Machine Learning Approaches for Detecting Intrusions in Network System

S.V.Shirbhate
Research scholar
Amravati

V.M.Thakare, PhD.
S.G.B.A.U
Amravati

S.S.Sherekar, PhD.
S.G.B.A.U
Amravati

ABSTRACT

As the speedy development of internet services and rising intrusion problem the traditional intrusion detection methods cannot work well for complicated intrusions. From the last decade intrusion detection is the type of security management system for computer and network which are attracted by computational intelligent society. To improve the performance of IDS nowadays many of the researchers are paying attentions towards the machine learning algorithms. It is one of the major concerns in the research of intrusion detection.

This paper focuses on the various approaches for feature selection of intrusion detection system, analyses the various intrusion detection techniques based on machine learning including SVM, outlier mining, Bayesian method and data mining**.**

Keywords

Intrusion detection, SVM, outlier mining, naïve Bayesian and data mining

1. INTRODUCTION

Now a day, as the advances in computer network technology expand for communication, rate of intrusion increases more. Intrusion detection is an area growing in relevance where more sensitive data are stored and process in network system. Intrusion detection is the process of identifying the actives that attempts to compromise the confidential, integrity or availability of computer or network. Intrusion detection system can collects and analyze the information from a variety of system and network sources for signs of intrusions. Intrusion detection system can be host based or network based system. Host based intrusion detection system located in server to examine internal interfaces and network based intrusion detection system monitors the network traffic for detecting intrusions.

Intrusion detection system can be classified according to their detection strategy: misuse or signature based detection and anomaly detection [1]. However, today's commercially available intrusion detection systems are signature base. Signature based intrusion detection system performs pattern matching technique to match an attack pattern correspondingly to known attack pattern in the database. Though this technique have low false positive rate it is enable to detect unknown attack. On other hand anomaly based intrusion detection system builds a normal behavior and automatically detects anomalous behavior.

To improve the performance nowadays the researchers introduce the machine learning algorithm in intrusion detection system. It is one of the major concerns in the research of intrusion detection. So in this paper learn various kinds of intrusion detection technique which are mainly used by researchers such as support vector machine, outlier mining, naïve Bayesian and data mining.

The remainders of the paper are organized as follows. Section 2 presents the overview of IDS based on support vector machine, in section 3 introducing IDS based on outlier mining. In section 4 and 5, IDS based on Naïve Bayesian and data mining respectively are discussed. In Section 6, the analysis of these methods is presented. Finally, our conclusions are mentioned in Section 7.

2. INTRUSION DETECTION SYSTEM BASED ON SUPPORT VECTOR MACHINE (SVM)

Now days in intrusion detection system, the machine learning based approaches such as NN, GA, and HMM is used to build the detection models. However using these machine learning approaches have some limitations such as

1. Tremendous amount of archived audit data is required by machine learning algorithm.
2. The time consuming training process is another problem for tradition machine learning algorithm to get significant performance.

Support Vector Machine (SVM), proposed by Vapnik in 1995, is a Machine Learning Theory based on Statistical Learning Theory. It is become one of the popular tools in IDS, due to its remarkable characteristics such as good generalization performance, the absence of local minima and the sparse representation of solution. By using SVM, the performance of IDS in training time and generalization achieves certain progress [2]. Support vector machine has recently has been introduced as new technique for solving variety of learning, classification and prediction problems. The basic support vector machine is deals with two class problem in which data are separated by number of support vectors. Support vectors are subset of training data used to define the boundary between two classes. In situation where SVM cannot separate two classes, it solves this problem by mapping input data into high dimensional feature spaces using kernel function.

Author in his [3] paper proposed a novel model based on KPCA and SVM. In this method author added a pre-process module before the classifier. Use principal components extracted from the input data using kernel personal component analysis (KPCA), which is the main part of the pre-process module for feature selection. After selecting the features SVM classifier that differentiates the normal and abnormal actions. By extracting features and SVM in intrusion detection system has good performance in both accuracy and generalization.

3. INTRUSION DETECTION SYSTEM BASED ON OUTLIER MINING

Outlier mining is the process of identifying outliers in set of data. In this paper [4], authors propose a new framework of unsupervised anomaly NIDS based on the outlier detection technique using random forests algorithm. Since supervised anomaly NIDS needs attack free training data and such kind of data is difficult to obtain in real world network environment. The framework builds the patterns of network services over datasets labeled by the services. With the built patterns, the framework detects attacks in the datasets using the outlier detection algorithm. Due to large population of datasets used in NIDSs, the process to detect outliers is very time-consuming and costs a large amount of memory. To improve the performance, authors modify the original outlier detection algorithm to reduce the calculation complexity, under the assumption that each network service has its own pattern for normal activities. This approach detects novel intrusion with attack-free training data having high detection rate with low false positive rate.

In this paper [5] authors proposed a new outlier mining algorithm based on index tree, named TreeOut, designed to detect the outliers. Outliers have the weight greater than the threshold. In this method the upper and lower bound of the weight of each record is calculate for r-region and index tree to avoid unnecessary distance calculation. This algorithm is easy to implement, and more suitable to detect intrusions in the audit data.

The outlier detection technique is effective to reduce false positive rate with desirable detection rate.

4. INTRUSION DETECTION SYSTEM BASED ON BAYESIAN CLASSIFICATION THEREOM

Naïve Bayesian classifier is one of the most popular data mining algorithms for classification, which provides an optimal way to predict the class of an unknown example. In this paper [6], author proposed a new learning algorithm

for mining network logs to detect network intrusions through naïve Bayesian classifier, which clusters the network logs into several groups based on similarity, and then calculates the prior and conditional probabilities for each group of logs. For classifying a new log, the algorithm checks in which cluster the log belongs and then use that cluster's probability set to classify the log. It improves detection rate and reduced the false positive for different type of network attack.

5. INTRUSION DETECTION SYSTEM BASED ON DATA MINING

Using data mining into intrusion detection system improve the performance has become one of the major concerns in the research of intrusion detection. Data mining generally refers to the process of extracting descriptive models from large stores of data. The recent rapid development in data mining has made available a variety of algorithm, drawn from the various fields. Nowadays, data mining methods have become indispensable tools for analyzing large volume of network logs or audit data to identify the patterns of the normal behaviors and pattern of the intrusions in computer network that are useful in classifying network intrusions. The main motivation of using data mining methods in intrusion detection is automation. Data mining technologies have been widely used to analyze network logs to gain intrusion related knowledge to improve the performance of IDS in last decades. To apply data mining techniques in intrusion detection, first the collected network logs or audit data needs to be preprocessed and converted to the format that suitable for mining. Next, the reformatted data will be used to develop a clustering or classification model. Data mining provide decision support for intrusion management, and also help IDS for detecting new vulnerabilities and intrusions by discovering unknown patterns of attacks or intrusions.

In this paper [7], authors propose a new data-mining based technique for intrusion detection using an ensemble of binary classifiers with feature selection and multiboosting simultaneously. This model employs feature selection so that the binary classifier for each type of attack can be more accurate, which improves the detection of attacks that occur less frequently in the training data. Based on the accurate binary classifiers, model applies a new ensemble approach which aggregates each binary classifier's decisions for the same input and decides which class is most suitable for a given input. During this process, the potential bias of certain binary classifier could be alleviated by other binary classifiers' decision. Also this model makes use of multiboosting for reducing both variance and bias.

In this paper [8] author describe a new data mining based method for intrusion detection based on network connection features. This method attempts to separate different kinds of intrusions from normal activities by using Improved Iterative Scaling (IIS). Improved Iterative Scaling (IIS) is invented by members of the machine translation group at IBM's T.J. Watson Research Center. It is a hillclimbing algorithm for computing maximum likelihood estimates of the parameters of Conditional Exponential Model (CEM) Also in this method Chi-squared is used for selecting relevant connection features to improve the performance.

6. ANALYSIS AND DISCUSSION

The various methods of IDS based on machine learning are studied and analyze in following table

S.no	Authors name	Technique for feature selection	Technique for detection	Advantages	Performance measured
1	Yuan-Cheng Li, Zhong-Qiang Wang	KPCA(kernel principal component analysis)	LS-SVM(Least Square Support Vector Machine)	Using SVM the performance of IDS in training time and generalization achieved certain progress	Accuracy and generalization
2.	Jiong Zhang and Mohammad Zulkernine	Random forest algorithm	Random forests algorithm(outlier mining technique)	It detects novel intrusion with attack-free training data having high detection rate with low false positive rate.	Detection rate and false positive rate
3.	Nannan Wu, Liang Shi, Qingshan Jiang, and Fangfei Weng	-----	TreeOut algorithm (outlier mining algorithm)	It is unsupervised algorithm. So no training sets are needed	Detection rate and false positive rate
4.	Dewan Md. Farid, Nouria Harbi, Suman Ahmmed, Md. Zahidur Rahman, and Chowdhury Mofizur Rahman	Not specified	Naïve Bayesian classification	It improves detection rate and reduced the false positive for different type of network attack.	Detection and false positive rate
5.	Christine Dartigue, Hyun Ik Jang, and Wenjun Zeng	Bayesian network & classification and regression trees(CART)	Ensemble of binary classification	I. Feature selection improved overall accuracy. Reduced the no of false positives and improved the detection of intrusions with low frequency in training data. I. By using multiboosting variance and bias are reduced.	Accuracy, detection rate and cost
6.	Xin Jin, Ronghuai Huang, Rongfang Bie	Chi square method	Improved iterative scaling (data mining)	Feature selection makes the intrusion detection engine more concise & faster .	Detection rate and false positive rate

Yuan-Cheng et al. apply KPCA for feature extraction which maps the high dimension features in the input space to low dimension eigenspace and then use LS-SVM classifier to estimate whether intrusion occurred or not. By applying feature extraction and machine learning (LS-SVM) in IDS gives the good performance in accuracy and generalization. Due to the feature extraction training time and testing time decreased. Jiong Zhang et al. apply random forest algorithm for feature selection. Then apply outlier detection technique provided by random forests algorithm in anomaly intrusion detection. By applying this approach it detects novel intrusion with attack-free training data having high detection rate with low false positive rate. Nannan Wu et al applied unsupervised TreeOut outlier mining algorithm to detect intrusion. Comparative study shows that this algorithm can have higher detection rate and lower false positive rate. Dewan Md. Farid et al applies Naïve Bayesian classification to improves detection rate and reduced the false positive for different type of network attack and explain the importance of feature selection. Christine Dartigue et al used Bayesian network & classification and regression trees (CART) for feature selection and ensemble of binary classification for detecting intrusions. Feature selection improved overall accuracy. Reduced the number of false positives and improved the detection of intrusions with low frequency in training data. Xin Jin et al applies chi squared method for feature selection

and Improved iterative scaling for intrusion detection. Due to feature selection the intrusion detection engine becomes more concise & faster.

7. CONCLUSION

Intrusion Detection System can help for detecting new vulnerabilities and intrusions by discovering unknown patterns of attacks or intrusions. In this paper various approaches for detecting intrusions are studied. It is found that to apply any techniques in intrusion detection, first the collected network logs or audit data needs to be preprocessed and converted to the format that are suitable for classifying. Next, the reformatted data will be used to develop a clustering or classification model. So in intrusion detection system feature selection is considered an important asset in building classification or clustering models. Elimination of useless features enhances the accuracy of detection and speeding up computation. Thus feature selection improves the overall performance of detection mechanism.

REFERENCES

[1] S. V. Shirbhate, Dr V. M. Thakare, Dr S. S. Sherekar**, "**Data Mining Approaches For Network Intrusion Detection System", International Journal of Computer Technology and Electronics Engineering (IJCTEE), ISSN 2249-6343,PP. 41-44,2011.

[2] Liu Hui, CAO Yonghui, " Research Intrusion Detection Techniques From the Prespective Of Machine Learning", second International Conference On Multimedia and Information Technology, 978-0-7695-4008-5/10, IEEE Computer Society, PP.166-168,2010.

[3] Yuan-Cheng Li, Zhong-Qiang Wang, "An Intrusion Detection Method Based On SVM And KPCA" Proceedings of the 2007 International Conference on Wavelet Analysis and Pattern Recognition, Beijing, China, 1-4244-1066-5/077 IEEE ,PP.1462-1466, 2007.

[4] Jiong Zhang and Mohammad Zulkernine, "Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection", Proceeding of IEEE workshop on Information Assurance United status Miltary Acadamy, PP.1-6,2005.

[5] Nannan Wu, Liang Shi, Qingshan Jiang, and Fangfei Weng, " An Outlier Mining-Based method for anomaly Detection", 1-4244-1035-5/07IEEE,PP.152-156, 2007.

[6] Dewan Md. Farid, Nouria Harbi, Suman Ahmmed, Md. Zahidur Rahman, and Chowdhury Mofizur Rahman, "Mining Network Data for Intrusion Detection through Naïve Bayesian with Clustering", proceeding in World Academy of Science, Engineering and Technology,PP.341-345, 201

[7] Christine Dartigue, Hyun Ik Jang, and Wenjun Zeng, "A New Data-Mining Based Approach for Network Intrusion Detection", Seventh Annual Communications Networks and Services Research Conference, 978-0-7695-3649-1/09 DOI 10.1109/CNSR, IEEE computer Society,PP.372-377,2009.

[8] Xin Jin, Ronghuai Huang, Rongfang Bie, "Detecting Network Attacks via Improved Iterative Scaling", 1-4244-0865-2/07, IEEE, PP.113-118, 2007.