Simulation of Distributed Intrusion Detection Based on Ensemble of Classifier

A. Anbumozhi
PG Scholar
Department Of Computer Science and Engineering
Mepco Schlenk Engineering College,
Sivakasi, India
aanbumozhi.08@gmail.com

Dr. K. Muneeswaran
Professor
Department Of Computer Science and Engineering
Mepco Schlenk Engineering College
Sivakasi, India
kmuni@mepco.ac.in

Abstract **Detecting Intruders around Networks plays an important role. Security in network aggregation is not an easy task. Network consists of nodes whose operation can be controlled by underlying network. In this paper, a traditional online Adaboost process is used where decision stumps are used as weak classifiers. In the second algorithm, an improved online Adaboost process is proposed, and online Gaussian mixture models (GMMs) are used as weak classifiers. In the second algorithm, an improved online Adaboost process is proposed where online Gaussian mixture models (GMMs) are used as weak classifiers. An algorithm based on particle swarm optimization (PSO) and support vector machines (SVM) is proposed. PSO and SVM based algorithm effectively combines the local detection models into the global model in each node, the global node in a node can handle the intrusion types. Both the algorithms outperform existing intrusion detection algorithms. It is also shown that our PSO, and SVM based algorithm effectively combines the local detection models into the global model in each node, the global model in a node can handle the intrusion types that are found in other nodes, without sharing the samples of these intrusion types.**

***Index Terms*—Dynamic distributed detection, network intrusions, online Adaboost learning, parameterized model.**

I. INTRODUCTION

Network intrusion detection aims at distinguishing the attacks on the Internet. Network attack detection is one of the most important problems in network information security. Network Intrusion Detection System (NIDS) / Network Intrusion Prevention System (NIPS) and Unified threat management devices is proposed to detect attack in the network. Current network intrusion detection systems (IDS) lack adaptability to the frequently changing network environment. Intrusion Detection System (IDS) focuses on machine learning based Network Intrusion Detection System (NIDS). Machine learning based intrusion detection methods can be classified as statistics based, data mining based, and classification based detection method. Network environments and the intrusion training data change rapidly over time. Most existing algorithms for training intrusion detectors are offline.

The KDD Cup 99 dataset has been the point of attraction for many researchers in the field of intrusion detection from the last decade. Many researchers have contributed their efforts to analyze the dataset by different techniques. Analysis can be used in any type of industry that produces and consumes data,

of course that includes security. This paper analysis 10% of KDD cup'99 training dataset based on intrusion detection, focused on establishing a relationship between the attack types and the protocol used by the hackers. Analysis of data have used the Oracle 10g data miner as a tool for the analysis of

dataset and build 1000 clusters to segment the 494,020 records. The investigation revealed many interesting results about the protocols and attack types preferred by the hackers for intruding the networks. The protocols that are considered in KDD dataset are
TCP, UDP and ICMP that are explained below: and attack types preferred by the hackers for intruding the networks.
TCP: TCP stands for "Transmission Control Protocol". TCP is an important protocol of the Internet Protocol Suite at the Transport Layer which is the fourth layer of the OSI model. It is a reliable connection oriented protocol which implies that data sent from one side is sure to reach the destination in the same order. TCP splits the data into labeled packets and sends them across the network. TCP is used for many protocols such as HTTP and Email Transfer.

UDP: UDP stands for "User Datagram Protocol". It is similar in behavior to TCP except that it is unreliable and connection less protocol. As the data travels over unreliable media, the data may not reach in the same order, packets may be missing and duplication of packets is possible. This protocol is a transaction oriented protocol which is useful in situations where delivery of data in certain time is more important than loosing few packets over the network. It is useful in situations where error checking and correction is possible in application level.

ICMP: ICMP stands for "Internet Control Message Protocol". ICMP is basically used for communication between two connected computers. The main purpose of ICMP is to send messages over networked computers. The ICMP redirect the messages and it is used by routers to provide the up to date routing information to hosts, which initially have minimal routing information. When a host receives an ICMP redirect

message, it will modify its routing table according to the message.

The current realistic solutions for NIDS used in industry are misuse based methods that make use of signatures of attacks to detect intrusions by modelling each type of attack. As typical misuse detection methods, pattern matching methods search packages for the attack features by utilizing protocol rules and string matching. Pattern matching methods can effectively detect the well known intrusions. But they rely on the suitable generation of attack signatures, and fail to detect novel and unknown attacks. In the case of rapid proliferation of novel and unknown attacks, any defence based on signatures of known attacks becomes impossible. Moreover, the increasing diversity of attacks obstructs modeling signatures. Adaboost based classifiers are generally encouraging. In our framework, a hybrid of online weak classifiers and an online Adaboost process results in a parameterized local model at each node for intrusion detection. The parametric models for all the nodes are combined into a global intrusion detector in each node using a small number of samples, and the combination is achieved using an algorithm based on particle swarm optimization (PSO) and support vector machine (SVMs). The computation complexity for constructing the decision stumps is very low, and online updating of decision stumps can be easily implemented when new training samples are obtained**.**

II. RELATED WORK

Classification based methods construct a classifier that is used to classify new connections as either attacks or normal connections. For instance, Mukkamala *et al*. [30] use the support vector machine (SVM) to distinguish between normal network behaviors and attacks, and further identify important features for intrusion detection. Mill and Inoue [31] propose the TreeSVM and ArraySVM algorithms for reducing the inefficiencies that arise when a sequential minimal optimization algorithm for intrusion detection is learnt from a large set of training data. Zhang and Shen [7] use SVMs to implement online intrusion detection. Kayacik *et al*. [5] propose an algorithm for intrusion detection based on the Kohonen self organizing feature map (SOM). Specific attention is given to a direct labeling of SOM nodes with the connection type. Bivens *et al*. [26] propose an intrusion detection method, in which SOMs are used for data clustering and multilayer perceptron (MLP) neural networks are used for detection. Hierarchical neural networks [28], evolutionary neural networks [29], and MLP neural networks [27] have been applied to distinguish between attacks and normal network behaviors. Hu and Heywood [6] combine SVM with SOM to detect network intrusions. Khor *et al*. [55] propose a dichotomization algorithm. As the number of training samples increases, the accuracy of the online ensemble classifier gradually increases until it approximates to the accuracy of the offline ensemble classifier [43].

III. LOCAL DETECTION MODEL

The classical Adaboost algorithm [37] carries out the training task in batch mode. A number of weak classifiers are constructed using a training set. Weights, which indicate the importance of the training samples, are derived from the classification errors of the weak classifiers.

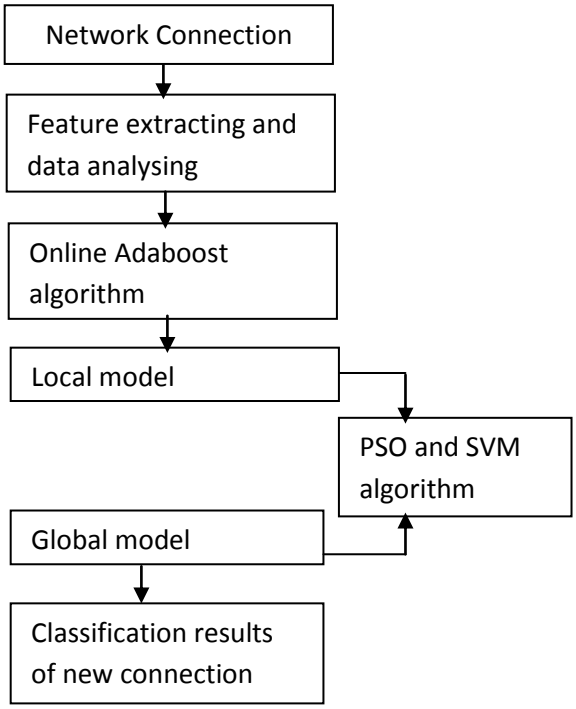

Fig. 1.1 System Design Modules

The final strong classifier is an ensemble of weak classifiers. The classification error of the final strong classifier converges to 0.

However, the Adaboost algorithm based on offline learning is not suitable for networks. We apply online versions of Adaboost to construct the local intrusion detection models. It is proved in [38] that the strong classifier obtained by the online Adaboost converges to the strong classifier obtained by the offline Adaboost as the number of training samples increase. Machine learning deals with automatically inferring and generalizing dependencies from data to allow extrapolation of dependencies to unseen data. Machine learning methods for intrusion detection model both attack data and normal network data, and allow for detection of unknown attacks using the network features [60]. This paper focuses on machine learning based NIDS. The machine learning based intrusion detection methods can be classified as statistics based, data mining based, and classification based. All the three classes of methods first extract low level features and then learn rules or models that are used to detect intrusions. A brief review of each class of methods is given below,

1) Statistics based methods construct statistical models of network connections to determine whether a new connection is an attack. For instance, Denning [1] construct statistical profiles for normal behaviors. The profiles are used to detect anomalous behaviors that are treated as attacks. Caberera et al. [2] test to compare observation network signals with normal behavior signals, assuming that the number of observed events in a time segment obeys the Poisson distribution. Li and Manikopoulos [22] extract several representative parameters of network flows, and model these parameters using a hyperbolic

distribution. Peng et al. [23] use a nonparametric cumulative sum algorithm to analyze the statistics of network data, and further detect anomalies on the network.

2) Data mining based methods mine rules that are used to determine whether a new connection is an attack. For instance, Lee *et al*. [3] characterize normal network behaviors using association rules and frequent episode rules [24]. Deviations from these rules indicate intrusions on the network. Zhang *et al*. [40] use the random forest algorithm to automatically build patterns of attacks. Otey *et al*. [4] propose an algorithm for mining frequent item sets (groups of attribute value pairs) to combine categorical and continuous attributes of data. The algorithm is extended to handle dynamic and streaming datasets. Zanero and Savaresi [25] first use unsupervised clustering to reduce the network packet payload to a tractable size, and then a traditional anomaly detection algorithm is applied to intrusion detection. Mabu et al. [49] detect intrusions by mining fuzzy class association rules using genetic network programming. Panigrahi and Sural [51] detect intrusions using fuzzy logic, which combines evidence from a user's current and past behaviors.

3) Classification based methods construct a classifier that is used to classify new connections as either attacks or normal connections. For instance, Mukkamala et al. [30] use the support vector machine (SVM) to distinguish between normal network behaviors and attacks, and further identify important features for intrusion detection.

IV. OVERVIEW OF THE FRAMEWORK

In the distributed intrusion detection framework, each node separately constructs its own local intrusion detection model according to its own data. By combining all the local models, at each node, a global model is trained using a small number of the samples in the node, without sharing any of the original training data between nodes. The global model is used to detect the intrusions at the node. The global model in a node can handle the attack types that are found in other nodes, without sharing the samples of these attack types. Our framework is original in the following ways,

1) In the Adaboost classifier, the weak classifiers are constructed for each individual feature component, for both continuous and categorical ones, in such a way that the relations between these features can be naturally handled, without any forced conversions between continuous features and categorical features.
2) New algorithms are designed for local intrusion detection. The traditional online Adaboost process and a newly proposed online Adaboost process are applied to construct local intrusion detectors. The weak classifiers used by the traditional Adaboost process are decision stumps. The new Adaboost process uses online Gaussian mixture models (GMM) as weak classifiers. In both cases the local intrusion detectors can be updated online. The parameters in the weak classifiers and the strong classifier construct a parametric local model.
3) The local parametric models for intrusion detection are shared between the nodes of the network. The volume of communications is very small and it is not necessary to share the private raw data from which the local models are learnt.
4) We propose a PSO and SVM based algorithm for combining the local models into a global detector in each node. The global detector that obtains information from other nodes obtains more accurate detection results than the local detector.

V. IMPLEMENTATION AND RESULTS

For Adaboost based learning algorithms, the detection rate and the false alarm rate depend on the initial weights of the training samples. So we propose to adjust the initial sample weights in order to balance the detection rate and the false alarm rate. Although there is much work on intrusion detection, several issues are still open and require further research, especially in the following areas.

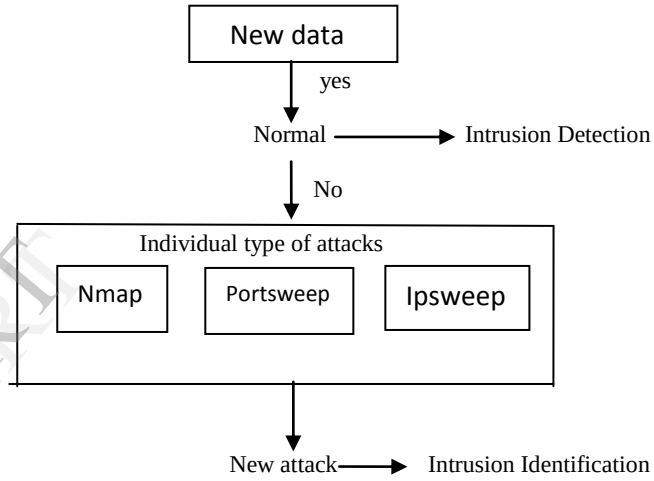

Fig 1.2 Implementation Flow Chart

Network environments and the intrusion training data change rapidly over time, as new types of attack emerge. In addition, the size of the training data increases over time and can become very large. Most existing algorithms for training intrusion detectors are offline. The intrusion detector must be retrained periodically in batch mode in order to keep up with the changes in the network. This retraining is time consuming. Online training is more suitable for dynamic intrusion detectors. New data are used to update the detector and are then discarded. The key issue in online training is to maintain the accuracy of the intrusion detector.

1.Dataset Preprocessing: This dataset has around 2 million connection records. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labelled as either normal or an attack. It is a predictive model capable of distinguishing between "bad" connections, called intrusions or attacks, and "good" normal connections. Attacks fall into four main categories:

- DOS : Denial of service.
- R2L : Remote to Local attack.

- U2R User to Remote attack.
- Probing.

VI. MODULE DESCRIPTION

1.Dataset Loading.

2.Local Model.

3.Global Model.

Module1: Dataset Loading

Table 1.1 Knowledge Discovery Dataset Attributes

Feature Name
Duration
Potocol Type
Service
Src_byte
Dst_byte
Flag
Land
Wrong_Fragment
Urgent
Hot
Num_failed_login
Logged_in
Num_compromised
Root_shell
Su_attempted
Num_root
Num__file_creations
Num_shells
Num_access_files
Num_outbound_cmds
Is_hot_login
Is_guest_login
Count
serror_rate
rerror_rate
same_srv_rate
diff_srv_rate
srv_count
srv_serror_rate
srv_rerror_rate
srv_diff_host_rate

Adaboost algorithm is one of the most popular machine learning algorithm. Uses Multiple iterations to generate a single composite strong classifier. It corrects the misclassification made by weak classifier. This algorithm results in an parameterized local model at each node for intrusion detection. Parametric model for all nodes are combined into a global model for detection purpose. Global detector in each node are detected using small number of samples. In the Adaboost classifier, the weak classifier are constructed for each individual feature component. New algorithm are designed for local intrusion detection.

- ✓ Traditional online Adaboost process are applied to construct local intrusion detectors.
- ✓ Newly proposed online Adaboost process are applied to construct local intrusion detectors. Local parametric model for intrusion detection are shared between the nodes of the network.

Module 2 : Local Model

1. Distributed Node Creation.

2. Decision Stumps.

3. Online GMM

1. Distributed Node Creation : Adaboost splits the nodes into various request. In this paper , 5 node has 60 request.

2. Decision Stump : A decision stump is constructed for each component of the network connection data. Decision stump decides how much request in the dataset. KDD CUP 99 has been most widely used in attacks on network. The simulated attack falls in one of the following four categories [9].The request are as follows

Total request in the dataset – 430

HTP request – 93

FTP request – 3

SMTP request – 2

Attack request – 332

Node 1`	HTTP Request	FTP Request	SMTP Request	Attack Request
	2	0	0	58
Node 2	HTTP Request	FTP Request	SMTP Request	Attack Request
	36	0	0	24
Node 3	HTTP Request	FTP Request	SMTP Request	Attack Request
	22	0	0	38
Node 4	HTTP Request	FTP Request	SMTP Request	Attack Request
	12	0	0	48
Node 5	HTTP Request	FTP Request	SMTP Request	Attack Request
	12	0	0	58

3. Online GMM : differentiates by each attack in the dataset.

1. Denial of Service Attack (DOS): In this category the attacker makes some computing or memory resources too busy or too full to handle legitimate request, or deny legitimate users access to machine. DOS contains the attacks: 'neptune', 'back', 'smurf', 'pod', 'land', and 'teardrop'.
2. Users to Root Attack (U2R): In this category the attacker starts out with access to a normal user account on the system and is able to exploit some vulnerability to obtain root access to the system. U2R contains the attacks: 'buffer_overflow', 'load module', 'root kit' and 'perl'
3. Remote to Local Attack (R2L): In this category the attacker sends packets to machine over a network but who does not have an account on that machine and exploits some vulnerability to gain local access as a user of that machine. R2L contain the attacks: 'warezclient', multihop',' ftp_write', 'imap', 'guess_passwd', 'warezmaster', 'spy' and 'phf' .
4. Probing Attack (PROBE): In this category the attacker attempt to gather information about network of computers for the apparent purpose of circumventing its security. PROBE contains the attacks: 'portsweep', 'satan', 'nmap', and 'ipsweep'
5. The major objectives performed by detecting network intrusion are stated as recognizing rare attack types such as U2R and R2L, increasing the accuracy detection rate for suspicious activity, and improving the efficiency of real time intrusion detection models. This detects that the training dataset consisted of 494,019 records, among which 97,277 (19.69%) were 'normal', 391,458(79.24%) DOS, 4,107 (0.83%) Probe, 1,126 (0.23%) R2L and 52 (0.01%) U2R attacks. Each record has 41 attributes describing different features and a label assigned to each either as an 'attack' type or as 'normal'.

VII. RESULTS AND DISCUSSION

The construction of a local detection model at each node includes the design of weak classifiers and Adaboost based training. Each individual feature component corresponds to a weak classifier. In this way, the mixed attribute data for the network connections can be handled naturally, and full use can be made of the information in each feature. The Adaboost training is implemented using only the local training samples at each node. After training, each node contains a parametric model that consists of the parameters of the weak classifiers and the ensemble weight.

The attacks are :

1. Smurf attack DOS attack uses ICMP protocol, in which it creates more network traffic.

2. Neptune attack DOS attack uses TCP protocol. In which implementation becomes attacked.

3. Back attack DOS attack uses TCP protocol, in which slows down all the activities and will get recovered after the attack stops.

4. Ipsweep attack Probe attack uses TCP protocol, sends packets to various hopes.

GMM identifies by placing specific attacks.

Smurf Attack Protocol	: ICMP/n
Neptune Attack Protocol	: TCP
Back Attack Protocol	: TCP
Ipsweep Attack Protocol	: TCP
Nmap Attack Protocol	: UDP

VIII. CONCLUSION

In this paper, online Adaboost based intrusion detection algorithms is proposed, in which decision stumps and online GMMs were used as weak classifiers. The results of the algorithm using decision stumps and the traditional online Adaboost were compared with the results of the algorithm using online GMMs and our online Adaboost. We further proposed a distributed intrusion detection framework, in which the parameters in the online Adaboost algorithm formed the local detection model for each node, and local models were combined into a global detection model in each node using a PSO and SVM based algorithm is proposed.

ACKNOWLEDGEMENT

The authors wish to express their sincere thanks to the department of computer science and engineering of Mepco Schlenk College, Sivakasi for providing valuable guidelines, good support and encouragement during this work. They are also thankful to the management and principal for their constant support and encouragement to carry out this part of the project work successfully.

REFERENCES

[1]D. Denning, “An intrusion detection model,” IEEE Trans. Softw. Eng., vol. SE 13, no. 2, pp. 222–232, Feb. 1987.
[2]J. B. D. Caberera, B. Ravichandran, and R. K. Mehra, “Statistical traffic modeling for network intrusion detection,” in Proc. Modeling, Anal. Simul. Comput. Telecommun. Syst., 2000, pp. 466–473.
[3] W. Lee, S. J. Stolfo, and K. Mork, “A data mining framework for building intrusion detection models,” in Proc. IEEE Symp. Security Privacy, May 1999, pp. 120–132.
[4] M. E. Otey, A. Ghoting, and S. Parthasarathy, “Fast distributed outlier detection in mixed attribute data sets,” Data Ming Knowl. Discovery, vol. 12, no. 2–3, pp. 203–228, May 2006.
[5]H. G. Kayacik, A. N. Zincir heywood, and M. T. Heywood, “On the capability of an SOM based intrusion detection system,” in Proc. Int. Joint Conf. Neural Netwo., vol. 3. Jul. 2003, pp. 1808–1813.
[6] P. Z. Hu and M. I. Heywood, “Predicting intrusions with local linear model,” in Proc. Int. Joint Conf. Neural Netw., vol. 3, pp. 1780–1785, Jul. 2003.
[7] Z. Zhang and H. Shen, “Online training of SVMs for real time intrusion detection,” in Proc. Adv. Inform. Netw. Appl., vol. 2, 2004, pp. 568–573.
[8]H. Lee, Y. Chung, and D. Park, “An adaptive intrusion detection algorithm based on clustering and kernel method,” in Proc. Int. Conf. Adv. Inform. Networking Appl., 2004, pp. 603–610.

[9]W. Lee and S. J. Stolfo, "A framework for constructing features and models for intrusion detection systems," ACM Trans. Inform. Syst. Security, vol. 3, no. 4, pp. 227–261, Nov. 2000.
[10]A. Fern and R. Givan, "Online ensemble learning: An empirical study," in Proc. Int. Conf. Mach. Learning, 2000, pp. 279–286.