IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Analyzing network traffic to detect anomaly based intrusion using SVM.

Bhopale Nandkishor G.[1], Sangve Sunil S[2]

Research Scholar ME student, Department of Computer Engineering, ZEAL College of Engineering and Research, Narhe, Pune, India [1]

Assistant Professor, Department Department of Computer Engineering, ZEAL College of Engineering and Research, Narhe, Pune, India [2]

Abstract: In the growing Era of Information Technology, the use of Internet is the basic need of all, but it also caused new problem called Network Security. To solve these problem Intrusion Detection Systems (IDSs) are used. IDS provide the information related to malicious activities and protect network from vulnerabilities and provide better information security. Intrusion detection system provides security for huge network like Internet. Huge amount of data is transferred on the network these days hence system should analyze and detect the malicious activities. The system should analyze and recognize intrusion in effective way and in less time. The main hurdle in network security is to enhance the intrusion detection system. In this paper we proposed Hybrid anomaly Intrusion Detection System Using Outlier Mining with Support Vector Machine on the basis of TCP transmission control protocol header information and other attributes. Here we are using different approaches from that one is k-mean clustering and another is one-class support vector machine to formulate and model different sessions already presented in the dataset provided by the MIT DARPA 99 dataset. After that we provide the testing set to the model for predicting the attack scenarios of session.

Keywords: IDS, k-means clusterig, one class SVM, outlier minning, TCP.

I. INTRODUCTION

INTRUSION detection system is used for detecting the malicious activity in the computer network and prevents it from threat. As day by day the technology changing, it also cause the new risks for the network and hence the new type of attacks came in picture, an intrusion detection system should able to tackle with such newly attacks effectively and provide security. Intrusion detection system identifies the malicious activities in the network environment and traffic in the network and make the network secure from threats. In this era of big data the huge amount of data is generated and transfer over the network daily and also collected on the network for analysis [1].

The transmission control protocol is commonly used for the purpose of exploiting the misuse and different forms of network attacks. In the network traffic there may be chance of having the malicious information executed with the help of TCP/IP protocols without being noticed or blocked by the firewall software and traditional IDS unable to rectify the new attacks. Hence the requirement of IDS that should be deploy easily for the TCP protocol data detection for attack. Research are going on the anomaly intrusion detection system, but no any globally used in practice system not in picture till now [2,3].The important reasons behind the anomaly intrusion detection systems are under development are that the deployments of such systems are difficult. Such system most of the time uses machine learning approaches and the data mining techniques for the input features and that contains very complicated features for their inputs and make the task difficult. Feature extraction from the raw TCP network traffic is challenging task. Most usable example is of the popular dataset that is KDD 99 dataset [4]

where 13 out of 41 features are extracted based on the domain knowledge. The data extraction from such a huge dataset is challenging task and no freely tools are available to extract such features form raw TCP data. To avoid this type of problem instead of taking the complicated feature vectors we use TCP raw data which can be made easily available with the help of tcpdump tool. The researchers found that different patterns of TCP flags that are associated with anomalies. In this research we examine the TCP header information with the help of TCP/IP packets which are normally used for the high speed data transfer and network status detection.

Many anomaly-based IDS based on raw TCP data has been proposed in last decade, some of them like PHAD/ALAD [5] are based on the packet. These system mainly based on the information of individual packets for detecting the malicious activities and so these systems unable to detect high level anomalous events that could available in the session having the many number of packets even the single packet looks like fine. To overcome the drawback of such issues the session based NATE [6] approach is used. In our proposed method we would like to implement session based approach. The dataset used in our approach is MITS DARPA 99 tcpdump dataset, to verify the performance of the model. We are

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

considering the dataset for the proof of concept (POC) of the proposed approach of Intrusion detection system. Our approach is applied on the tcpdump files. The main aim of the research is to provide useful and practical intrusion detection system to the world which is also easy to deploy and very secure and can be used with firewall and other security mechanism. We are also using the combination of the multiple machine learning algorithms for improving the accuracy of the system dynamically.

II. LITERATURE SURVEY

A. Transmission Control Protocol (TCP)

For enabling reliable and ordered transmission of traffic in Internet TCP protocol is used which is connection oriented protocol. For Establishment of trusted connection TCP header information is important. Unique identifiers of the connection to the client and server the TCP sequence and acknowledgement numbers used and also they provide confirmation about data reception. To control state of the TCP connection TCP flags are used. Three way handshakes must be completed between client and server before any data exchange takes place. Party can reset the connection with TCP RST packets by initiating termination of the connection. Thus, we have to focused on weakness of protocol where intruders look to exploit and can come with various attacks.

B. Intrusion Detection Systems

To monitoring network activities for malicious activities and to generate alarms and logs an intrusion detection system device/tool is used. It is divided into two types based on work.

Signature-based IDS

For pattern matching on known signatures leads to high accuracy for detecting threat's a rule-based approach used. But it fails to detect novel attacks as novel attack signatures are not available for pattern matching. Drawback of this IDS is its downfall to detect new attack and also it neglect minor variation of known patterns as well as an overhead costs in order to maintain signature databases.

Anomaly-based IDS

For mining information from available data a statistical approach used in which analyses the deviation from normal activities. It can detect novel attacks by comparing suspicious ones with normal traffics but has high false alarm rate due to difficulty of generating practical normal behaviour profiles for protected systems[9].for anomaly based IDS recently different statistical techniques have been applied. When there are no structured knowledge about the pattern of data these techniques used. In anomaly detection many data mining/machine learning algorithms are considered good for attack detection, in which decision tree is considered one of the most powerful and effective ways [10,3]. In addition, k- nearest neighbour (KNN), naive Bayes, artificial neural network (ANN), support vector machine (SVM),random forest, and fuzzy logic are known to be used for learning and classifying the network traffic data as normal or attacks[10].for improving result of instruction detection as compared to the result of a single classier ensemble learning used, which is a combination of multiple machine learning approaches.

C. Anomaly-based IDS using Outlier Mining

It can be regarded as an observation as to arouse suspicions that it was generated by a different mechanism. Researchers have tried two approaches either to model only normal data or both normal and abnormal data for finding intruders. For modeling both normal and abnormal data system allowed to be tight on false positive and false negative as both the normal and abnormal data needs to be modeled; but limitations come across in modeling abnormal patterns.Similarly only for normal patterns system allows to model the boundaries of normal data but due to unknown nature of the boundaries of abnormal data, it gives a possibility of false positive as well if it comes to overlap with the normal data. For defining the threshold of the system to balance the number of true positives and false negatives proper tuning needs to be done. For finding intrusions outlier mining is a popular technique [6,11].

Network analysis of anomalous traffic events(NATE)[6]is a good example of an anomaly based IDS using outlier mining. It's a low cost network IDS, which measures only TCP packets headers information.For generating several concise descriptions using clustering to model normal network traffic behaviors. The measure of detection is the deviations of new TCP header data from the existing clusters,which allows for high speed network traffic monitoring once the normal behavior base is built. By performing a cluster analysis and found 7 such clusters that represent their normal TCP sessions. To perform boundary analysis chebyshev's inequality was used. To detect intrusions in their test sessions mahalanobis distance or euclidean distance was used. They have identified that false positives could be a potential problem with this approach, and hence, they have used significance level as a tuning parameter in response to decrease the false positives.

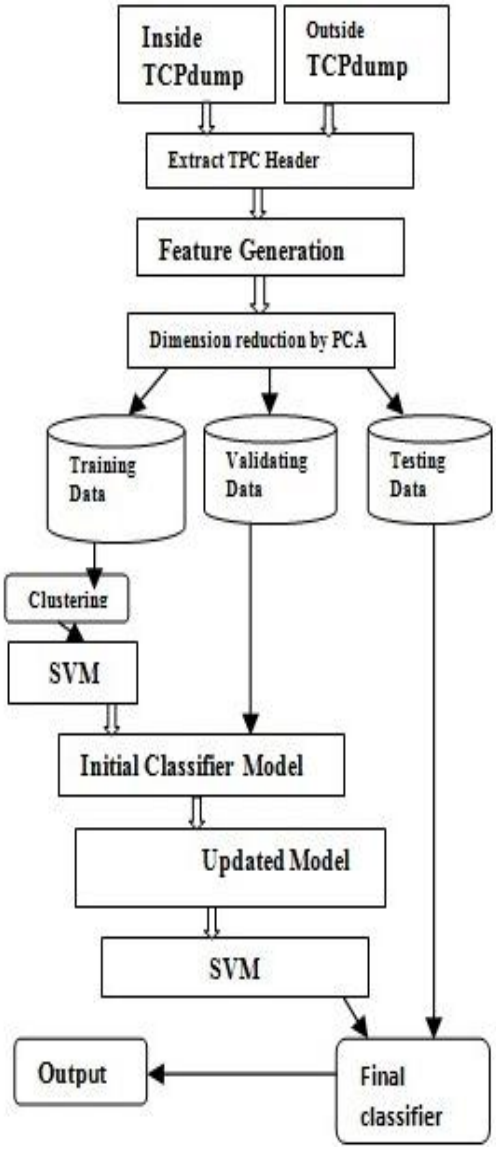
III.PROPOSED METHOD

The main objective of proposed method is to model the normal network data and after that prediction of the abnormality on the basis of deviation from the normal type of pattern. There may be different types of normal data and contains lot of variations. Hence the main task is to analyze the normal patterns and then try to distinguish it from the abnormal attacks data. In this research we are trying to exploit the TCP protocol for the network data and also going through the TCP header for understanding the normal patterns. To reduces the analysis complexity TCP headers are used as it can use for monitoring high speed network traffic. The main goal of the system is to formulate model for normal behavior data for the base and from that base identify the anomalous data. We partially

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

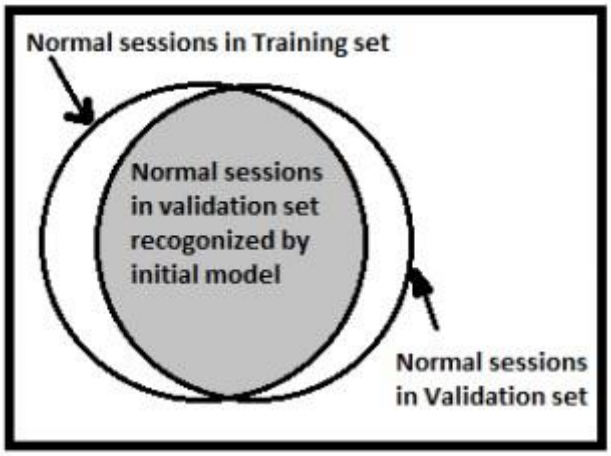
inspired by the method named NATE for defining the TCP sessions and the outlier mining is based on the TCP header information and the k-mean clustering. Different clustering improvements such as clustering quality analysis, one-class SVM, and model validation and updating are used. The result given by our method is much better than NATE. This method is using TCP header for detecting the anomaly intrusion in real time.

For accurate model we tried the system with different clustering and one-class classifier with different model parameter for improvement of the system. Finally we got the model which has high accuracy and perfect recall having high consistency. The architecture of the system is shown in the Figure 1(i). We accept the tcpdump data as input, analyse and process the TCP headers, and then finally predict the intrusions. The four steps involving in our method, namely (i) preprocessing, (ii) initial model creation, (iii) model validation and updating, and (iv) anomaly prediction are briefly described below.



(I)

Fig:1 System architecture (I,II,III)



(II)

Sr.	Description
1	Packet count
2	Avg. packet length
3	Avg. defragment (DF) value
4	Avg. more fragment (MF) value
5	Avg. reserved value
6	Avg. time to live (TTL) value
7	Avg. type of service (TOS) value
8	Avg. checksum value
9	Avg. push (PSH) value
10	Avg. explicit congestion notification echo (ECE) value
11	Avg. congestion window reduced (CWR) value
12	Avg. urgent (URG) value
13	Avg. acknowledgement (ACK) value
14	Avg. reset (RST) value
15	Avg. synchronization (SYN) value
16	Avg. finish (FIN) value
17	Forward segment redundancy (FSR) value
18	Session duration

(III)

D. Preprocessing

From the input tcpdump data, we extract TCP headers of individual packets Then, following the assumptions by Taylor and Alves-Foss [6], we define session, a group of packets having a unique combination of source IP and port number to destination IP and port number. This definition varies from the standard one practiced in network traffic, which is one complete TCP connection beginning with SYN flags from source, response of SYN/ACK flags from destination and ending along with FIN flag from source, FIN/ACK flags from destination. The reason for defining our session is due to the fact that it is very difficult to extract such independent connections as many connections do not complete until more than one FIN flags are sent out. Since it is technically difficult and time consuming to capture such activities, instead, we collect and analyze the aggregated traffic between the unique pairs of senders and receivers. This allows for a quick and real-time analysis of the network traffic. The important attributes in an

individual TCP packet's header include its packet length, DF, MF, and reserved flags, time-to-live (TTL), type of service (TOS), checksum, and all 8 TCP flags. Once a session is defined, the number of packets in it is counted and the average values of the above attributes from the session's packets are computed. In addition, the attribute FSR (forward segment redundancy) is derived as below. FSR shows the percentage of connection control flags in total packets of the session and observed to have high values for abnormal traffics.

FSR = (FIN +SYN +RST) = Packet count (1)

The resultant feature vector containing 18 attributes, hown in Figure 1(iii), represents a session's behavior. Since the values of attributes vary significantly from one attribute to another, we use a natural logarithmic transformation as a form of normalization. In order to avoid negative infinite values caused by logarithm on zeros, we modify transformation function as:normalize(x) = log(1+x).After that, we apply principal component analysis (PCA) technique to reduce the dimensionality of our feature vectors. The first c number of principal components is taken. (c is an empirically determined integer).

E. Initial Model Creation

We split the whole dataset into 3 portions as follows:

1. Training set: contains normal sessions only; used for initial model creation.
2. Validation set: contains both normal and attack sessions; used for model update after removing attack zone from the validation set.
3. Testing set: containing both normal and attack sessions; for prediction The c-dimensional feature vectors epresents normal sessions, , in the training set are divided into clusters by means of k means clustering in order to model the overall behaviors of the normal sessions. The highest average silhouette value is taken among the different values of k are tried. Since the k means algorithm is affected by first initialized center values, prior to deciding with k, we make sure we get the reliable value of k, by executing the clustering process multiple times for each value of k. Then, Euclidean distance is used to measure the distances from k cluster centers to each training data point. From this, new k-dimensional feature vectors, each containing k distance values, are constructed. Then, these new feature vectors in the training set are fed to oneclass SVM, resulting in the initial classier model.

F. Model Validation and Updating

By calculating its Euclidean distances to the k cluster centers obtained before, a similar k-dimensional feature vector is also constructed for each of the normal sessions in the validation set. Then, each validation set's feature vector is sent to the initial classier model for classification in order to validate the initial model. Misclassified (as attacks) are removed from the data pool, as they do not agree with the initial model defined by the normal sessions in the training data. Finally, the original c-dimensional feature vectors of the correctly classified normal session in the validation set (depicted as the gray area in Figure 1(ii) used to build another one-class SVM model. This results in the final classifier model, which reflects both the training and validation sets.

G. Anomaly Prediction

We first need to construct its c-dimensional feature vector, as a new unknown TCP session is to be classified. Then, in order to predict whether it is a normal or an attack session it is sent to the one-class SVM of the final classifier model.

IV. EXPERIMENTAL RESULTS

H. Dataset Used

MIT DARPA '99 dataset [7] used by us which is publicly available. For a long time , it has been one of the well-known benchmarks for IDS. One drawback with DARPA dataset concerns with its simulated nature as pointed out by McHugh [8]. However, our results would be comparable with others when we use a standard and recognized dataset that is why we chose to proceed with this dataset. We use the three weeks of data starting from third, fourth and fifth. The third week of dataset consists of only normal packets and hence is used for training purpose. For validation fourth dataset and for testing (evaluation) purposes fifth week's datasets, containing both normal and attack packets used. The Portsweep, Satan, Neptune, Mailbomb, nmap, etc are different types of TCP attacks included in DARPA '99 dataset. In our experiment, we look at the distributions of the numbers of sessions over the three weeks' data. We similarly define our session, based on the assumption presented by Taylor and Alves-Foss [6]. The number of normal and attack sessions in training, validation, and testing sets.

I. Comparison with NATE

In the NATE method presented by Taylor and Alves-Foss [6], among the 4 selected types of attacks explored, they were able to detect the Portsweep, Satan, and Neptune attacks, but not Mailbomb (ref. Tables 4 and 5 in [6]). That means their recall is less than 100%. In our case, we are able to detect all attacks presented in the 4th week of DARPA '99 dataset with a perfect recall of 100%. Our model overcomes the problem of NATE, which requires absolutely normal packets for normal session modeling. Since we have used the validation set to crosscheck our normal session modeling, our final model reflects both the normal session and the attack session (by means of using the negative space).

V. CONCLUSION

In this research, we successfully model normal sessions and detect some attack sessions obtain from the 99 DARPA dataset. This research provides a very good practical solution for IDS that is based on the outlier mining method for TCP headers. In this era the threats are

increasing very high for the network security and this system will help network security professionals for analyzing and detecting threats in TCP header and perform anomaly detection in real-time.

VI. FUTURE WORK

We are currently using the TCP header information for processing in future we will like to work on derived information from the different connection log analysis and that might help us to further reduce false positive rate of the system. We shall also look forward to compare our system with another system like ALAD/PHADs.

ACKNOWLEDGMENT

We are pleased to express our sentiments of gratitude to all who rendered their valuable guidance to us. We express our appreciation and thanks to the Principal of our college. We are also thankful to the Head of Department and guide Prof. S.M.Sangve. We thank to the reviewers for their valuable comments.

REFERENCES

[1] S. Suthaharan and T. Panchagnula, Relevance features selection with data cleaning for intrusion detection system, Proc. 2012 IEEE SECon, 2012.
[2] X. Zhang, L. Jia, H. Shi, Z. Tang, X. Wang, The application of machine learning methods to intrusion detection, Proc. 2012 S-CET, 2012.
[3] M. Kumar,M. Hanumanthappa, T.V.S. Kumar, Intrusion detection system using decision tree algorithm, Proc. 14th IEEE ICCT, 2012
[4] M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani, A detailed analysis of the KDD CUP 99 data set, Proc. 2nd IEEE CISDA, 2009.
[5] M. V. Mahoney, P. K. Chan, Learning non stationary models of normal network traffic for detecting novel attacks, Proc. 8th ACM KDD, 2002.
[6] C. Taylor, J. Alves-Foss, NATE-network analysis of anomalous traffic events, a low-cost approach, Proc. 2001 NSPW, 2001.
[7] M. V. Mahoney, P. K. Chan, An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection, Proc. 6th RAID, 2003.
[8] J. McHugh, Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory, ACM Transactions on Information System Security 3, 2000.
[9] F. Gharibian, A. Ghorbani, Comparative study of supervised machine learning techniques for intrusion detection, Proc. 5th CNSR, 2007.
[10] H. Sarvari, M. M. Keikha, Improving the accuracy of intrusion detection systems by using the combination of machine learning approaches, Proc. 2010 SoCPaR, 2010.
[11] S. A. Hofmeyr, S. A. Forrest, Architecture for an artificial immune system, Evolutionary Computation 8 , 2000.

BIOGRAPHIES

Nandkishor G. Bhopale, pursuing Master of Engineering Degree from Computer Engineering Department, in Zeal College of Engineering Research, Pune. He has 1completed his B. Tech from Viswakarma Institute of Technology, Pune, Maharashtra, India.

Asst. Prof. S.M.Sangve, working as Assistant Professor and Head of Computer Engineering Department, in Zeal College of engineering Research, Pune. He has total 17 years of experience in teaching. He has published many research papers in National and International conferences and International journals. His area of interest is network security, data mining and software engineering. He is a research scholar from SGGSCOET, Nanded, Maharashtra,India

DOI 10.17148/IJARCCE.2016.55171