International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

General Article

Detecting Threats in IDS using Data Mining Techniques

Sukhleen[Á*] and Gurpreet Kaundal[Á]

[Á]Department of Computer Science and Technology, Lovely Professional University, Phagwara, India

Accepted 01 April 2014, Available online 10 April 2014, **Vol.4, No.2 (April 2014)**

Abstract

Achieving security has become one of the most critical factors as more and more sensitive data and information is being maintained and manipulated online. Intrusion Detection System (IDS) is one of the most popular methods which is used to detect malicious activities and maintains the security of the system. IDS can use either anomaly based approach or misuse based approach. In order to detect the malicious activities large amount of data is analyzed. For analyzing data using data mining techniques are best way to achieve the required objective. This paper discusses the various data mining techniques such as clustering, classification and association rules that can be used with IDS so that huge amount of data can be analyzed and attacks can be detected.

Keywords: *Data Mining, Knowledge Discovery, Intrusion Detection, Misuse Detection, Anomaly Detection, Clustering, Classification, Association.*

Introduction

Today we rely so much on computer based information system that it has become an integral part of our life and servers various functions and activities of our business as well as in our daily routines. Among all security becomes an indispensable factor which ensures the integrity, confidentiality and availability of information. Various virus detecting software, firewalls are available to protect the system against attacks. An attacker may be either outsider or insider to an organization.

Therefore we need to have some effective components which will protect data from attack. In recent days IDS has been deployed as a security component which acts as a firewall between the user computer system or data and the network. As its name describes it is used to detect the intrusion and generates alarm for the user. IDS come in two flavors: Network Intrusion Detection System (NIDS) and Host based Intrusion Detection System (HIDS). NIDS works at network level where as HIDS works at host level. The network based IDS is placed along a network boundary and analysis all the traffic exists on that boundary. Hence, they run on assigned machines that monitor the flow of network or also perform analysis with the help of firewall. On the other hand HIDS can be installed on many different types of machines namely servers, workstations and notebook computers. In fig. 1, both the NIDS and HIDS have been described. All the traffic will pass through the NIDS on the network. It will detect the attacks by analyzing the network packets. On the other hand, HIDS are installed as an agent on the host and these can look into the systems in order to detect the malicious activity.

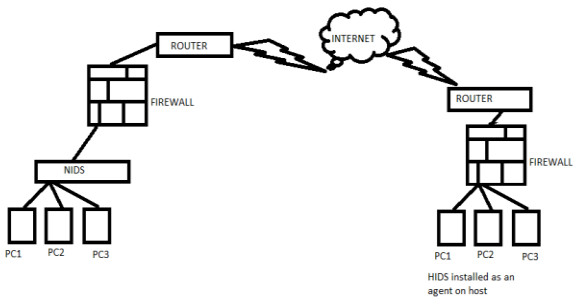

Fig. 1*:* NIDS and HIDS

But the existing system has the following drawbacks (Wang Pu and Wang Jun qing *et al*, 2011):

- Data overload. Since large amount of data is generated everyday over the network like system logs etc., in order to analyze such a large amount of data ct a large amount of information traditional IDS face difficulties and not able to give better results.
- False positives. Traditional IDS generates alarm even if the attack does not take place. It sometimes by mistake generates alarm when a normal system activity takes place.
- False negatives. Another weakness is IDS system does not alert the user when there is an actual attack has occurred in the system.

So to overcome these problems of traditional IDS data mining techniques plays very important role.

Data Mining and Intrusion Detection

Data Mining Technology

*Corresponding author: **Sukhleen**

Data Mining refers to extraction or mining of useful information or knowledge from large amount of data (Ming Xue and Changjun Zhu *et al*, 2009) Data Mining process also known as knowledge mining from data. Thus we can say that it is the process of discovering interesting knowledge from large amount of data stored in large databases, data warehouses or in other repositories. People often term Data Mining as Knowledge Discovery in Database (KDD) which is process of discovering knowledge in databases. (Jiawei Han and Micheline Kamber *et al*, 2006) It consists of an iterative sequence as shown in fig.2:

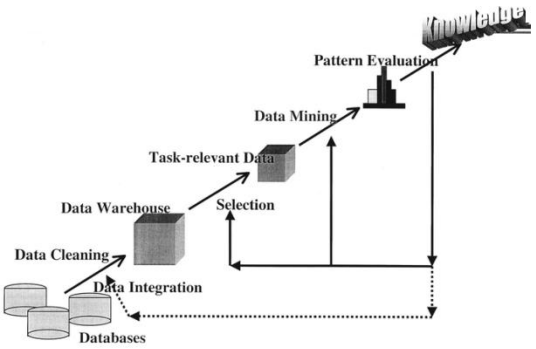

Fig.2: Knowledge Discovery Process

- Data cleaning: remove noise or irrelevant data.
- Data integration: where multiple data sources may be combined.
- Data selection: where the relevant data are retrieved from the database.
- Data transformation: where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- Data mining: an essential process where intelligent methods are applied in order to extract data patterns.
- Pattern evaluation: to identify the truly interesting patterns representing knowledge based on some interestingness measures.
- Knowledge presentation: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Intrusion Detection Technology

Intrusion Detection Systems (IDS) is a combination of software and hardware that attempts to perform intrusion detection. It is a process of analyzing the information for intrusion and raises the alarm when a possible intrusion occurs in the system. The network data source of intrusion detection consists of large amount of textual information, which is difficult to comprehend and analyze. The intrusion based on two types based on data analyses: Misuse detection and Anomaly detection (Ming Xue *et al*, 2009).

Misuse Detection

Misuse Detection detects the attack by comparing them with the previously stored behavior in audit trails. It increases the speed of detecting attacks and lowers down the false alarm generation. However, if any mistake happens while recording the signatures, it will increase the false alarm rate. In signature based intrusion detection known intrusion patterns have to be hand coded and they are unable to detect any future (unknown) intrusions that have no matched patterns stored in the system. (A.A. Ghorbani, W. Lu, M. Tavallaee *et al, 2010)* The basic idea of Misuse detection is to first collect the data from various data sources including network traffic, audit trails and system call trace. Then the collected data is converted into a form which is understandable by other components of the system. A system profile is maintained which has certain features such as when packet enters the network, at which time the connection is established and the port being used as shown in fig. 3:

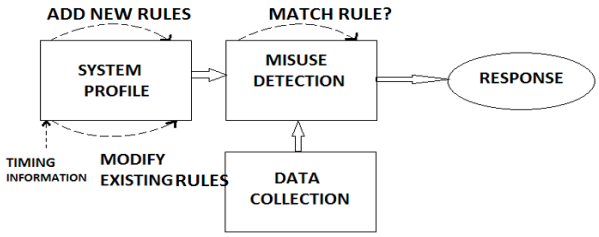

Fig. 3: Misuse Detection

Anomaly Detection

Anomaly detection observes the ongoing system activities and then makes decision which activities are normal and which are intrusive. (A.A. Ghorbani, W. Lu, M. Tavallaee *et al, 2010)* The anomaly detection model has four components as shown in fig. 4, Data collection, normal system profile, anomaly detection and response. All the data including network data as well as user data is collected by data collection module and normal system files are created using specific technique. After that the next component anomaly detection will decide some threshold value that after what percentage value the ongoing activity in the system will be flagged as abnormal. At last the response will give report to the user.

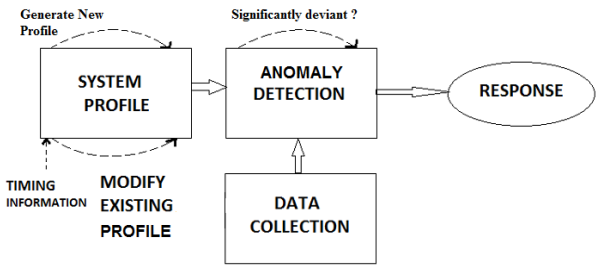

Fig. 4: Anomaly Detection

Data Mining Techniques and Intrusion Detection

Data mining is the process of monitoring and verifying the data and events that takes place in a computer network system so as to detect the attacks (Jiawei Han *et al*, 2011).

Therefore by combining both the data mining and IDS will overcome the existing weaknesses of traditional IDS as well as better results will come out. Different data mining technique like clustering, classification and association rules can be used with IDS. Now a day data mining is becoming an important component which analyzes the network data and flagged out the intrusion related information for the user.

Clustering

It arranges the data into meaningful cluster or groups of objects which have similar characteristics. Basically it defines the classes and in each class it puts the similar data under one class. There are four types of clustering: k-means clustering, fuzzy c-means clustering, Mountain clustering, and Subtractive-clustering (A. M. Chandraprakash and K. Raghveer *et al, 2012)*. K-means clustering is one of the main clustering algorithms which marks the data points with a random number and assign to k cluster and after that the centers of clusters are calculated and data points are assign to the clusters which are closest to the calculated center. Fuzzy c-mean clustering is an improved version of k-means clustering algorithm. In this algorithm the centre of clusters and allocation of data points are evaluated on the basis of Euclidian distance. Mountain clustering algorithm is very simple to find the clusters there is no need to pre specify the number of clusters as in K-means and Fuzzy c-mean clustering, this techniques find the clusters based on mountain function which means area with high density value will be considered as cluster and so on. Subtractive clustering technique is an extension of mountain clustering in which instead of making clusters bases on the density function, it uses data points to calculate the density function. In other words it considers each data point as a cluster and calculates the data point density around it and forms the cluster. After clusters are formed they are labeled as normal or abnormal clusters. Since clustering is an unsupervised method finding useful patterns of information. Once the clusters have been defined, the data points or records which go beyond the defined clusters will be considered as an attack.

Classification

This technique is based on machine learning and used to classify each item in a dataset according to predefined set of classes or groups. This technique makes use of mathematical techniques such as linear programming, decision trees, neural network and statistics. In classification, the software or a model is developed which tells how to classify the data items into groups. It is somehow related to the clustering in which it also forms the groups of related data. But in classification method the way of classifying or analyzing the data must be known to the user. The main goal of classification is to analyze the new records and they will be classified either as normal or abnormal. A classification model is build using different algorithms and training data sets. The training data set contains labelled sequence of normal as well as abnormal data. After the model is developed it is used to classify the record as normal or abnormal.

The classification model can be rule-based, decision-tree based, association-rule based, Bayesian-network based, or neural network based (Chang-tien Lu, Arnold P. Boedihardjo and Prajwal Manalwar *et al*, 2005). In rule based classification method RIPPER (Repeated Incremental Pruning to Produce Error Reduction) system is mostly used to build the classifier. As the name specifies it learns iteratively and generate rule set directly from the training dataset (Mlungisi Duma, Bhekisipho Twala and Tshilidzi Marwala *et al*, 2010). It is very fast and efficient algorithm for dealing with large and noisy dataset. On the other hand in decision tree based classification method a decision tree is build which has flow chart like structure in which the central node is known as root node and internal node consists of a condition on attributes and branch represents the result of the condition. At last the leaf nodes represent the class labels. The mostly used decision based classification modules are ID3 and C4.5. A neural network based method is another way to exploit the classification algorithm in which system detects the new intrusion by learning the sequence of commands executed by a normal user. This capability of learning lead to the detection of new types of attacks.

In order to detect the intrusion by following any of the classification method first of all we need to built a classifier using pre defined labeled training dataset which means either a normal or abnormal sequence of data. After that the classifier will analyze the new input data. If the data is matched with the normal list it will mark the data as normal otherwise abnormal.

In contrast to the clustering techniques classification is not used often because in order to built a classifier large amount of data is needed as training dataset to train the classifier. In addition, it works well when used for known attacks otherwise the false alarm generation is high.

Association rules

Association rules are used to discover patterns on the basis of relationship between the various items of the same transactions and are also called as relation technique and are used in market basket analysis. The task of association mining is to discover the association rules which have two measurements i.e. Support and Confidence. (Jiawei Han and Micheline Kamber *et al*, 2006) Such as, Rule: $X \rightarrow Y$ or antecedent (X) implies consequent (Y). Support = the number of time a rule shows up in a database. Confidence = Conditional probability of Y given X. (Chang-tien Lu ,Arnold P. Boedihardjo and Prajwal Manalwar *et al*, 2005) The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule where as confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent to the number of transactions that include all items in the antecedent.

Firstly all the combinations of associated data items are considered whose support is more than the decided min_sup (minimum support). All the selected data items will be referred as frequent item sets. Second using the

generated frequent item sets rules are generated. These rules must satisfy the user defined min_conf (minimum confidence) and support.

Association rules analyze the incoming data traffic and classify as either normal or intrusive data. (Zulaiha Ali Othman and Entisar E. Eljadi *et al*, 2011) There are many association rule techniques are available among which Apriori, fuzzy Apriori and FP-growth are the best association rule techniques of data mining.

Conclusion

In this paper various data mining techniques has been discussed in conjunction with intrusion detection system. Also various limitations of existing IDS has been discussed which lead to deployment of data mining techniques with IDS so that both the known and unknown attacks can be detected. A lot of work has already been done but still there are some limitations in existing data mining techniques which can be overcome with further study. Moreover by using hybrid techniques with IDS we can improve the efficiency as well as the performance of the system and get better results.

References

Wang Pu and Wang Jun qing (2011), Intrusion Detection System with the Data Mining Technologies, *IEEE*, 490-492.

Ming Xue and Changjun Zhu (2009), Applied Research on Data Mining Algorithm in Network Intrusion Detection, *International Joint Conference on Artificial Intelligence*, 275-277.

Jiawei Han and Micheline Kamber, Second Edition (2006), Data Mining Concepts and Techniques.

A.A. Ghorbani, W. Lu, M. Tavallaee (2010), Network Intrusion Detection and Prevention Concepts and Techniques, *Springer*, 27-53.

Jiawei Han ,Micheline Kamber (2011), Data Mining: Concepts and Techniques, Morgan Kaufmann.

A. M. Chandrashekhar and K. Raghveer (2012), Performance evaluation of data clustering techniques using KDD Cup-99 Intrusion detection data set, *IJINS*, vol.1, no.4, 294-305.

Chang-tien Lu, Arnold P. Boedihardjo and Prajwal Manalwar (2005), Exploiting Efficient Data Mining Techniques to Enhance Intrusion Detection Systems, *IEEE*, 512-517.

Mlungisi Duma, Bhekisipho Twala and Tshilidzi Marwala (2010),Improving the Performance of the Ripper in Insurance Risk Classification: A Comparitive Study Using Feature Selection.

Zulaiha Ali Othman and Entisar E. Eljadi (2011), Network Anomaly Detection Tools Based on Association Rules, *International Conference on Electrical Engineering and Informatics*, 1-7.