2017 14th International ISC (Iranian Society of Cryptology) Conference on Information Security and Cryptology (ISCISC)
September 6-7, 2017; Shiraz University - Shiraz, Iran

Hybrid Intrusion Detection: Combining Decision Tree and Gaussian Mixture Model

Marzieh Bitaab
Computer Science and Engineering Dept., ECE
School Shiraz University
Shiraz, Iran
bitaab@cse.shirazu.ac.ir

Sattar Hashemi
Computer Science and Engineering Dept., ECE
School Shiraz University
Shiraz, Iran
s_hashemi@shirazu.ac.ir

***Abstract*—Nowadays, cybercrimes have become a major threat for computer networks. Many researchers considered Network Intrusion Detection System (NIDS) as a layer of defense and proposed new methods for detecting malicious network traffics. In this paper, we propose a hybrid method for detecting intrusion in networks. Using hybrid techniques exploits the strength of both misuse and anomaly detection methods. In our technique, we use decision tree for the misuse detection component and Gaussian Mixture Model (GMM) for anomaly detection. The advantage of using GMM is that it can recognize the attacks, which are similar to the normal distributions. The proposed technique's performance is evaluated on NSL-KDD dataset. Our empirical observations indicate that the proposed technique is a method of choice by offering higher accuracy and AUC while preserving lower false positive rates.**

Keywords—component; formatting; style; styling; insert (key words)

I. Introduction

Nowadays the rate of using computational devices has increased. Interconnected computers and network devices are often used in an environment where most protocols and applications are vulnerable to intrusion. Hence, it is important to deploy and improve the security of computer systems in a continuous manner. The presence of computer security systems attempt to assure confidentiality, integrity and availability of information resources [1].

An intrusion detection system (IDS) is a security system, which monitors network traffics or processes activities on systems to detect malicious activities or intrusions. According to Mukherjee [2], the term of intrusion detection is defined as identification of users who attempt to use a computer system without authorization, or authorized users who are abusing their privileges.

Intrusion detection systems are categorized into two types based on the place they look for the attack signature, host based intrusion detection system (HIDS) and network based intrusion detection system (NIDS). HIDS monitor incoming and outgoing traffic, users' activities, running processes, etc. on a single host or device. Hence, they cannot detect attacks on any other part of the network. On the other hand, NIDS monitors network traffics and looks for malicious signatures or behaviors to detect network based attacks [3].

Anomaly detection and misuse detection are the main types of IDS. A misuse detection system takes advantage of the previously known attacks pattern in their detection process, thus these methods have a low false positive rate and low detection rate. However, they are incapable of detecting unseen attacks which are not similar to their stored signature of known attacks. On the other hand, anomaly detection systems create a profile for normal behavior and they presume that intruders have different behavior from normal user. Anomaly detection methods are capable of detecting unseen attack since they only consider normal behaviors for their detection. However, these methods have high false positive rate since their hypothesis is not always valid. In other words, it is not true to consider the patterns with different behavior than normal activities as an attack, it might be a new normal behavior [4].

To alleviate the disadvantage of these two methods, recent techniques mostly apply hybrid approaches that applies both misuse and anomaly detection methods for their detection [5].

In this paper, we propose a hybrid intrusion detection method, which employs misuse detection model to improve anomaly detection. In general, anomaly detection methods create profile for all normal data. However, the proposed method creates a profile for each subset of normal dataset. As a result, it is possible to find a more accurate profile. In the proposed method decision tree is employed to separate known attack samples from normal ones, then a GMM model is learned for each normal leaf of the tree.

The rest of the paper is organized as follows: Section 2 reviews previous intrusion detection approaches. In Section 3 description of proposed method is presented in detail. Section 4 provides our experimental settings and finally Section 5 concludes the paper.

II. Related Work

Early approaches of intrusion detection are mostly based on one of anomaly detection or misuse detection methods.

Mukkamala et al. [6] present an ensemble approach for intrusion detection by combining artificial neural network,

978-1-5386-6560-2/17/$31.00 ©2017 IEEE

support vector machines and multivariate adaptive regression splines. The evaluation in this experiment is based on DARPA data of MIT Lincoln Laboratory. Jemili et al. [7] propose a framework using Bayesian network for adaptive intrusion detection. It is called adaptive since it adds new intrusion signatures to learning dataset. Cannady analyzes the applicability of neural network for intrusion detection. The advantage of this method is using neural network that leads to flexibility of network and ability to learn characteristic of new attack behaviors. However, the disadvantage of using neural network cannot be ignored.

The disadvantages of misuse detection and anomaly detection are inability in detecting unknown attack and having high false positive rate, respectively. Due to these disadvantages, use of hybrid intrusion detection systems have been growing lately and most recent studies focus on this type. Aydin et al. [5] proposed a hybrid method that combines SNORT as misuse detection and NETAD and PHAD as anomaly detection component. Another hybrid method proposed by Govindarjan et al. [8] uses bagging ensemble method with MLP and RBF neural networks and aggregates their results. This method works better than standalone MLP or RBF neural network.

Most intrusion detection systems based on machine learning techniques suffer from high time complexity mainly caused by large data and high dimensions. A solution to this problem is reducing the number of train data which effects both time complexity and accuracy. Chun Guo et al. [9] proposed a method for extracting important data to improve detection of malicious packets and therefore achieve more accurate results.

Kim et al. [10] created a C4.5 decision tree based on training data and used one-class SVM for each subset of normal dataset. It flags newly received data as unknown attack if corresponding one-class SVM classify it as not normal. Furthermore, Guo [11] proposed a two-level IDS. At the first stage, the anomaly detection takes advantage of K-means and “Add One In”. The second stage, composed of two parallel components, uses K-Nearest Neighbors algorithm. Using AOI, it is possible to explore the variation of cluster centers when a new data instance is added.

III. Proposed Method

In the current research a hybrid IDS method is proposed through combining C4.5 decision tree and Gaussian mixture model. The key assumption is that normal data takes different variations due to different protocol types, services etc. That is why we take advantage of decision tree to split data into different partitions, then for each normal partition, we use GMM to model normal data.

A. Decision Tree

Decision tree is one of the most used classifiers in machine learning [10]. A decision tree can be transformed into sets of rules to classify new instance and consists of a root, set of internal nodes and leaves. To classify new data instances, it is checked with rules according to the decision tree till it reaches a leaf. The label of data instance will be assigned according to leaf’s label. C4.5 is a kind of decision tree that was developed by Quinlan [12]. It chooses best attributes for each division according to information gain. Information is a measure of purity and information gain is amount of information we gain by splitting data using a specific attribute. Information gain is calculated using (1), node *p* is splitted into *k* partitions according to a specific feature and n_i is the number of records in portion *i*:

$$Information\ Gain = Entropy(p) - (\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)) \qquad (1)$$

Where *Entropy(p)* is amount of impurity of parent node *p* and (2) is sum of impurity on all child nodes. Higher information gain lead to higher purity of child nodes, which means the chosen attribute, is good enough. *Entropy* is calculated using formula (3):

$$\textstyle\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \qquad (2)$$

$$Entropy(X) = -\sum_{i=1}^{k} P(x = i) \log_2 P(x = i) \qquad (3)$$

B. Gaussian Mixture Model

A mixture of distributions for observations is known as mixture model. Gaussian Mixture Model (GMM) is an algorithm for clustering and density estimation. GMM assumes the observations are from several Gaussian distributions with unknown parameters, which will be estimated with EM algorithm. GMM models the observation using the following likelihood function [13]:

$$p(x) = \sum_{k}^{K} \pi_k \mathrm{N}(\mathrm{x}|\mu_\mathrm{k}, \Sigma_\mathrm{k}) \qquad (4)$$

in which the $N(x|\mu_k,\Sigma_k)$ represents multinomial normal distribution:

$$\mathrm{N}(\mathrm{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)'\Sigma(x-\mu)} \qquad (5)$$

Expectation Maximization (EM) is an algorithm to find estimates of parameters for maximum likelihood (ML) or maximum a posteriori (MAP) where the models depends on unobserved latent variables. This algorithm has two steps, E-step calculates the expectation of the Gaussian component assignments for each data point and M-step maximizes the expectations calculated in previous step. This procedure continues until it converges [14].

C. Proposed Hybrid Intrusion Detection Method

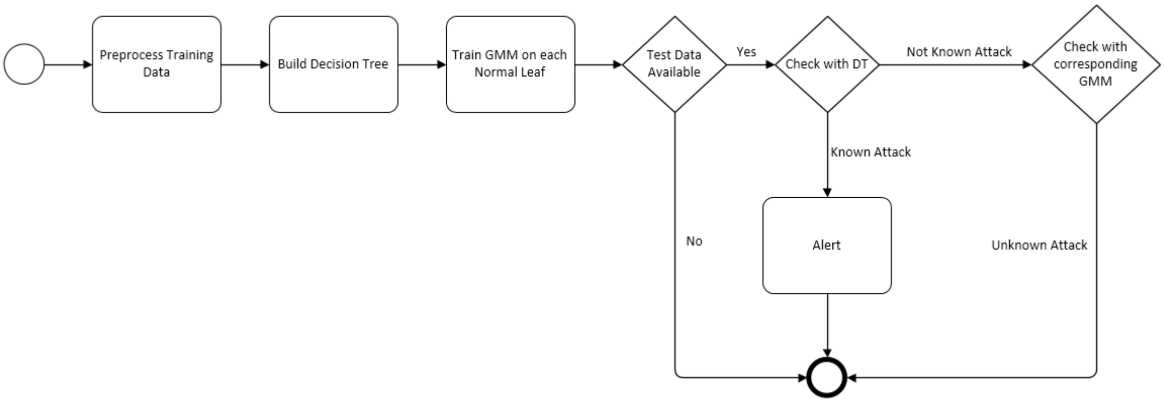

Fig. 1. Diagram of proposed hybrid intrusion detection

Our proposed hybrid intrusion detection method, learns the decision tree model with training data consist of known attacks and normal data. Decision tree will detect known attacks, if a new arrival instance isn't similar to known attacks the decision tree will end up labeling it as normal, however it might be an unknown attack but its behavior is similar to normal behavior. To resolve this problem, for each normal leaf a GMM is trained to find their distribution. Its framework is in fig. 1.

One advantages of using GMM for finding normal boundary is that there might be more than one peak in the distribution of data in each leaf as there are different types of normal data that may belongs to different distributions. Thus, modeling multimodal data as a mixture of many unimodal Gaussian distributions intuitively makes sense while trying to fit a multimodal distribution with unimodal model will give poor fit. Fig. 2 shows two different modes if we try to fit one Gaussian distribution, it'll misclassify attack data between the two modes, but if we fit two Gaussian distributions, attacks between these two components will be detected correctly. Another advantage of GMM is that it is computationally inexpensive.

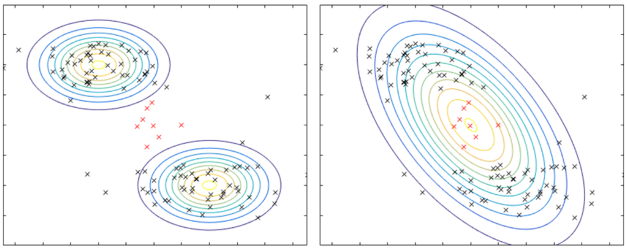

Fig. 2. Comparision of multimodal and unimodal GMM

As we use GMM on each normal leaf of the decision tree there may be small sample size problem, so, we set minimum number of data for each leaf of the decision tree. In addition, we prune the decision tree according to a given confidence interval to avoid overfitting. The detailed process is showed in fig3.

IV. Experimental Setup

A. Dataset

We apply our proposed method on NSL-KDD dataset that is a well-known publicly available dataset and present the achieved results in the following section. NSL-KDD is the new version of KDD'99 that consists of network connections and has 42 attributes with a label that determines connection type into 5 main categories. Table 1 attack classes namely DoS, probe, R2L and U2R with detailed discription of attacks [15][16]:

TABLE I. Attack Descriptions

Attack Class	Description	Attack Type
DOS	attempt to make system unavailable	back, land, neptune, smurf, pod, teardrop
Probe	collect information about the host by scanning network	ipsweep, nmap, portsweep, satan
R2L	send packets to a remote system over internet to exploit priviledges and gain access to system	guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster, ftp_write
U2R	normal user attempt to gain root access	buffer_overflow, loadmodule, perl, rootkit

B. Evaluation Result

To evaluate our IDS method, we use detection rate, accuracy and false positive rate:

1) *Detection Rate:* $\frac{TP}{TP+FP}$

2) *False Positive Rate:* $\frac{FP}{FP+TN}$

3) *Accuracy:* $\frac{TP+TN}{TP+FP+TN+FN}$

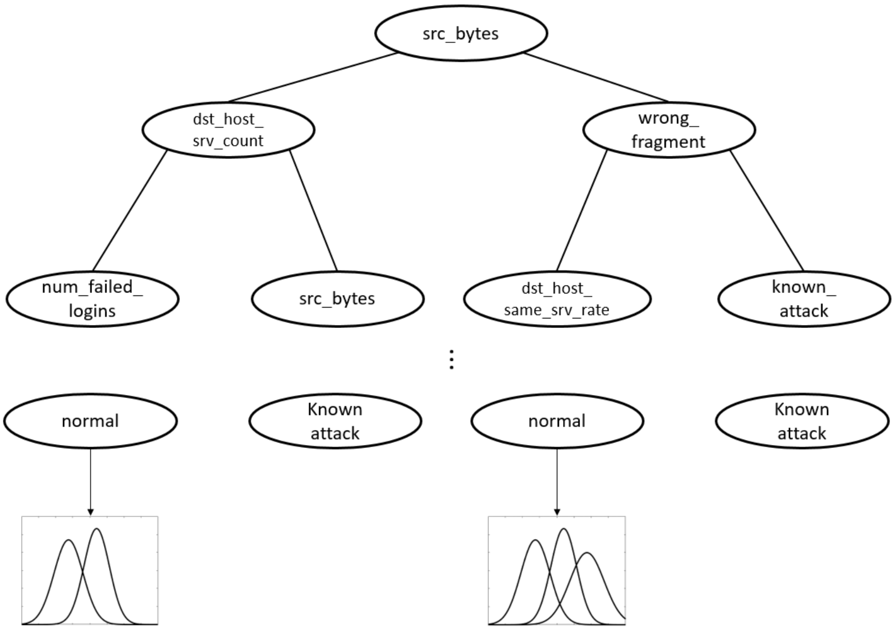

Fig.3. Detailed training process of proposed hybrid method

Where TP (True Positive) shows the number of attacks predicted correctly, FP (False Positive) represents the number of normal behaviors predicted as attack incorrectly, TN (True Negative) equals the number of normal behaviors predicted correctly and FN (False Negative) stands for the number of attacks predicted as normal incorrectly [17].

Our model needs several parameters to work properly. Each of decision tree and GMM has different parameters. For decision tree we have used 10% confidence interval and minimum number of data in each leaf is 1% of total number of training data. These parameters for decision tree prevents overfitting and also we have enough data in each leaf to perform GMM (the number of data must be greater than number of attributes). To perform GMM on each leaf we must determine the number of Gaussians to fit on leaf's data. We have set number of Gaussians to 10 for normal data. The number of Gaussian components can be determined by the amount of training data available [13].

Fig. 4 and 5 represent receiver operating characteristic (ROC) curve for the proposed method in comparison with with the hybrid method proposed in [10] that combines decision tree and SVM. The method has been tested in two different mod of dataset. First one is the modified version of NSL-KDD dataset. First we split test set to two part known attacks and unknown attacks. We combine known attacks of test set with train dataset after shuffling, we evenly divide this combined data. Unknown attack of test dataset is combined with a one part of known attack, and it is used as new test set [10]. The process of creating new train and test set is shown in fig. 6. The purpose of manipulating dataset is that to make characteristic of train and test data sufficiently similar to each other. The second dataset is the original NSL-KDD dataset.

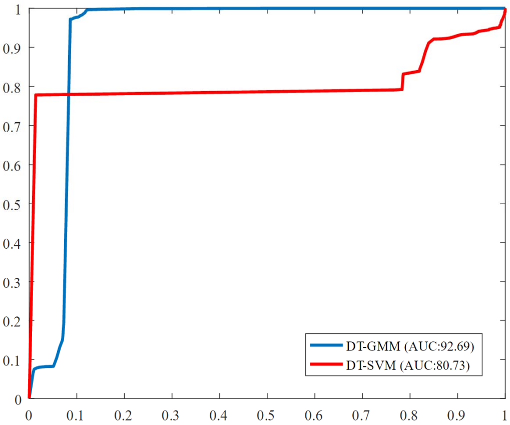

Fig. 5. Performance comparison using ROC curve on NSL-KDD

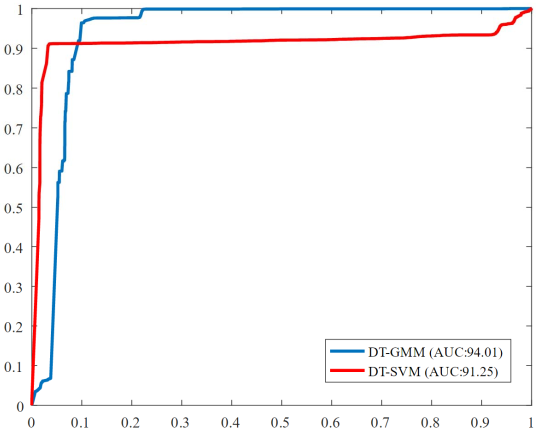

Fig. 4. Performance comparison using ROC curve on baseline

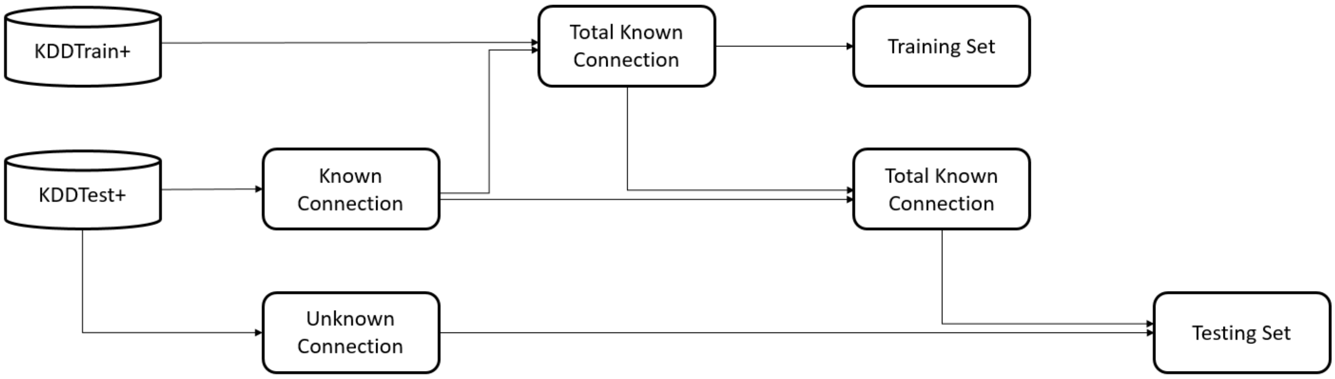

Fig.6. Creating testing and training set

We provide a comparison of detection performance of our proposed method with [10] in table 2. As it is depicted in table 2, the propose method reduce false positive rate it significantly improve accuracy and detection rate. It is due to the fact that not only we take advantage of misuse detection component to detect known attacks and decompose normal data, but also we use a multimodal technique in our anomaly detection component. Since a multimodal learner fit different distribution on data it is possible to detect attacks which are between distributions.

TABLE II. COMPARISONS OF RESULTS OBTAINED BY THE PROPOSED METHOD AND DT-SVM METHOD

Measure/ Method	Dataset 1		Dataset 2	
	DT-SVM	DT-GMM	DT-SVM	DT-GMM
Detection Rate	96.82	**97.21**	**97.65**	96.72
False Positive Rate	18.53	**8.59**	15.10	**9.37**
Accuracy	89.05	**94.28**	92.16	**94.10**

REFERENCES

[1] -J. L. a, C.-H. R. L. a n, and Y.-C. L. a b, “Intrusion detection system: A comprehensive review,” *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 16–24, 2012.

[2] B. Mukherjee, L. T. Heberlein, and K. N. Levitt, “Network intrusion detection,” *IEEE Network*, vol. 8, no. 3. pp. 26–41, 1994.

[3] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, “Network Anomaly Detection: Methods, Systems and Tools,” *Commun. Surv. Tutorials, IEEE*, vol. 16, no. 1, pp. 303–336, 2014.

[4] A. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” *IEEE Commun. Surv. Tutorials*, vol. PP, no. 99, p. 1, 2015.

[5] M. A. Aydin, A. H. Zaim, and K. G. Ceylan, “A hybrid intrusion detection system design for computer network security,” *Comput. Electr. Eng.*, vol. 35, no. 3, pp. 517–526, 2009.

[6] S. Mukkamala, A. H. Sung, and A. Abraham, “Intrusion detection using an ensemble of intelligent paradigms,” *J. Netw. Comput. Appl.*, vol. 28, no. 2, pp. 167–182, 2005.

[7] F. Jemili, M. Zaghdoud, and M. Ben Ahmed, “A Framework for an Adaptive Intrusion Detection System using Bayesian Network,” *2007 IEEE Intell. Secur. Informatics*, no. March, pp. 66–70, 2007.

[8] M. Govindarajan and R. Chandrasekaran, “Intrusion detection using neural based hybrid classification methods,” *Comput. Networks*, vol. 55, no. 8, pp. 1662–1671, 2011.

[9] C. Guo, Y. J. Zhou, Y. Ping, S. S. Luo, Y. P. Lai, and Z. K. Zhang, “Efficient intrusion detection using representative instances,” *Comput. Secur.*, vol. 39, no. PART B, pp. 255–267, 2013.

[10] G. Kim, S. Lee, and S. Kim, “A novel hybrid intrusion detection method integrating anomaly detection with misuse detection,” *Expert Syst. Appl.*, vol. 41, no. 4 PART 2, pp. 1690–1700, 2014.

[11] C. Guo, Y. Ping, N. Liu, and S. S. Luo, “A two-level hybrid approach for intrusion detection,” *Neurocomputing*, vol. 214, pp. 391–400, 2016.

[12] J. R. Quinlan, “Constructing decision tree,” *C4*, vol. 5, pp. 17–26, 1993.

[13] D. a Reynolds, “Gaussian Mixture Models,” *Encycl. Biometric Recognit.*, vol. 31, no. 2, pp. 1047–64, 2008.

[14] T. K. Moon, “The Expectation-Maximization Algorithm,” *IEEE Signal Processing Magazine*. pp. 47–60, 1996.

[15] H. H. Pajouh, G. Dastghaibyfard, and S. Hashemi, “Two-tier network anomaly detection model: a machine learning approach,” *J. Intell. Inf. Syst.*, no. 2, pp. 1–14, 2015.

[16] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the KDD CUP 99 data set,” *IEEE Symp. Comput. Intell. Secur. Def. Appl. CISDA 2009*, no. Cisda, pp. 1–6, 2009.

[17] W. C. Lin, S. W. Ke, and C. F. Tsai, “CANN: An intrusion detection system based on combining cluster centers and nearest neighbors,” *Knowledge-Based Syst.*, vol. 78, no. 1, pp. 13–21, 2015.