Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

An effective approach on Intrusion Detection using SVM and ANN

Karuturi Venkata AbhiRam
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India.

Kancharla Sai Prajith
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India.

Pappuri Jithendra Sai
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India.

Adith Sreeram A S
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India.

***Abstract*—The scaling growth of voluminous data and its impact over the CIA (confidentiality-integrity-availability) factor in Industry 4.0 era have increased the importance of network security. To effectively secure the networks, IDS has been put into deployment as a second line of defense in recent times. IDS is an application or device that monitors and analyzes information and network-flow to identify anomalous/intrusive data within the system or network. It may also detect the malicious activities that sometimes can't be identified by network/host firewall. High volume & high-speed traffic of the network generated within a specific environment have made the intrusion detection process difficult, that motivated researchers to develop ML-based statistical intrusion detection models. Hence, this paper proposes a new technique with 2 stages namely selecting features and classification. Performance of the model has been validated with NSL-KDD dataset with standard measures.**

***Keywords*—IDS,MGA,SVM,ANN,MLP**

I. Introduction

The cybercrimes are increasing day by day due to advancement in technology. The data should be protected efficiently in order to avoid attacks. By monitoring the data, IDS identify different types of attacks. The attacks are broadly classified into two types: host-based attacks and network-based attacks. In Host-based attacks the attacker targets a particular host and tries to gain unauthorized access to that particular host. In Network-based attacks the attacker intrudes into a particular network and captures the packets in that network. stability is an important measure for assessing the IDS. Initially researchers used rules-based expert systems and statistical methods for designing IDS. These methods gave poor results for larger datasets. Later on, ML methods have been used for overcoming the problem with statistical methods.

There are two working stages in the project:

1. Selecting Features
    - a. Preprocessing
        - 1.Numericalization
        - 2.Normalization
    - b. Initialization
    - c. Calculating the fitness
    - d. Selecting the parents
    - e. Crossover, mutation

2. Classification (ANN)
    - a. Agent Representation
    - b. Agent evaluation
    - c. Agent's update

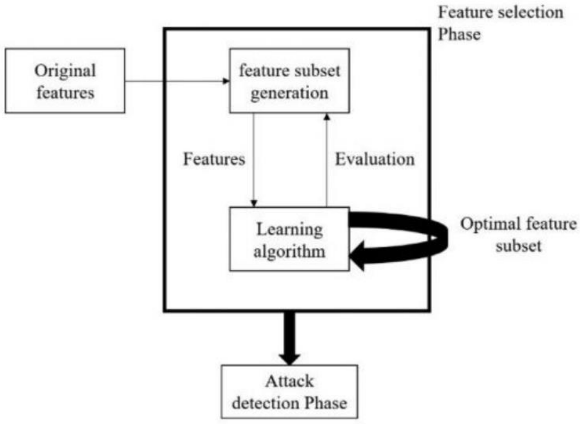

Fig.A. General Schema of SVM-MGA-ANN

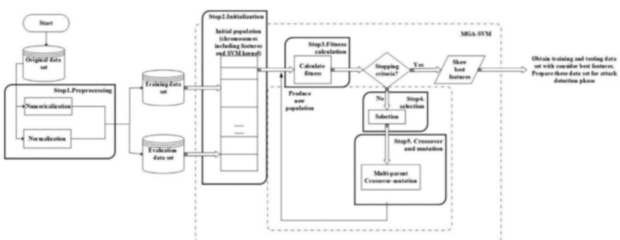

Fig.B. Work Flow Chart of SVM-MGA

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

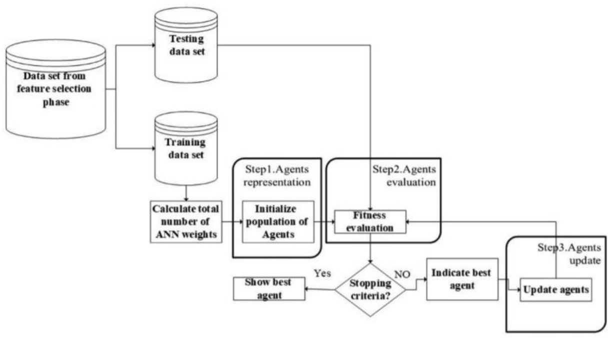

Fig.C. Work Flow Chart of ANN

II. METHODOLOGY

A. Datasets

The Datasets used in this project are NSL-KDD, Attack Dataset. NSL-KDD is used for anomaly detection. It has 25192 records, and 42 features. The Attack dataset contains 2 features.

B. Preprocessing

The dataset has to be preprocessed before it can be sent through a model. If the dataset is not preprocessed, the model will not fit correctly and there might be a drastic drop in accuracy. It has to be cleaned and checked for missing values, null values and noise. The preprocessing techniques applied on the dataset in this project are Numericalisation and Normalization.

Data Numericalization is converting the non-numerical features into numerical features.

Each feature in the dataset might have values in different ranges. While applying Machine Learning algorithms the features having larger value may dominate the features having lower value and this affects the accuracy. In order to avoid this the data is normalized in the range [-1,1].

C. Algorithms

This project has 2 stages namely Selecting features and classification. For selecting features SVM and MGA is used. For classification ANN is used.

SVM is used for classifying the data. It works by transforming the data into relatively high dimension and then finding the hyper-plane that differentiates the classes. The data is transformed to higher dimensions using kernel functions. The kernel functions only calculate the relationships between every pair of points as if the points are in higher dimension, they don't actually do the transformation. This is called kernel trick. The kernels are of 2 types namely polynomial kernel and radial kernel. Radial kernel is used in this project.

Genetic Algorithm is an abstraction of real biological evolution. It focuses mainly on optimization. It is based on the concept of survival of the fittest. In general, we extract 2 chromosomes (parents) from the population set based on their fitness values and generate offspring by performing crossover and mutation techniques. But in MGA we use 3 chromosomes (parents). In this project we are using MGA to extract best features. The fitness calculation is given by

Fitness= α * (1 - Detection rate) + β * |S_F| / |T_F|

Where,
α is random value in between 0 and 1
β = 1-α
Detection rate is accuracy of SVM algorithm
S_F is total no of genes having value greater than 0.5 in the chromosome
T_F is total no features Selection of the chromosome is based on the roulette wheel mechanism

Selection = 1/fitness

The offspring calculation is given by
Crossover calculation

O1= P1 + θ * (P2 - P3)
O2= P2 + θ * (P3 - P1)
O3= P3 + θ * (P1 - P2)

Mutation calculation

M = O3 + θ * (P3 - O3)

where,
O1, P1 are offspring 1 and parent 1
O2, P2 are offspring 2 and parent 2
O3, P3 are offspring 3 and parent 3
M is mutated offspring.

Here we perform mutation in order to avoid repetition of offspring.

The MLP is a class of ANN. The flow of data is always in the forward direction so it is called a feedforward neural network. It contains a series of nodes or neurons which uses weights as parameters for the transfer of inputs through different layers. MLP uses non-linear activation function for all its layers except for input layer in order to change the inputs to nonlinear outputs. For increasing accuracy or performance of the model MLP uses supervised learning technique called backpropagation. It calculates the error and adjusts the weights.

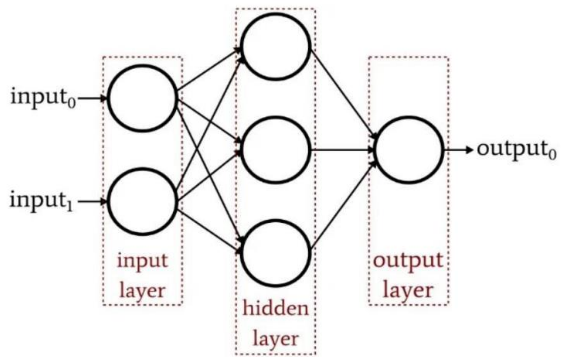

Fig.D. MultiLayer Perceptron

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

III. RESULTS AND DISCUSSION

Firstly, SVM machine learning algorithms is used to cross-validate and evaluate the indicators. The original training set is divided into training set and testing set according to the ratio of 75:25 and the values of evaluation metrics like accuracy, precision, recall and F1 Score of SVM are given below in table-1

Algorithm	Accuracy	Precision	Recall	F1 Score
SVM	97.824	97.698	97.824	97.704

Table 1. Results for SVM classifier

```
In [30]: # MLP part

         from sklearn.neural_network import MLPClassifier
         from sklearn.metrics import classification_report, confusion_matrix
         mlp = MLPClassifier(hidden_layer_sizes=(10, 10, 10), max_iter=1000)
         mlp.fit(X_train1, y_train1.ravel())

Out[30]: MLPClassifier
         MLPClassifier(hidden_layer_sizes=(10, 10, 10), max_iter=1000)

In [31]: predictions = mlp.predict(X_test1)
         print('prediction: ')
         print(predictions)

         prediction:
         ['normal' 'normal' 'normal' ... 'normal' 'dos' 'dos']
```

Fig.E. MLP Predictions

The predictions of the dataset using MLP which was run in google colab are shown in the above figure E

Algorithm	Accuracy	Precision	Recall	F1 Score
MLP	95.125	94.199	95.125	94.611

Table 2. Results for MLP classifier

Methods	Number of features	Precision	Recall	F1-Score
DT	41	0.897	0.904	0.900
GA-SVM	10	0.930	0.934	0.932
Chi-SVM	31	0.891	0.910	0.905
GA-ANN	13	0.900	0.920	0.913
PSO-ANN	41	0.912	0.927	0.918
SVM-MGA-ANN	**4**	**0.955**	**0.953**	**0.954**

Table 3.1 The number of features selected

Above Table 3.1 shows the number of features selected from 41 original features by each of the compared methods which was previously done by the other researchers along with the corresponding values of precision, recall and F1-Score. GA-ANN, GA-SVM, and the proposed method have lower number of selected features because they follow a feature selection approach. The best four features selected by the proposed method are presented in the below table 3.2. The result in table 3.1 shows that the proposed method is the best performing method on all criteria.

Number	Name of feature	Brief description
1	Service	Model of network service on the target
2	Flag	Indicates normal or error condition of connection
3	Protocol type	Type of the protocol
4	dst_host_same_srv_rate	Percentage of sessions to the same service

Table 3.2 Best Features

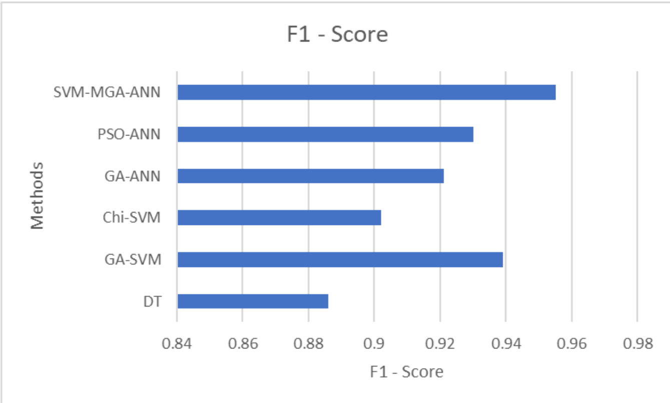

Fig-F. F1 Score Comparison

Fig – F Shows the F-measure comparison results for the considered methods and proposed methods. The results reveal that the proposed method achieves the best value of F-measure. In addition, it can be seen from Fig – F, that PSO – ANN outperforms Chi-SVM and DNA because the feature selection approach used in the latter method cannot select proper features and remove redundant ones, therefore it can't help classifiers to detect attacks. Whereas PSO-ANN with no feature selection approach can properly detect attacks. Also, PSO helps ANN to escape from the local optimum.

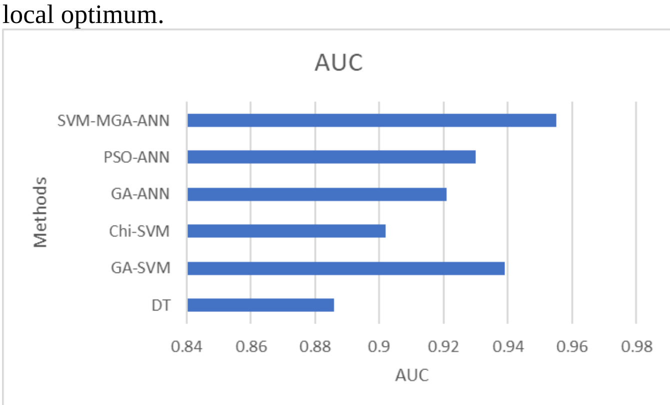

Fig.G. AUC Score Comparison

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

Fig G shows AUC for different methods AUC of DT, and Chi-SVM is lesser than the other methods which means their performance is comparatively low. SVM requires extensive training time and performs well with properly pre-processed datasets. As NSL-KDD is a preprocessed dataset, this results in a higher performance of the proposed method compared to GA-ANN and DT in detecting attacks. ANN is a highly weight-depended classifier and performs well for classifying rough datasets like image data when proper values for its weights are set.

IV. Conclusion

In the paper, a new hybrid method SVM-MGA-ANN has been proposed for attack detection. Multi parent crossover and mutation has been ensembled to increase the performance of Genetic Algorithm (GA). The back-propagation technique used in MLP has significantly increased the performance of ANN. The overall performance of the model is assessed by standard measures such as accuracy, precision. The proposed method has given an accuracy of 95.12% and it takes less time for training and testing. Overall experimentation was carried out in google co-laboratory environment. Optimization techniques like Particle Swarm Optimization (PSO), and Hybrid Gravitational Search (HGS) will further be studied and experimented as a future work for optimizing the ANN model.

References

[1] K.L. Fox, R.R. Henning, J.H. Reed, R.P. Simonian, “A neural network approach towards intrusion detection,” 1990.

[2] J. Ryan, M.-J. Lin, R. Miikkulainen, Intrusion detection with neural networks, Adv. Neural Inf. Process Syst. (1998) 943–949.

[3] G. Wang, J. Hao, J. Ma, L. Huang, A new approach to intrusion detection using artificial neural networks and fuzzy clustering, Expert Syst. Appl. 37 (2010) 6225–6232.

[4] C. Manikopoulos, S. Papavassiliou, Network intrusion and fault detection: a statistical anomaly approach, IEEE Commun. Mag. 40 (2002) 76–82.

[5] A.A. Elngar, A. Dowlat, F.F. Ghaleb, A fast accurate network intrusion detection system, Int. J. Comput. Sci. Inf. Secur. 10 (2012) 29.

[6] Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, K. Dai, An efficient intrusion detection system based on support vector machines and gradually feature removal method, Expert Syst. Appl. 39 (2012) 424–430.

[7] V.M. Hashemi, Z. Muda, W. Yassin, Improving intrusion detection using genetic algorithm, Inf. Technol. J. 12 (2013) 2167.

[8] F. Kuang, S. Zhang, Z. Jin, W. Xu, A novel SVM by combining kernel principal component analysis and improved chaotic particle swarm optimization for intrusion detection, Soft Comput. 19 (2015) 1187–1199.

[9] B. Aslahi-Shahri, R. Rahmani, M. Chizari, A. Maralani, M. Eslami, M. Golkar, et al., A hybrid method consisting of GA and SVM for intrusion detection system, Neural Comput. Appl. 27 (2016) 1669–1676.

[10] T. Dash, A study on intrusion detection using neural networks trained with evolutionary algorithms, Soft Comput. 21 (2017) 2687–2700.

[11] I.S. Thaseen, C.A. Kumar, Intrusion detection model using fusion of chi-square feature selection and multi class svm, J. King Saud Univ.-Comput. Inf. Sci. 29 (2017) 462–472.

[12] N.T. Pham, E. Foo, S. Suriadi, H. Jeffrey, H.F.M. Lahza, Improving performance of intrusion detection system using ensemble methods and feature selection, in: Proceedings of the Australasian Computer Science Week Multiconference, 2018, p. 2.

[13] S. Aljawarneh, M. Aldwairi, M.B. Yassein, Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model, J. Comput. Sci. 25 (2018) 152–160.

[14] O.F. Rashid, Z.A. Othman, S. Zainudin, Features selection for intrusion detection system based on DNA encoding, in: Intelligent and Interactive Computing, Springer, 2019, pp. 323–335.