Analysis of Hybrid Intrusion Detection System Based on Data Mining Techniques

Prathibha K S [1], Pankaj Kumar[2] Shyni T S [3]
[1]Post Graduate Student, [2,3]Assistant Professor, [2]FISAT, Angamaly, [1,3]School of Computer Sciences, M G University

***Abstract*—The rapid growth of network based activities makes computer security is a more crucial issue. Many security methods are developed and used, but they are unfit to detect novel intrusions. Therefore, we propose a hybrid intrusion detection framework based on data mining classification and clustering techniques. In the proposed hybrid framework, improves the detection rate by taking the advantages of misuse and anomaly detection. In case of misuse detection, intrusion patterns are built automatically from a training data by the use of the random forest classification method. Then comparing this pattern against network activities to detect intrusions. In case of anomaly detection, the network activities processed to several clusters using weighted k means technique to detect novel intrusions. The whole process is evaluated over KDD'99 dataset.**

Keywords —Network security, Intrusion detection, Data mining, Random Forest and Weighted K-Means.

I. INTRODUCTION

Computer network plays an important role in our day to day life to exchange sensitive information between computer devices. For example business strategy document, voice conversation, etc. Based on National Institute of Mental Health & sciences (NIMHANS) study, 73% youngsters are addicted to the internet. Although the network crimes [1] are rapidly increasing that means current security methods like encryption, authentication, access control etc. are unrealistic. Therefore, in order to improve the detection rate of intrusion detection system by the use of data mining techniques.

Intrusion Detection System (IDS) [2], analyzing normal and malicious network access by monitoring the network traffic. And the system will raise an alarm, if any kind of unauthorized access will find. Intrusion detection system is a process used as a countermeasure for monitoring the network and analyzing them to sign or intrusion. The system raises the alarm, if intrusions occur. Intrusion detection techniques are classified into two types: misuse and anomaly detection. The misuse detection system finds known attack based on extensive knowledge of sign or patterns. Pattern matching, state transition analysis is some methods based on misuse detection. Anomaly detection finds a significant deviation from normal system behavior. Statistical methods, expert system is some methods based on anomaly detection. The performance of intrusion detection system depends on its detection rate and false positive rate. The drawback of misuse detection is (i) it cannot detect novel attack (ii) bad detection rate. The drawback of anomaly detection is very bad false positive rate. So the hybrid intrusion detection framework is proposed to improve the detection rate and false positive rate by strengthening the advantages of both misuse and anomaly detection.

Data mining [3] is the process of discovering useful patterns from a large quantity of the dataset. Data mining also called data or knowledge discovery. Data mining is a powerful tool for analyzing and summarizing patterns from many different perspectives. Data mining includes (1) Extract, renovate and load transaction data (2) store and manage the data (3) provide data access (4) analyze the data (5) Present the data in understandable structure. In recent years data mining is widely used in different fields such as electrical power engineering, educational research, biomedical, and human genetic etc. This paper proposes new systematic framework that applies to two important data mining techniques called a random forest classification algorithm and weighted k-means algorithm in the hybrid intrusion detection system. The random forest algorithm is an ensemble approach for classification and regression. The weighted k-means algorithm is an unsupervised learning approach for clustering large data's to specific number of clusters. It's a simple and an iterative method for clustering.

The current intrusion detection system has many challenges. One important challenge is feature selection. Because the number of feature selection highly affects the effectiveness of the classification process. Feature selection determines the most relevant feature of the data. Intrusion detection systems mainly focus on two feature selection measures. They are correlation feature selection (CF) and minimal redundancy maximal relevance (mRMR). Another challenge of intrusion detection system is an imbalance between real and trained data. Misuse detection has a key advantage is their high rate of accuracy in detecting known attacks. Their main drawback is the inability to detect novel attacks. Anomaly detection, built profiles based on normal behavior. Anomaly detection, detect novel attacks (new types of intrusions). Hybrid intrusion detection system, combine the advantages of misuse and anomaly detection. Therefore, the hybrid detection system achieves a high degree of performance and can detect novel attacks by the use of a supervised method.

We present the experimental result on Defense Advanced Research Project Agency (DARPA) dataset, which are developed by the ACM special interest group on knowledge discovery and data mining 1999 (KDD'99) contest. The KDD dataset consist of 41 attributes and around 5 lacks data that is 10% original data set.

II. RELATED WORK

Many intrusion detection systems, a set of rules are used to describe intrusions. The detection techniques are applied in misuse and anomaly detection. The main drawback of these two techniques resides in the encoded models, which defines normal and malicious activities. In [4], is an open source network security tool for intrusion detection in the campus network environment. Two data mining techniques that are random forest and k-means algorithms are used in misuse, anomaly and hybrid detection [5]. A random forest algorithm is used in misuse, anomaly and hybrid detection [6]. The overview of research in building rare class prediction models for identifying known intrusions and their variations and anomaly outlier detection schemes for detecting novel attacks whose nature is unknown [7]. The design and experiences with the ADAM (Audit Data Analysis and Mining) system, which used as a tested test bed to study how useful data mining techniques can be in intrusion detection [8].

Manikandan *R, Oviya P, Hemalatha C* [9] proposed a new ensemble boosted decision tree approach for intrusion detection system. Intrusion detection system evaluations are designed to focus research efforts on core technical issues and provide unbiased measurement of current performance levels [10]. The goal of an intrusion detection system (IDS) is to provide another layer of defense against malicious uses of computer systems by sensing a misuse or a breach of a security policy and alerting operators to an ongoing attack [11].

Data mining can automate the process of finding relationships and patterns in raw data, and can deliver results that can be either utilized in an automated decision support system or assessed by a human analyst thus enhancing the quality of the intrusion detection process [12]. Analysis of 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, using clustered data [13]. Mining Audit Data for Automated Models (MADAM) [14] for intrusion Detection uses a data mining algorithm to compute activity patterns from system audit data and extracts predictive features from the patterns. It then applies machine learning algorithms to the audit records that are processed according to the feature definitions to generate intrusion detection rules.

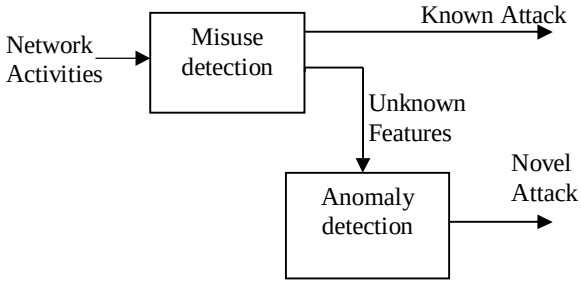

Fig. 1. Overview of Hybrid Intrusion Detection

III. HYBRID INTRUSION DETECTION

The Hybrid Intrusion Detection System is a combination of Misuse and anomaly detection. The main drawback of anomaly detection is very bad false positive rate (12.6%). In contrast to the misuse detection model, achieves bad detection rate (92.73%). Therefore, the proposed Hybrid Intrusion Detection System, combining the misuse and anomaly detection to resolve the drawbacks of both misuse and anomaly detection by taking the advantages of misuse and anomaly detection. Overview of hybrid detection approach (see Fig. 1) has two phases: an online phase and offline phase. In online phase misuse detection detects known intrusion. The unknown features are sent to anomaly detection component. In offline phase, anomaly detection detects novel attacks.

The hybrid detection system has several key advantages. Because it's taking the advantages of misuse and anomaly detection. In misuse detection, random forest classification algorithm is used to achieve high speed. Accuracy of misuse detection is also high. The performance level of anomaly detection tends to be reduced, due to a large number of connections. In order to overcome this problem known attacks are detected in the misuse detection component. Thus, the number of attacks can be reduced significantly. Proposed hybrid detection system, as shown in Fig. 2. Misuse detection work in online phase and anomaly detection work in offline phase. In the online phase, network traffic is captured using network sensors. After some preprocessing the captured packets are stored in a data storage unit. Misuse detection component detects the known attack by comparing the network traffic with intrusion patterns generated previously in the offline phase. If any attack is detected, misuse detection will generate an alarm. If attack features do not match with any attack, it is considered as uncertain data, and it will be stored in anomaly database. A Random forest classification algorithm is used in misuse detection part for classifying intrusion patterns.

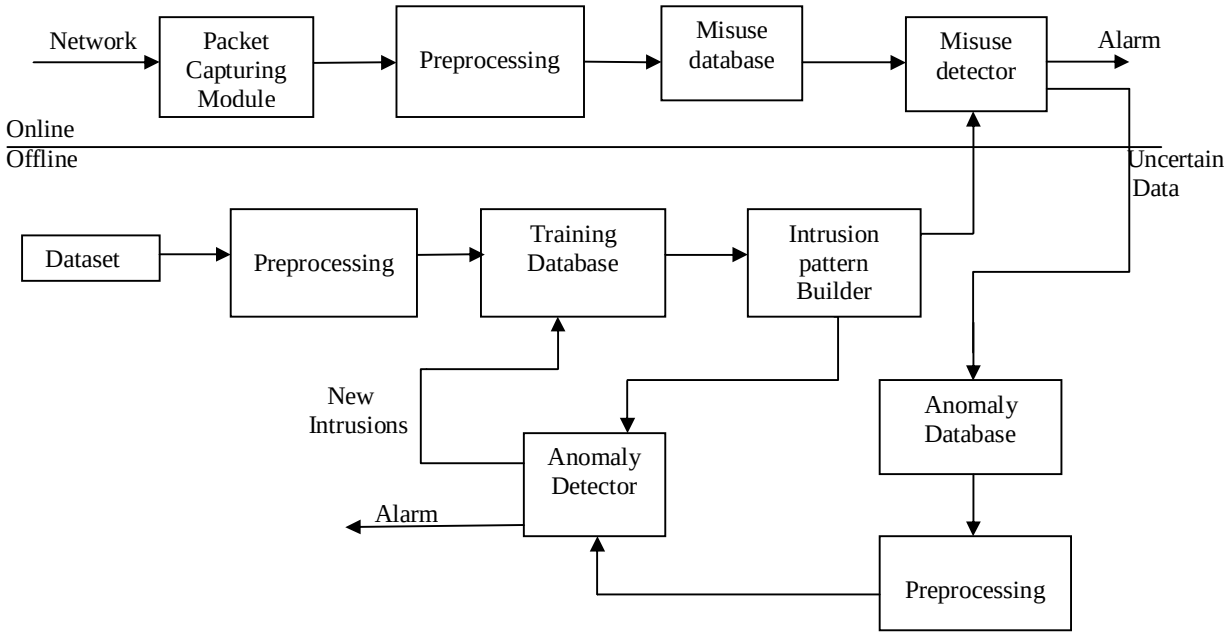

Fig. 2. Block diagram of Hybrid Detection System

A. Random Forest Algorithm

A random forest algorithm is a data mining, classification algorithm. It is an ensemble learning method for classification and regression. The random forest was developed by Leo Breiman and Adele Cutler. Random Forest grows many classification trees. Random forest algorithm as follows:

1. If the number of training set is N, and the number of variable in the classifier be M.
2. The number of input variables m is selected randomly for each node of the tree. M should be much less than N.
3. Calculate the best split of this m in the training set.
4. During the forest growing, the value of m is held constant.
5. Each tree is grown to the largest extent and there is no pruning.

The offline phase uses the training dataset to build intrusion patterns used by the online phase. The pattern builder output the feature importance values used in the anomaly detection part. The offline phase also processing the uncertain items that are stored in the anomaly database. Weighted k-means algorithm as a data mining clustering algorithm is used to cluster the connection data. The anomaly detector determines the anomalous and normal clusters based on the count of known intrusions into them. The anomaly detection system generates an alarm, if unknown intrusions are detected. And add this new intrusion pattern to the intrusion pattern builder of the misuse detection part.

B. The Weighted k-means algorithm

Weighted k-means algorithm is a data mining clustering algorithm. It is a modified version of k-means algorithm by adding a weight of the data features. Algorithm as follows [16]:

1. Initially, K clusters are picked randomly as a “centroids”.
2. Assigning each node to its closet centroids.
3. Calculating mean of assigned points. Relocating each centroid based on the mean value.
4. Repeat step 2and 3, until convergence will occur.

The weighted k-means algorithm requires minimum and maximum value of each data in our KDD’99 dataset for calculating the weight.

IV. EXPERIMENTAL RESULT

The hybrid detection approach is evaluated using KDD’99 dataset. KDD”99 has been most widely used in network attacks. It has 494,019 records. Each record has 41 attributes and labeled as normal and attack. Table I list the connection type in the 10 % dataset. Intrusion patterns are built from KDD’99 dataset. With the built patterns, we use, the misuse detection for detecting intrusions over the test set.

The network connections in KDD’99 contain four categories of attacks. They are DOS, U2R, R2L and PROBE.

a. Denial of Service Attack (DOS): This attack is to deny the victim access to particular resources. Eg.'neptune', 'back', 'smurf', 'pod', 'land' and 'teardrop'.

b. User to Root Attack (U2R): Attackers try to access normal user account and gain root access information of the system. Eg. 'buffer_overflow', 'loadmodule','rootkit' and 'perl'.

c. Remote to Local Attack (R2L): 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. Eg. 'multihop', 'imap' and 'warezmaster'.

d. Probing Attack (PROBE): Attacker attempt to gather information about network of computers for the apparent purpose of circumventing its security. Eg. 'satan', ' nmap' and 'ipsweep'.

The KDD'99 dataset contains 494,019 records and three of them are categorical features. They are protocol_type, service and flag. Table II shows the categorical feature and their values. These features are encoded to binary valued features.

- Protocol_type:defines the protocol of the connection
- Service_type : defines the service of the connection
- Flag: defining the normal and the error status of the connection.

A Euclidean distance function is used in weighted k-means clustering algorithm for calculating the distance between each point.

TABLE I. NETWORK CONNECTION IN 10% KDD'99 DATASET BASED ON CONNECTION TYPE

Connection Type	10% Dataset
Normal	97,277(19.69%)
DOS	391,458 (79.24%)
Probe	4,107 (0.83%)
R2L	1,126 (0.23%)
U2R	52 (0.01%)

TABLE III. CATEGORICAL FEATURES AND THEIR VALUES IN 10% KDD'99 DATASET

Feature Name	Feature Values		
protocol_type	icmp	tcp	udp
flag	OTH	REJ	RSTO
	RSTOS0	RSTR	S0
	S1	S2	S3
	SF	SH	
service_type	aol	auth	bgp
	courier	csnet_ns	ctf
	daytime	discard	domain
	domain_u	echo	eco_i
	ecr_i	efs	exec
	finger	ftp	ftp_data
	gopher	harvest	hostnames
	http	http_2784	http_443
	http_8001	imap4	IRC
	iso_tsap	klogin	kshell
	ldap	link	login
	mtp	name	netbios_dgm
	netbios_ns	netbios_ssn	netstat
	nnsp	nntp	ntp_u
	other	pm_dump	pop_2
	pop_3	printer	private
	red_i	remote_job	rje
	shell	smtp	sql_net
	ssh	sunrpc	supdup
	systat	telnet	tftp_u
	tim_i	time	urh_i
	urp_i	uucp	uucp_path
	vmnet	whois	X11
	Z39_50		

V. CONCLUSION AND FUTURE WORK

In this paper, two data mining techniques are used in misuse and anomaly detection. A random forest classification algorithm is used in misuse detection part. And weighted k-means clustering algorithm is used for cluster the data. Random forest is a powerful algorithm for building the patterns automatically instead of coding rules manually. The proposed approaches are evaluated over 10% KDD'99 dataset.

In misuse detection framework, intrusion patterns are built in the offline phase. The main characteristic of misuse detection techniques is in comparing network traffic against a predefined intrusion pattern in order to decide whether it is considered an attack. In case of anomaly detection techniques involves any significant deviation of a system from normal behavior. Hybrid framework, we used advantages of both misuse and anomaly detection, thus offering speed and accuracy to detect the intrusion. To improve the performance of clustering, we are modifying the clustering algorithm by including a weight of data feature. The result shows that our framework achieves a higher detection rate and low false positive rate, compared to other approaches. In the hybrid

framework, in order to improve the performance of the anomaly detection component, misuse detection is applied first to filter out the known intrusions from the datasets. Thus, the number of connections in the anomaly detection component is significantly reduced. The limitation of the hybrid system is to keep the intrusion patterns in the dataset need to be much less than normal data. Another problem associated with our hybrid system, in anomaly detection, some intrusions cannot detect if it has a high degree of similarity.

In future, more advanced data mining algorithms could be investigated to overcome the earlier limitations. And try to make all process online. The performance of weighted k-means algorithm is strongly depending on the value of k clusters. Try to find the best method for deciding the value of k.

ACKNOWLEDGMENT

We are grateful to to Dr. R. Vijayakumar, Professor, School of Computer Sciences, M G University for providing all facilities. This work was supported by High Performance Computing Center of Fedaral Institute of Science And Technology(FISAT), Angamally.

REFERENCES

[1] CSI/FBI Computer Crime and Security Survey. (2004). Computer Security Inst., San Francisco, CA. http://www.issa-sac.org/docs/FBI2004.pdf

[2] R.Bane, N.Shivsharan, "Network intrusion detection system (NIDS)", 2008, pp.1272-1277.

[3] S. T. Brugger, "Data mining methods for network intrusion detection", 2004, pp. 1-65.

[4] Snort Intrusion detection system.(2006). www.snort.org

[5] Reda M. Elbasiony, Elsayed A. Sallam, Tarek E. Eltobely, Mahmound M. Fahmy, "A hybrid network intrusion framework based on random forest and weighted k-means," Ain Shams Engineering Journal, 2013.

[6] Jiong Zhang, Mohammad Zulkernine, and Anwar Haque, "Random-forest-based network intrusion detection systems," IEEE transactions on systems, man, and cybernetics-part c: Applications and Reviews, Vol. 38, No. 5, September 2008

[7] Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava and Pang-Nig Tan, "Data Mining for Network Intrusion Detection".

http://www-users.cs.umn.edu/~kumar/papers/nsf_ngdm_2002.pdf

[8] D. Barbara, J. Couto, S. Jajodia, L. Popyack, and N. Wu, "ADAM: Detecting intrusions by data mining," in *Proc. 2nd Annu. IEEE Workshop Inf. Assur.Secur.*, New York, Jun. 2001, pp. 11-16.

[9] Manikandan R, Oviya P, and Hemalatha C, "A new data mining based network intrusion detection model," Journal of Computer Applications, vol.5, February 2012.

[10] M. Mahoney and P.Chan, "An analysis of the 1999 DARPA/Lincoln laboratory evaluation data for network anomaly detection," in *Proc. Recent Adv. Intrusion Detect.(RAID)*, Pittsburgh, PA, Sep. 2003, Lecture Notes in Computer Science, vol. 2820, pp. 220-237.

[11] Salvatore j. Stolfo, Wenke Lee, Philip K. Chan, Wei Fan and Eleazar Eskin, "Data mining-based intrusion detectors: An overview of the Columbia IDS Project," Columbia University, September 2001.

[12] Amit Sharma, S. N. Panda and Ashu Gupta, "data mining techniques and their role in intrusion detection systems", *J. Acad. Indus. Res.* Vol.1 (4), September 2012.

[13] Mohammad Khubeb Siddiqui and Shams Naahid, " Analysis of KDD CUP 99 dataset using Clustering based Data Mining", International Journal of Database Theory and Application, Vol. 6, No.5, 2013, pp. 23-34.

[14] Wenke Lee , Salvatore j. Stolfo, " A framework for constructing features and models for intrusion detection systems," *ACM Transactions on information and system security*, vol. 3, No. 4, pp. 227-261, November 2000.

[15] Salvatore j. Stolfo, Wenke Lee, Philip K. Chan, Wei Fan and Eleazar Eskin, "Data mining-based intrusion detectors: An overview of the Columbia IDS Project," SIGMOD Record, Vol. 30, no.4, December 2001.

[16] Data mining Algorithms In R/Clustering/K-Means.

http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/K-Means Kdd99