Volume 4, Issue 4, April 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Intrusion Detection System with Supervised Learning Algorithms

Miss. M. R. Yadav
Computer Engineering & Pune University
India

Prof. P. B. Kumbharkar
Research Scholar, JJT University, Rajasthan
India

Abstract— Normally, The Internet grows rapidly and most useful in each domain but network vulnerability and intrusions are still an important issue that causes attacks. A good system to detect the illegal user is to monitoring the packets and using the different algorithms, methods and applications which are created and implemented to solve the problem of detecting the attacks in intrusion detection systems. We consider network intrusion detection using supervised learning algorithm to classify attacks in the datasets. Fuzzy rule is a machine learning algorithm that can classify network attack data and protect the system from damage, while a genetic algorithm is an optimization algorithm that can help finding appropriate fuzzy rule and give the optimal solution. We consider both well-known KDD99 dataset. We evaluate our IDS in terms of detection speed, detection rate and false alarm rate.

Keywords— Fuzzy genetic algorithm; multilayer perceptron algorithm; Ada-boost algorithm; intrusion detection; network security

I. INTRODUCTION

Internet grows rapidly but intrusion detection is an important issue in computer security. Information technology has become a critical component in the organization that manages the huge amount of data. Securing those systems from different actions should be the main goal when applying security, but the evolution of technologies makes this task very difficult. Information security is dependent on the things like "Protection, Detection, Reaction, and Recovery". Intrusion detection is a crucial part of information security. IDS are the software or hardware tools that automatically scan the events that take place in network, looking for intrusion. Any activity aimed at disrupting a service or making resource unavailable or gaining unauthorized access can be termed as an intrusion. IDS can be deployed to detect the attack. Generally, IDS necessary to detect the network attacks before they damage the whole system. Organization uses IDS system for different purpose, such as identifying problems with security policies, threats and individual from violating security policies. IDS typically record the information related to observed events, Notify security administrators of important observed events, and produce report and uses the several response techniques which involve the IDS stopping the attack and protect the system. Intrusion detection is the detection of network behaviours that violate network security. Intrusion detection is distinguishing between network attacks and normal network behaviours or further distinguishing between different categories of attacks. Different types of IDS are available like "Host-based, Network-based, Stack-based Intrusion Detection System [1]. Most classification techniques for intrusion detection can be based on various classification algorithms. They are classified into two groups, which are grouped into supervised learning approach and Unsupervised learning approach.

Contribution Work

In summary, there are many different algorithms for network intrusion detection. Most of them considered well-known KDD99 dataset. Therefore, we propose to use a FGA algorithm to detect intrusions on KDD99 dataset. The FGA algorithm work on kdd99 dataset the algorithm then classify the data into attack and normal from used datasets. Thus we classify the data sets into attacks and normal. This is the overview of the system which is implemented using eclipse. Thus it is intrusion detection system using the supervised learning algorithm on the datasets.

The rest of this paper is presented as follows. In section II we represent the Literature survey. In section III, We discuss on Implementation with algorithm, in section IV we present the Results with dataset preparation and performance evaluation criteria. In section V we give a conclusion of this research work and future work. document is a template. An electronic copy can be downloaded from the Journal website. For questions on paper guidelines, please contact the journal publications committee as indicated on the journal website. Information about final paper submission is available from the conference website.

II. LITERATURE SURVEY

There is no standard and actual definition of intrusion detection. Intrusion detection is the detection of network behaviours that violate network security. Intrusion detection is distinguishing between network attacks and normal network behaviours or further distinguishing between different categories of attacks. Different types of IDS are available like "Host-based, Network-based, Stack-based Intrusion Detection System [1].

A. Host-based intrusion detection

It consists of an agent on a host that identifies intrusions by analyzing system calls, application logs, file-system modifications are binaries, password files, capability databases, Access control lists.[1] and other host activities and state. In a HIDS, sensors usually consist of a software agent. Some application-based IDS are also part of this category.

B. Stack-based intrusion detection

This type of system consists of an evolution to the HIDS systems. The packets are examined as they go through the TCP/IP stack and, therefore, it is not necessary for them to work with the network interface in promiscuous mode. This fact makes its implementation to be dependent on the Operating System that is being used.

C. Network-based intrusion detection

Network intrusion detection system is an independent platform that identifies intrusions by examining network traffic and monitors multiple hosts, developed in by Pete R. Network intrusion detection systems gain access to network traffic by connecting to a network hub, network switch configured for port mirroring, or network tap[1]. In a NIDS, sensors are located at choke points in the network to be monitored, often in the demilitarized zone (DMZ) or at network borders. Sensors capture all network traffic and analyse the content of individual packets for malicious traffic. An example of a NIDS is Snort.

Most classification techniques for intrusion detection can be based on various classification algorithms. They are classified into two groups, which are grouped into supervised learning approach and Unsupervised learning approach.

A. Supervised learning-based approach

In supervised learning approach, the instances consist of input attributes and desirable output and the algorithm would produce an inferred function, which is called a classifier or regression function. This approach has high accuracy, low false- alarm with fast computing time. Supervised learning methods for intrusion detection can only detect known intrusions. In these methods from machine learning and pattern recognition have been utilized to detect intrusions. For supervised learning for intrusion detection, there are mainly supervised neural network (NN)-based approaches [3], [4], and support vector machine (SVM)-based approaches [5].

B. Unsupervised learning-based approach

Unsupervised learning methods can detect the intrusions that have not been previously learned. Examples of unsupervised learning for intrusion detection include *K*-means-based approaches and self-organizing feature map (SOM)-based approaches [6], [7].

Previous discussions review the related work. Different approaches for intrusion detection have the following:

A. J. Gómez and E. León IDS System

J. Gomez and E. Leon [8] proposed fuzzy and genetic algorithm to classify behaviour of intrusion. The input data is KDDCup99 dataset which consists of 42 features. The fuzzy rule is automatically adapted using evolutionary technique and genetic algorithm. The algorithm can classify the data into 5 classes including DoS, Probe, R2L, U2R and Normal.

B. Solution by T.P. Fries

T. P. Fries [9] proposed a fuzzy genetic algorithm approach. In the pre-processing phase, they used clustering algorithm and genetic algorithm to find significant attributes in KDD99 dataset. In the detection phase, they used fuzzy GA algorithm. The algorithm has high performance in terms of speed, memory consumption and robust for large problems.

C. T. Komviriyavut et al Dataset Methods

T. Komviriyavut et al [10] proposed a method to preprocess dataset in actual network environment within 2 seconds. The pre-processed data has 12 attributes. Then, they used a decision tree algorithm to classify data (output classes are DoS, Probe and Normal). This technique is efficient to be used in actual network environment.

D. M.-Y. Su etal Real-time IDS

M.-Y. Su etal. [11] Proposed a Real-time IDS for large-scale attacks by using fuzzy association rules. The technique pre-processed packet header into 16 attributes from opened network environment in every 2 seconds (the network that connects to internet and allow every packets flow through it). Then, each record will be sent to another computer in order to update new rule. However this technique does not show the detection rate and is able to detect only DoS attack.

E. Work by P. Kachurka and V. Golovko

P. Kachurka and V. Golovko [12] proposed a neural network approach to real-time network intrusion detection; they collected the network traffic by using an open source intrusion detection system (Bro IDS). This technique is able to detect unknown attack in real time. However, it can classify only 2 classes attack and normal.

Thus we construct the "Fuzzy genetic algorithm" for the IDS which have supervised learning approach. Which has high accuracy, low false- alarm with fast computing time.

III. IMPLEMENTATION

In this section, we explain our FGA algorithms on KDD99 dataset and performance evaluation criteria. According to the characteristics of the FGA algorithm and the intrusion detection problem we implement the intrusion detection using the FGA.

A. Design

The design and process of this experiments that were conducted are described as follow.

1. We are going to create IDS System for detection of attacks in that we will use KDD99 dataset which gives an input to our application.
2. After that we pass this dataset file to our Fuzzy Genetics Algorithm in that we will calculate FP, FN and DR.
3. We extend our application for detect the attacks and classify the types of attack in kdd99 dataset.

4. Additionally we will compare with supervised learning Algorithm for detect and classify the attacks from KDD99 and online dataset.

The process flow for the system architecture is described using Fig. 1 which is the sequence of process which is followed sequentially in system flow.

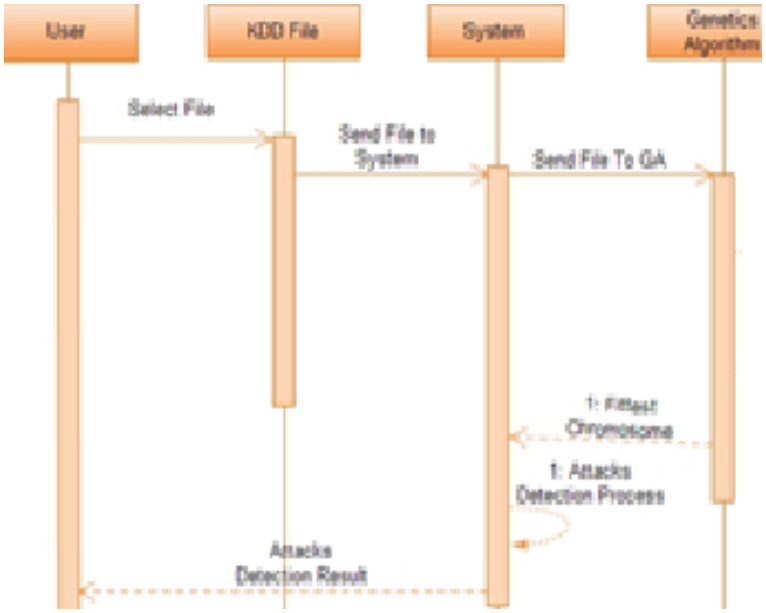

Fig. 1 Process flow

B. Algorithms

1) Fuzzy Genetic Algorithm

We consider using a FGA which is similar to the work proposed by T.P. Fries [9]. The FGA can classify two classes which are normal class and attack class. The steps to use FGA are described as follows.

The Fuzzy genetic algorithm starts from a population of individuals generated randomly. Each individual is an "if then" fuzzy rule. In order to optimize the set of fuzzy rules already generated in the first stage, a genetic algorithm process which consists of selection, crossover and mutation operators are applied on the individuals. The

FGA for intrusion detection are defined in follow:

1. Select the natural selection of dataset find the fittest chromosomes.
2. Normalize each attribute of data to be a real number in the range 0.0 -7.0, where the maximum and minimum values among overall attributes from the training data are set to 7.0, and 0.0, respectively. The normalization is given by using (1) and applied in order to set attribute numerical values in the range [0.0, 7.0].

$$x' = \frac{x - \min a}{maxa - mina}(n\_maxa - n\_mina) + n\_mina \qquad (1)$$

Where *x*: is the numerical attribute value, *mina* is the minimum value that the attribute *x* can get and *maxa* is the maximum one and *n_maxa* and *n_mina* are the new values of attribute x.

3. Encode the each attribute value into binary format and convert it into 0.0 – 7.0 octal number. As a, b, c, d format as shown in Fig. 2.

010	011	100	101
a	b	c	d

Fig.2 Fuzzy encoding for each attribute (where a ≤ b≤ c ≤ d)

4. Find probability of each record for each detection rule and count for true negative and true positive as shown in the pseudo code below. A data record contains information of the attribute in the dataset.

```
for each record
  {  for each rule
       {  for each attribute
                          {     prob = fuzzy();
                                totalprob = totalprob + prob;  }
          If (totalprob > threshold)
          {         class is attack;
                    true negative ++;
          }
        Else
```



```
{
        class is normal;
        true positive++;
}
```

We used each rule to calculate a probability of being an attack of each data record. The system will read the record data one by one and evaluate each block of data to calculate a probability to be the attack using the trapezoidal fuzzy rule shape. Then, we gather a probability of each attribute and find the average probability. If the average probability is greater than a predefined threshold, the system would classify this record as an attack. We repeat this process for every record.

We used a trapezoidal shape to measure a probability of being an attack identified by each attribute. The fuzzy logic is encoded into four parameters which are a, b, c and d. The probability is calculated as shown in Fig. 3 where its meaning described below.

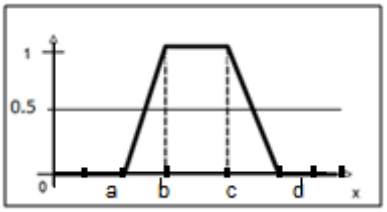

Fig. 3 A trapezoidal fuzzy rules with 4 parameters

From Fig .3, we calculate probability of being an attack from condition below

$$
\begin{aligned}
&if(attribute\_value\ is\ between\ "b"to\ "c") \\
&\qquad then\ prob = 0.0 \\
&else\ if\ (attribute\_value\ between\ "a"to\ "b"\ \ ) \\
&\qquad then\ prob = \frac{attribute\_value\ -\ a}{b\ -\ a} \\
&else\ if\ (attribute\_value\ between\ "c"to\ "d"\ \ ) \\
&\qquad then\ prob = \frac{d\ -\ attribute\_value}{d\ -\ c} \\
&else\ then\ prob = 1.0
\end{aligned}
$$

4. Preserve rule that has the highest optimal solution which is the best detection rule.
5. Use evolutionary GA method to find the next rules.

C. Performance Evaluation

With the help of the algorithms, we detect the attack and classify them and further provide the security against the attack and calculate,

1. Detection rate (DR) is the percentage of normal and attack correctly classified from total number of data records.
2. False-negative rate (FN) is the percentage that attack is misclassified from total number of attack records.
3. False-positive (FP) is the percentage that normal data is classified as attack from total number of normal data records.
4. The speed of detection is the time that our system uses, measured right after the records arrive until the system classifies the data and gives output classes in to normal or attack.

Thus using the "Fuzzy Genetic Algorithm" [14] we can detect the attack and classify the attack. The "Fuzzy genetic algorithm" is in two parts 1.Genetic algorithm 2.Rule creation. The FGA is easier for the attacks to understand the flow of algorithm hence he can easily implement different technique to work the system so it's so secure and rule creation process is easily so it is not time consuming and open source. The dataset had detected the five types of attacks. That is DoS, Probe, U2R, R2L and Normal attack only.

D. Comparisons with Different Supervised Learning Algorithms

There are the different supervised learning algorithms like MLP algorithm, Ada-Boost algorithm Random forest algorithm and so on. All are the machine learning supervised learning algorithms. We compare with it detection rate and the accuracy in Fig. 4.

The Multilayer Perceptron Algorithm for intrusion detection system on the datasets. The Multilayer Perceptron algorithm is supervised learning algorithm. The Multilayer Perceptron is an example of an artificial neural network that is used into number of layer and using weka tool, so it is more secure and easily implemented than other algorithms.The Multilayer Perceptron algorithm is supervised learning algorithm [15]. The Multilayer Perceptron is an example of an artificial neural network that is used extensively for the solution of a number of different problems, including pattern recognition and interpolation. It is a development of the Perceptron neural network model that was originally developed with layers that are input layer, output layer, and hidden layer.

In our AdaBoost-based algorithm for intrusion detection, decision stumps are used as weak classifiers. The decision rules are provided for both categorical and continuous features. Experimental results show that our algorithm has low computational complexity and error rates, as compared with algorithms of higher computational complexity, as tested on

the benchmark sample data. AdaBoost-based algorithm possesses the lowest computational complexity in the published learning algorithms for intrusion detection. This property is very attractive and promising because the classifiers for intrusion detection should be retrained very quickly in practice and fast detection is essential for an effective defense against intrusions.

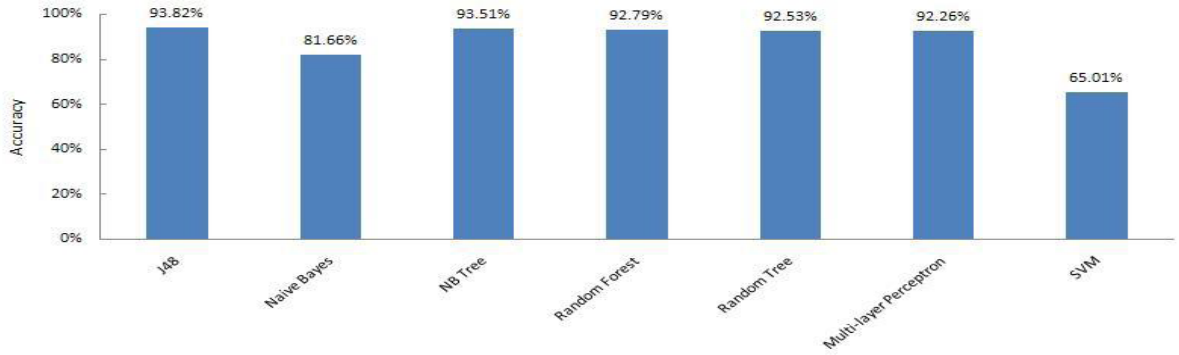

Fig.4 comparisons with different algorithm

IV. RESULTS

Building any classifier involves two phases, i.e., training and testing phases. Training phase in our approach involves learning the parameters of the model from a KDD99 dataset, and apply the algorithm on that datasets profiles this data and uses this information to test datasets. To build a classifier we need to have labeled data for training and testing. Data sets released by DARPA [13] were used to train and test our datasets.

A. Current Results

Intrusion detection system with FGA algorithm and on KDD99 dataset the detection rate is 94.65%. This is current stage results of the system shown in Table.1 as fallow.

Table.1 Results obtained for Kdd99 data sets

Detection Rate	False alarm Rate
94.654%	12.83%

B. KDD99 Dataset

The KDD99 dataset is a benchmark dataset which was simulated in military network environment in 1998 then derived to KDD99 dataset in 1999. The dataset package was gathered and preprocessed into 41 attributes. In this work, we consider using only eight important attributes from the dataset. The selected attributes are attributes "duration, src_bytes, num_failed_logins, root_shells, num_acess_files, srv_count, serror_rate, and same_srv_rate" from 41 attribute and apply the algorithms on that dataset and classify the attack and normal data.

V. CONCLUSIONS

We have proposed a Fuzzy Genetic Algorithm for intrusion detection. Fuzzy genetic algorithm for network intrusion detection, using well-known KDD99 dataset. Our Fuzzy genetic algorithm and online dataset has more recent attack types of DoS, Probe, U2R, R2L attack. Moreover, it has more current behaviours of network activities than those in the KDD99 dataset. The experimental results show that our fuzzyGA can efficiently detect where 2 seconds belong to the preprocessing time and less than a second for the detection time, while it takes only a fraction of a second to detect attacks in the KDD99 dataset. That shows that our Fuzzy Genetic algorithm is able to classify both KDD99 dataset with high accuracy and low false alarm rate. The experiments illustrated detection rate of each attack. We can see that the fuzzy genetic algorithm can mostly distinguish behaviour of each attack type in both KDD99 dataset and online dataset with low false negative rates. Our future work will focus on the following aspects.

1. To create our own dataset means captured data and then perform the experiments, so we have to perform without storing of data.

ACKNOWLEDGMENT

Our heartfelt thanks go to Siddhant College of Engineering for providing a strong platform to develop our skills and capabilities. We would like to thank to our guide & respected teachers for their constant support and motivation for us. Last but not least, we would like to thanks all those who directly or indirectly help us in presenting the paper.

REFERENCES

[1] S. Chebrolu, A. Abraham, and J. P. Thomas, "Feature deduction and ensemble design of intrusion detection systems," *Comput. Secur.*, vol. 24, no. 4, pp. 295–307, Jun. 2005.

[2] W. Lee and S. J. Stolfo, "A framework for constructing features and models for intrusion detection systems," *ACM Trans. Inf. Syst. Secur.*, vol. 3, no. 4, pp. 227–261, Nov. 2000.

[3] Y.-H. Liu, D.-X. Tian, and A.-M. Wang, "Annids: Intrusion detection system based on artificial neural network," in *Proc. Int. Conf. Mach. Learn. Cybern.*, Nov. 2003, vol. 3, pp. 1337–1342.

[4] C. Zhang, J. Jiang, and M. Kamel, "Intrusion detection using hierarchical neural networks," *Pattern Recognit. Lett*, vol. 26, no. 6, pp. 779–791, May 2005.

[5] Z. Zhang and H. Shen, "Online training of SVMs for real-time intrusion detection," in *Proc. Int. Conf. Adv. Inf. Netw. Appl.*, 2004, vol. 1, pp. 568–573.

[6] M. O. Depren, M. Topallar, E. Anarim, and K. Ciliz, "Network-based anomaly intrusion detection system uses SOMs," in *Proc. IEEE 12th Signal Process. Commun. Appl. Conf.*, Apr. 2004, pp. 76–79.

[7] H. G. Kayacik, A. N. Zincir-Heywood, and M. I. Heywood, "On the capability of an SOM based intrusion detection system," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2003, vol. 3, pp. 1808–1813.

[8] J. Gómez and E. León, "A fuzzy set/rule distance for evolving fuzzy anomaly detectors," IEEE International Conference on Fuzzy Systems, ART. No. 1682017, pp. 2286-2292.

[9] T.P. Fries, "A fuzzy-Genetic approach to network intrusion detection," GECCO'08: The 10th Annual Conference on Genetic and Evolutionary Computation, 2008, pp. 2141-2146.

[10] T. Komviriyavut, et al., "Network intrusion detection and classification with decision tree and rule-based approaches," 9th International Symposium on Communications and Information Technology, Art.No. 5341005, pp. 1046-1050.

[11] M-Y. Su, et al., "A real-time network intrusion detection system for Large-scale attacks based on an incremental mining approach," Computers and Security 28 (5), pp. 301-309.

[12] P. Kachurka, V. Golovko., "Neural network approach to real-time Network intrusion detection and recognition," The 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS'2011, Art.No. 6072781 , pp. 393-397.

[13] KDD99 dataset, a network dataset [online], http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[14] M. Saniee Abadeha, , J. Habibia, C. Lucasb, "Intrusion detection using a fuzzy genetics-based learning algorithm", Journal of Network and Computer Applications 30 (2007) 414–428.

[15] Norouzian M.R., Merati. S., "Classifying Attacks in a Network Intrusion Detection System Based on Artificial Neural Networks" Proceedings of the Advanced Communication Technology (ICACT), 2011 13th International Conference on Publication Year: 2011, Page(s): 868 - 873

