED IN IDS

Several clustering algorithms have been used for intrusion detection. All these algorithms reduce the false positive rate and increase the detection rate of the intrusions. The detection rate is defined as the number of intrusion instances detect ed by the system divided by the total number of intrusion instances present in the data set. The false positive rate is defined as total number of nor mal instances that were incorrectly classified as intrusions defined by the total number of nor mal instances. Some of the clustering techniques such as K-Means Clustering, Y-Means Clustering and Fuzzy C-Means Clustering are discussed below:

A. K-Means Clustering

K-Means algorithm is a hard partitioned clustering algorithm widely used due to its simplicity and speed. It uses Euclidean distance as the similarity measure. Hard clustering means that an item in a data set can belong to one and only one cluster at a time. It is a clustering analysis algorithm that groups items based on their feature values into K disjoint cluster s such that the items in the same cluster have similar attributes and those in different clusters have different attributes. The Euclidean distance function used to compute the distance (i.e. Similarity) between two items is given follows

.....1

where, p = (p1, p2,…, pm) and q = (q1, q2,…,qm) are the two input vectors with m quantitative

attributes. The algorithm is applied to training datasets which may contain nor mal and abnormal traffic without being labeled previousl y. The main idea of this approach is based on the assumption that nor mal and abnormal traffic for m different cluster s. The data may also contain outliers, which are the data items that are very different from the other items in the cluster and hence do not belong to any cluster. An outlier is found by comparing the radiuses of the data items; that is, if the

radius of a data item is greater than a give threshold, it is considered as an outlier. But this does not disturb the K-means clustering process as long as the number of outlier s is small.[2]

K-Means Clustering Algorithm is as follows:

i) Define the number of clusters K. For example, if K=2, we assume that normal and abnormal traffic in the training data for m two different clusters.

ii) Initialize the K cluster centroids. This can be done by randomly selecting K data items from the data set.

iii) Compute the distance from each item to the centroids of all the cluster by using the Euclidean distance metric which is used to find the similarity between the items in data set.

iv) Assign each item to the cluster with the nearest centroid. In this way all the items will be assigned to different clusters such that each cluster will have items with similar attributes.

v) After all the items have been assigned to different clusters re-calculate the means of modified clusters The newly calculated mean is assigned as the new centroid.

vi) Repeat step (iii) until the cluster centroids do not change.

vii) Label the cluster as normal and abnor mal depending on the number of data items in each cluster.

6. CONCLUSION

Data mining techniques are widely used because of their capability to drastically improve the performance and usability of intrusion detection systems. Different data mining techniques like classification, clustering and association rule mining are very helpful in analyzing the network data. Since large amount of network traffic needs to be collected for intrusion detection, clustering is mor e suitable than classification in the domain of intrusion detection as it does not require labeled data set thereby reducing manual efforts. Data mining techniques can detect known as well as unknown attacks. Data mining technology helps to understand nor mal behavior inside the data and use this knowledge for detecting unknown intrusions. Three clustering algorithms namely K-means, Y-means and Fuzzy C-means have been discussed. Each of these has both advantages and disadvantages and are an improvement over the other. Among these Fuzzy C-Means clustering can be considered as an efficient algorithm for intrusion detection since it allows an item to belong to more than one cluster and also measures the quality of partitioning. The technique can be used for large data sets as well as data sets that have overlapping items

REFERENCES

[1] G.V. Nadiammai, M. Hemalatha. "Effective approach toward Intrusion Detection System using data mining techniques" Received 22 May 2013; revised 29 August 2013; accepted 27 October 2013Egyption informatics journal. Elsevier

[2] Shakiba Khademolqorani, Ali Zeinal Hamadani "An Adjusted Decision Support System through Data Mining and Multiple Criteria Decision Making" The 2nd International Conference on Integrated Information Elsevier 2013

[3] Zubair Md. Fadlullah, Hiroki Nishiyama, "An Intrusion Detection System (IDS) for Combating Attacks Against Cognitive Radio Networks" 2013 IEEE

[4] Mueen Uddin , Azizah Abdul Rehmanl etl "Signature-based Multi-Layer Distributed Intrusion Detection System using Mobile Agents" International Journal of Network Security, Vol.15, No.1, PP.79-87, Jan. 2013

[5] Sahilpreet Singh Meenakshi Bansa , "Improvement of Intrusion Detection System in Data Mining using Neural Network" Volume 3, Issue 9, September 2013 IJARCSSE

[6] Paul Dokas, Levent Ertoz,, "Data Mining for Network Intrusion Detection"

[7] Pise, P. J., 1982, Laterally loaded piles in a two-layer soil system., J. Geotech. Engrg. Div., 108(9), 1177–1181.

[8] Poulos, H. G., 1971, Behavior of laterally loaded piles-I: Single piles., J. Soil Mech. and Found. Div., 97(5), 711–731.

[9] Reese, L. C., and Matlock, H., 1956, Non-dimensional solutions for laterally loaded piles with soil modulus assumed proportional to depth., Proc., 8th Texas Conf. on Soil Mechanics and Foundation Engineering, Austin, Texas, 1–23

[10] Reese, L. C., and Welch, R. C., 1975, Lateral loading of deep foundations in stiff clay., J. Geotech. Engrg. Div., 101(7), 633–649.