Neural Network Training by Selected Fish Schooling Genetic Algorithm Feature for Intrusion Detection

Nomaan Jaweed Mohammed

ABSTRACT
In an ever-growing world of internet users, network security has become an important aspect of today's digital age. Due to a multitude of users accessing the internet for a plethora of reasons, it has become imperative to identify an appropriate and safe network for which, an Intrusion Detection System (IDS) solution has been proposed. The proposed IDS solution utilizes Fish Schooling Genetic Algorithm and an error backpropagation neural network. The genetic algorithm has been used for detecting the good feature set from the training dataset and the selected good features train the neural network. This combination of genetic algorithm and Neural network increases the detection accuracy of intrusion with a lesser number of training features, and the reduction of the feature set increases the learning accuracy of neural networks for intrusion detection. This experiment was done on a real dataset and the obtained results are better than the previous works done on different parameters that are highlighted in Section II below.

Keywords
Anomaly Detection, ANN, Clustering, Genetic Algorithm, Intrusion Detection.

1. INTRODUCTION
The digital network is a revolutionary architecture in the field of communication which challenges attackers to develop new threats to exploit the vulnerability in the network. Traditional techniques of prevention from these attacks such as firewalls and antivirus have their limitations and can't handle intelligently launched attacks.

Thus this communication of information needs high levels of protection from different types of cyberattacks, which is where the Intrusion Detection System plays an important role in detecting and generating an alarm for a suspicious session.

A typical Intrusion Detection System consists of basic rules to detect an intrusion activity and perform a counteraction based on the type of attack [1].

IDS are classified into two categories; Anamoly Based intrusion detection and misuse Based Intrusion detection.

The anomaly-based intrusion detection system can identify unseen attacks based on the behavior of an attack, but this sometimes leads to a false alarm when the session nature is falsely detected as an intrusion [2]. In the case of a Misuse based intrusion detection, a pattern of intrusions and normal sessions are used to train a mathematical model or a machine to detect and identify only a specific type of intrusion pattern. This leads to failure as any new intrusion class is not identified by the IDS [3]. This failure would call for the network administrator to either update or replace the Misuse based IDS to increase the true alarm rate.

Intrusion detection systems are classified based on their implementation; either on the host or a network [4]. If an IDS is implemented on a Host then it is considered as a Host Based Intrusion detection system, while an IDS that is implemented on a network is considered as a Network-based Intrusion Detection System [5].

This paper proposes a Misuse based Intrusion Detection System implemented on the network where the NSL-KDD dataset has been used for training and testing.

The remainder of this paper has been categorized into sections where Section II briefly highlights the work done by other researchers in the field of intrusion detection, section III explains the methodology of the proposed intrusion detection system with the help of a block diagram, section IV highlights the experimental values of a proposed model and an existing model [11] and finally, section V concludes the paper and summarizes the outcome of the proposed model.

2. RELATED WORK
Vajayanandet al. [6] have proposed an intrusion detection system where features were selected from the training dataset by utilizing a genetic algorithm and mutual information. Selected features from the training dataset were used to train the Support Vector Machine (SVM), model. This work resulted in showing that the reduction in features by the genetic algorithm increased their learning and classification accuracy of the SVM model.

Kabir et al. [7] have developed an OALSSVM model (Optimum Allocation Least Square Support Vector Machine). In their work, the optimum allocation term selected a session from the training or testing section of the dataset. These selected sessions or samples were used to train the support vector machine model. Thus, the output of the proposed OALSSVM depended on the selected session which increased the accuracy of the intrusion detection.

Chuanlong Yin et. al. [8] have proposed an artificial neural network-based intrusion detection system where the model works well for a two and multiclass partition. The input training dataset was pre-processed into the numeric part and this preprocessed data was normalized in the subsequent steps as some of the numeric values were very high or very low.

The normalized input dataset was passed through the recurrent neural network model for training the intrusion detection system. Pre-processing steps are similar to the ones in the testing phase as well.

Moukhafi et al. [9] have proposed a feature reduction model for increasing the detection accuracy of an intrusion in the network. This paper utilized a particle swarm optimization genetic algorithm for the selection of features from the input dataset as per the number of classes for detection. Selected features from the training dataset were used to train the support vector machine. This hybrid genetic and SVM model works well to detect DOS attacks.

Kaiyuan Jiang et. al. in [11] the author has proposed a training dataset optimization technique for intrusion detection where a minority of session class was increased artificially by Synthetic

Minority Over-sampling Technique to enhance the detection accuracy. This optimized dataset was used to train the convolution neural network. This work shows that the use of an optimized dataset increases the accuracy of the intrusion detection system.

3. METHODOLOGY

The entire work has been divided into two different modules based on the steps of training of the intrusion detection system.

Fig. 1below illustrates the entire proposed training model. The first module consists of selecting the features for the training of the neural network bypassing the input dataset into the Fish Schooling Genetic Algorithm. The second module consists of the learning of a selected feature data set that was done by error backpropagation neural network for the Intrusion detection system.

1. Pre-Processing

Here, the information like the protocol type, socket type, etc. are removed. As the presence of this information increases the confusion for the clustering process.

The raw data session is as follows;

{0,tcp,ftp_data,SF,491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,150,25,0.17,0.03,0.17,0.00,0.00,0.00,0.05,0.00,normal,20}

After applying pre-processing, the session is as follows;

{0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,150,25,0.17,0.03,0.17,0.00,0.00,0.00,0.05,0.00,normal}

PD←PreProcess_Dataset(ID, n) Eq. 1

In the above equation, ID is the Input dataset having an' number of sessions and PD is a Pre-processed dataset.

2. Generate Population

In this FSGA(Fish Schooling Genetic Algorithm), the set of probable solutions act like fish. This collection of Fish is termed as Population and consists of a set of features in the form of 1 or 0, where 1 denotes that the feature is available and 0 denotes that the feature has not been selected in the probable solution.

The Fish generation function is shown in Eq. 2.

P←Generate_Population(M, PD)-----Eq. 2

Where M is the number of probable solution and PD is the pre-processed dataset.

3. Fitness Function

The estimation of the fitness value of the probable solution depends upon the input feature training matrix as per P.

Hence this function performs the training of EBPNN from the input feature matrix. Then the testing of the trained neural network for the same input feature matrix is performed to find the accuracy of the training chromosome feature set.

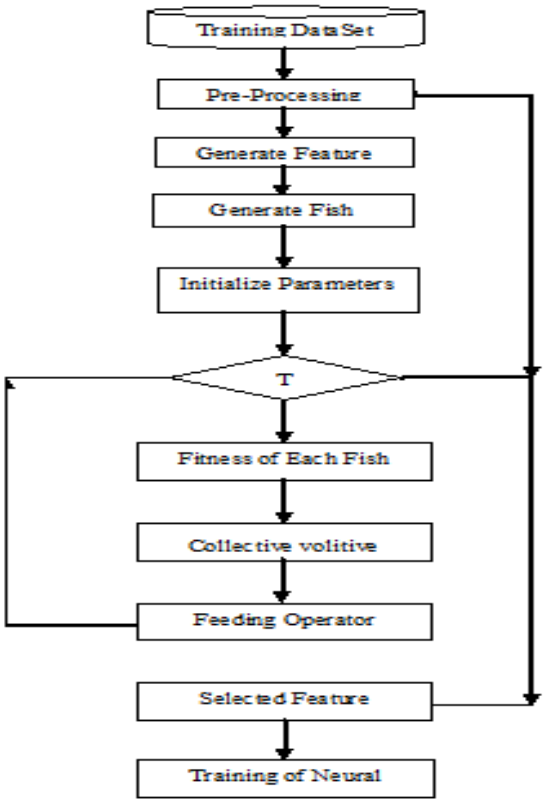

Fig. 1. Proposed Training Model

4. Parameters

Svol: Maximum displacement performs in an operator.

X(t): Random position of a document in the clusters for the iteration.

W(t): Summation of fitness value.

5. Collective volitive Movement

All fishes either reduce their distance if their barycenter is found, otherwise the dispersion movement is applied. Hence, if the sum of the fitness value of t+1 iteration is higher than the last't' iteration summation fitness value then, the movement operation is done by case1 otherwise, case2 applies.

Case1

$$\boldsymbol{x(t+1) = x(t) - S_{vol} \times R \times \left(\frac{x(t) - B(t)}{Euclidean(x(t), B(t))}\right)}$$

Case2

$$x(t+1) = x(t) + S_{vol} \times R \times \left(\frac{x(t) - B(t)}{Euclidean(x(t), B(t))}\right)$$

$$B(t) = \frac{\sum_{i=1}^{N} x(t) W(t)}{\sum_{i=1}^{N} W(t)}$$

6. Feeding Operator

$$W(t+1) = W(t) + \left(\frac{\Delta f}{max(|\Delta f|)}\right)$$

$$\Delta f = F(t+1) - F(t)$$

$$S_{vol} = S_{vol} - \frac{S_{vol}}{Iteration_{max}}$$

7. Crossover

In this work, population P was modified as per x(t) values which range from 0 to n, where n denotes the maximum number of feature values. Crossover operation generates a new chromosome and the selection of the good parent depends on the fitness value. Here, the fish with the best fitness values acts as a good parent throughout the crossover operation so that the other set of chromosomes undergo a crossover with the parents that have good chromosomes. In the crossover, one feature is randomly selected from the good chromosomes and its presence or absence status is replaced in other chromosomes to generate new chromosomes.

Features	F1	F2	F3	.....	Fn
Good Chromo	0	0	1		1
Chromo	1	1	1		0
If the random position is 2,and a feature is absent in the good chromosome then, in the new chromosome that feature is absent as well while the other features remain the same.					
New Chromo	1	0	1		0

8. Population Updation

As the crossover changes, the chromosome of the population and the retention of this fish depends on their fitness value. If the new fish has a good fitness value when compared to the parent fish's fitness value then, this means that the new fish was included in the population. If not, this means the parent fish will continue among the population. Hence in any situation, the population will never change from 'M'.

9. Error Back Propagation Neural Network

In this module elected dataset selected feature as per the genetic algorithm was utilized to train the EBPNN. Then, this trained neural network was utilized for testing the unknown attack session.

10. Input Feature

As per the Fish schooling genetic algorithm, the selected feature from the dataset was used to train the neural network. Considering the first input feature vector which consists of numeric values is arranged in the input matrix while the second desired output vector consists of the classes of sessions which was present in the dataset.

10.1. Training of Error Back Propagation Neural Network (EBPNN): The training input matrix passed in the input layer of the neural network while the desired output makes proper weight adjustment in the network. With a certain number of iterations or epochs, the neural network will get trained.

11. Proposed Intrusion detection Algorithm

Input: D // Dataset

Output: TNN, SF // Trained Neural Network,

SF: Selected Feature

- PD←Pre_Process(D) // Preprocessed Dataset
- P←Generate_Population(M, PD)
- IP←Initialize_Parameter()
- Loop 1:Itr // Itr : number of iteration
- F←Fitness(P, PD)
- IP←Volatile_Movement(IP)
- P←Crossover(IP, P, F)
- P←Update_Population(P)
- EndLoop
- F←Fitness(P, PD)
- SF←Best(F)
- TNN←EBPNN(PD, SF)

12. Testing of EBPNN

In this step, the input session is preprocessed as it was done in the training module. The selected feature vector was created for input in the neural network. Finally, the feature vector is input in the EBPNN which gives thesession intrusion class.

Now, the output is analyzed to determine whether the specified class is desired or not.

4. EXPERIMENT AND RESULTS

The Data Set For the evaluation of the entire work is NSL KDD [12] which has been explained in the above sections. The collection of all evaluating vectors look similar where the numeric values are used for feature learning followed by the respective corresponding class at the end of each vector.

The pre-processing step and its requirements have been explained in the sections above.

1. Evaluation Parameter

To test our results, this work uses the following measures Precision, Recall, and F-score. These parameters are dependent on the TP, TN, FP, and FN.

$$\Pr ecision = \frac{True\_Positive}{True\_Positive + False\_Positive}$$

$$\mathrm{Re}call = \frac{True\_Positive}{True\_Positive + False\_Negative}$$

$$F\_Score = \frac{2 * \Pr ecision * \mathrm{Re}call}{\Pr ecision + \mathrm{Re}call}$$

2. Results

Table 1. Precision value-based comparison of IDS

Data-Set Size	Intrusion Detection System	
	Previous Work [11]	Proposed Model
5000	0.851586	0.973943
7000	0.850054	0.973712
15000	0.846459	0.972634

Table 1 above shows that the proposed hybrid models FSGA and EBPNN have an increased precision value. This enhancement was achieved by the use of FSGA for classification of both; normal sessions and attack sessions.

Table 2. Recall value-based comparison of IDS.

Data-Set Size	Intrusion Detection System	
	Previous Work [11]	Proposed Model
5000	0.931049	0.977635
10000	0.930964	0.976331
15000	0.929145	0.977125

Table 2 above shows that the proposed hybrid model FSGA and EBPNN have an increased recall value. This enhancement was achieved by the use of EBPNN for further identification of intrusion classes. The use of selected feature values for the training of the neural network has resulted in an improvement of the parameter values.

Table 3. F-measure value-based comparison of IDS

Data-Set Size	Intrusion Detection System	
	Previous Work [11]	Proposed Model
5000	0.889546	0.975785
10000	0.888671	0.97502
15000	0.885877	0.974874

Table 3 above shows that the proposed hybrid model FSGA and EBPNN have increased the precision value. This enhancement was achieved by the use of FSGA for classification of both; normal sessions and attack sessions.

Table 4. Accuracy value-based comparison of IDS

Data-Set Size	Intrusion Detection System	
	Previous Work [11]	Proposed Model
5000	0.888022	0.974405
10000	0.886588	0.973432
15000	0.884013	0.973336

Table 4 above shows that the proposed hybrid model FSGA and EBPNN have increased accuracy values. This enhancement has been achieved by the use of EBPNN for further identification of intrusion classes. The use of selected feature values for the training of the neural network has improved the parameter values.

Table 5. Execution time (Seconds) value-based comparison of IDS

Data-Set Size	Intrusion Detection System	
	Previous Work [11]	Proposed Model
5000	57.0575	56.1001
10000	73.6251	69.4306
15000	89.8248	84.9307

Table 5above shows that the proposed hybrid model FSGA and EBPNN has reduced the testing time value. This enhancement has been achieved by the use of EBPNN for further identification of intrusion classes. The use of the selected feature values for the training of the neural network has reduced the parameter values.

5. CONCLUSION

Detection of intrusion in a network is an important issue as several researchers have proposed different models for its detection. This paper has proposed FSGA for the reduction of input training feature dataset. Hence, training of EBPNN by this reduced feature set has increased the detection of intrusion in a network session. This paper highlights that the proposed hybrid model FSGA and EBPNNreduced the testing time value. This experiment was performed on a real dataset NSL-KDD while the comparison was done by existing methods. Results show that the proposed FSGA&NN model increases the precision by 12.74%, while accuracy was enhanced by 8.98%. In the future, the researcher can reduce the training feature vector by using other genetic algorithms.

6. REFERENCES

[1] Koushal Kumar, Jaspreet Singh Batth "Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms" International Journal of Computer Applications (0975 – 8887) Volume 150 – No.12, September 2016

[2] Bharot, N., Verma, P., Sharma, S. Gupta. "Distributed Denial-of-Service Attack Detection and Mitigation Using Feature Selection and Intensive Care Request Processing Unit". Arab J Sci Eng 43, 959–967 (2018).

[3] PremansuSekhararath, Manisha Mohanty, Silva Acharya, Monica Aich "optimization of ids algorithms using data mining technique" International Journal of Industrial Electronics and Electrical Engineering, ISSN: 2347-6982 Volume-4, Issue-3, Mar.-2016

[4] Nitesh Bharot, S. Gupta" Mitigation Distributed Denial of Service Attack in Cloud Computing Environment using Threshold based Technique". Indian Journal of Science and Technology. Vol9 (38), Oct 2016.

[5] YU-XIN MENG," The Practice on Using Machine Learning For Network Anomaly Intrusion Detection" Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, 2011 IEEE.

[6] R. Vijayanand, D. Devaraj, and B. Kannapiran, ''A novel intrusion detection system for wireless mesh network with

hybrid feature selection technique based on GA and MI,'' J. Intell. Fuzzy Syst., vol. 34, no. 3, pp. 1243–1250, 2018.

[7] E. Kabir, J. Hu, H. Wang, and G. Zhuo, ''A novel statistical technique for intrusion detection systems,'' Future Gener. Comput. Syst., vol. 79, pp. 303–318, Feb. 2018.

[8] Chuanlong Yin, Yuefei Zhu, Jinlong Fei, And Xinzheng He. "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks" current version on November 7, 2017.

[9] M. Moukhafi, K. El Yassini, and S. Bri, ''A novel hybrid GA and SVM with PSO feature selection for intrusion detection system,'' Int. J. Adv. Sci. Res. Eng., vol. 4, pp. 129–134, May 2018.

[10] Liu Hui, CAO Yonghui "Research Intrusion Detection Techniques from the Perspective of Machine Learning" Second International Conference on MultiMedia and Information Technology 2010 IEEE.

[11] Kaiyuan Jiang, Wenya Wang, Aili Wang, And Haibin Wu. "Network Intrusion Detection Combined Hybrid Sampling With Deep Hierarchical Network". IEEE Access February 24, 2020.

[12] https://github.com/defcom17/NSL_KDD/blob/master/Original%20NSL%20KDD%20Zip.zip