International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET

Deep Neural Network based Intrusion Detection system using Principal Component Analysis Techniques

[1]B. Indra Devi, [2] Yarlagadda Sivaiah, [3] Kondaveeti Dihitha, [4] Vanteddu Eshitha, [5] Sayyad Irshad

[1]Assistant Professor, [2,3,4,5] UG Student

Department of CSE-AI & ML, Sri Vasavi Institute of Engineering & Technology, Nandamuru, Andhra Pradesh, India

ARTICLEINFO

Article History:
Accepted: 15 April 2024
Published: 25 April 2024

Publication Issue :
Volume 11, Issue 2
March-April-2024
Page Number :
467-472

ABSTRACT

With the evolution in wireless communication, there are many security threats over the internet. The intrusion detection system (IDS) helps to find the attacks on the system and the intruders are detected. Previously various machine learning (ML) techniques are applied on the IDS and tried to improve the results on the detection of intruders and to increase the accuracy of the IDS. This project has proposed an approach to develop efficient IDS by using the principal component analysis (PCA) and the random forest classification algorithm. Where the PCA will help to organise the dataset by reducing the dimensionality of the dataset and the random forest will help in classification. Results obtained states that the proposed approach works more efficiently in terms of accuracy as compared to other techniques like SVM, Naïve Bayes, and Decision Tree. The results obtained by proposed method are having the values for performance time (min) is 3.24 minutes, Accuracy rate (%) is 96.78 %, and the Error rate (%) is 0.21 %.

Keywords : IDS, Knowledge Discovery Dataset, PCA, Random Forest.

I. INTRODUCTION

Nowadays, the involvement of the internet in normal life has been increased rapidly. The internet has made a crucial place in everyone's life. The use of the internet has become very crucial for everyone. So with the increase in the use of the internet for personal activities, it is also necessary to keep secure the system from malicious activities. Different attacks are seen on the system or the network. The attacks like a black hole, grey hole, wormhole etc. are seen on the network system. These attacks are to steal the information from the system or to corrupt the data present over any system. To make misuse of the data, the intruders attack the system in various ways, some of the attacks are DoS, probe, snort, r2l etc. So to prevent the system from such attacks, the intrusion detection system was introduced. IDS keep track of attacks on the system and to prevent the system from these attacks. So to detect such attacks, the various works have done earlier by using various techniques. Here an intrusion detection system that makes use of the principal

component analysis is used along with the random forest technique. Both the methods work for a special purpose, where the PCA gives the granularity in the data, and the random forest helps the classification between the nodes for attack.

With the evolution in wireless communication, there are many security threats over the internet. Different attacks are seen on the system or the network. The attacks like a black hole, grey hole, wormhole etc. are seen on the network system. These attacks are to steal the information from the system or to corrupt the data present over any system. To make misuse of the data, the intruders attack the system in various ways, some of the attacks are DoS, probe, snort, r2l etc. So to prevent the system from such attacks, the intrusion detection system was introduced. IDS keep track of attacks on the system and to prevent the system from these attacks.

II.RELATED WORK

Authors here presented a mechanism to design the IDS for the IoT that is based on the classification of the traffic by making the use of deep learning model. They performed the binary and multi-class classification. The obtained accuracy for the presented system is high. The authors here gave a solution for the IDS as they applied the SVM and Naïve Bayes algorithms and proved that the SVM works better than the Naïve Bayes method. They carried the experiment on the KDD dataset, and they also give the results in terms like detection and false alarm rate. In this paper, the authors performed three different experiments. They applied the feature selection as well in the analysis. Also showed the naïve Bayes, adaptive boost and partial decision tree. They analysed all techniques for intrusion detection. In this paper, the authors have evaluated that the Artificial neural networks with the feature selection technique will provide better results as compared to the Support vector machine technique. They used NSL-KDD dataset for the experiment. The given approach worked well. Here the authors presented a review on the intrusion detection systems, which uses a machine-learning algorithm. The authors provided various machine learning a lgorith m's comparison based on their performance. They evaluated the survey based on the detection rates and false alarm rates. Authors have presented an approach for intrusion detection, which uses logistic regression and belief propagation. And the proposed method has proved that it provides better average detection time as compared to earlier techniques. The authors used an in-depth learning approach for the feature extraction from the dataset. They tried to extract the features from dataset to make a dataset efficient for use and in this way, they decided to provide better input to the intrusion detection system.

Intrusion could be a period of time that provides to urge into the machine with statistics within the machine. This intrusion into any machine can also damage the hardware of the machine. It's grown to be a motivating period to save lots of you the machine. This intrusion inner any machine might be managed or perhaps retaining the song of this intrusion could also be achieved with the help of the IDS. the various kinds of intrusion structures are used earlier, however, with inside the tip, the accuracy worries are apparent in each approach used.

The phrases, inclusive of detection price and therefore the fake alarm price, are analyzed for the assessment of the accuracy of the machine. These phrases need to be with inside the way that the fake alarm price needs to be minimized and therefore the development with inside the detection price has got to be there with inside the machine. therefore the random woodland at the sting of the PCA is administrated.

Random Forest: Random Forest is that the prevalent supervised technique. it's useful for mainly doing classification challenges and also regression challenges. RF is one amongst the classifiers which holds multiple decision trees in each subset of an assumed data set and computes the everyday value that enhances prediction accurateness for the dataset. The random forest doesn't depend upon decision trees. Instead, it gets a prediction

from every tree so forecasts the last result which is made upon polls of prevalence estimations. The more trees within the forest, the upper the accuracy and avoid overfitting problems. it's supported the ensemble technique concept, which mixes multiple classifiers to unravel a thorny problem and improves model performance.

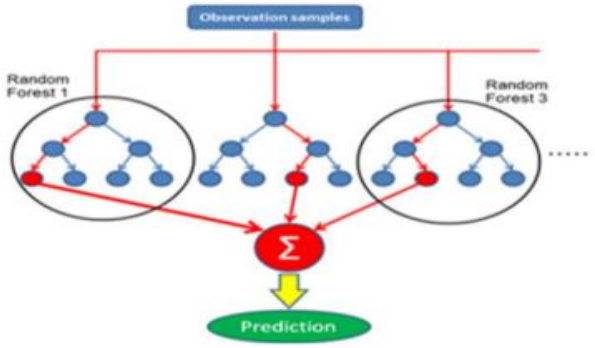

Fig 1: Random Forest Technique

III.PROPOSED SYSTEM

We proposed an approach to develop efficient IDS by using the principal component analysis (PCA) and the random forest classification algorithm. Where the PCA will help to organise the dataset by reducing the dimensionality of the dataset and the random forest will help in classification. Results obtained states that the proposed approach works more efficiently in terms of accuracy as compared to other techniques like SVM, Naïve Bayes, and Decision Tree. The results obtained by proposed method are having the values for Accuracy rate (%) is 96.78 %, and the Error rate (%) is 0.21%

Advantages:

- PCA will reduce the dimensionality of dataset, so we can get some more accuracy.
- Lesser training time is required.

3.1 MODULES DESCRIPTION

Pandas: pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

Numpy: NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

MatPlotLib: matplotlib.pyplot is a plotting library used for 2D graphics in python programming language. It can be used in python scripts, shell, web application servers and other graphical user interface toolkits.

Scikit-learn: scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

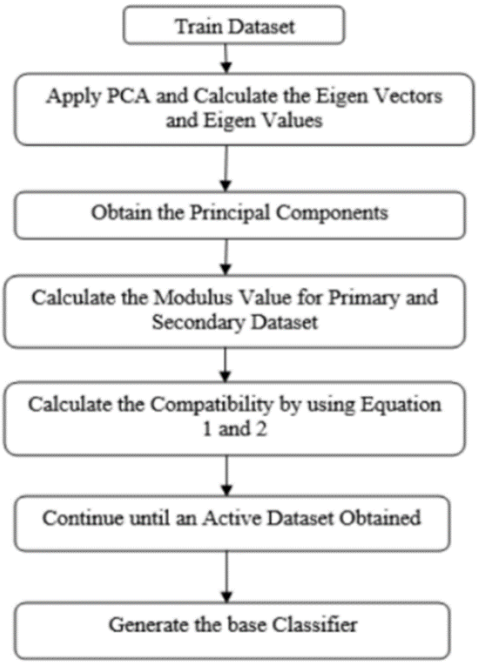

Fig 2: Flow chart for the proposed system

IV.RESULTS AND DISCUSSION

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# fetch the training file
file_path_20_percent = 'KDDTrain+_20Percent.txt'
file_path_full_training_set = 'KDDTrain+.txt'
file_path_test = 'KDDTest+.txt'

#df = pd.read_csv(file_path_20_percent)
df = pd.read_csv(file_path_full_training_set)
test_df = pd.read_csv(file_path_test)
```

Fig 3. Results Screenshot

Data transformations

The first transformations that we'll want to do are around the attack field. We'll start by adding a column that encodes 'normal' values as 0 and any other value as 1. We will use this as our classifier for a simple binary model that idenfities any attack.

```
# map normal to 0, all attacks to 1
is_attack = df.attack.map(lambda a: 0 if a == 'normal' else 1)
test_attack = test_df.attack.map(lambda a: 0 if a == 'normal' else 1)

df['attack_flag'] = is_attack
test_df['attack_flag'] = test_attack

# view the result
df.head()
```

rc_port_rate	dst_host_srv_diff_host_rate	dst_host_serror_rate	dst_host_srv_serror_rate	dst_host_rerror_rate	dst_host_srv_rerror_rate	attack	level	attack_fla
0.88	0.00	0.00	0.00	0.0	0.00	normal	15	
0.00	0.00	1.00	1.00	0.0	0.00	neptune	19	
0.03	0.04	0.03	0.01	0.0	0.01	normal	21	
0.00	0.00	0.00	0.00	0.0	0.00	normal	21	
0.00	0.00	0.00	0.00	1.0	1.00	neptune	21	

Fig 4. Results Screenshot

```
# get the intial set of encoded features and encode them
features_to_encode = ['protocol_type', 'service', 'flag']
encoded = pd.get_dummies(df[features_to_encode])
test_encoded_base = pd.get_dummies(test_df[features_to_encode])

# not all of the features are in the test set, so we need to account for diffs
test_index = np.arange(len(test_df.index))
column_diffs = list(set(encoded.columns.values)-set(test_encoded_base.columns.values))

diff_df = pd.DataFrame(0, index=test_index, columns=column_diffs)

# we'll also need to reorder the columns to match, so let's get those
column_order = encoded.columns.to_list()

# append the new columns
test_encoded_temp = test_encoded_base.join(diff_df)

# reorder the columns
test_final = test_encoded_temp[column_order].fillna(0)

# get numeric features, we won't worry about encoding these at this point
numeric_features = ['duration', 'src_bytes', 'dst_bytes']

# model to fit/test
to_fit = encoded.join(df[numeric_features])
test_set = test_final.join(test_df[numeric_features])
```

Fig 5. Results Screenshot

```
from sklearn.model_selection import train_test_split
# create our target classifications
multi_y = df['attack_map']
test_multi_y = test_df['attack_map']

# build the training sets
multi_train_X, multi_val_X, multi_train_y, multi_val_y = train_test_split(to_fit, multi_y, test_size = 0.3)
```

```
from sklearn.preprocessing import StandardScaler
ss = StandardScaler()
X_train_scaled = ss.fit_transform(multi_train_X)
X_test_scaled = ss.transform(multi_val_X)
y_train = np.array(multi_train_y)
```

```
pca = PCA(n_components=10)
pca.fit(X_train_scaled)
X_train_scaled_pca = pca.transform(X_train_scaled)
X_test_scaled_pca = pca.transform(X_test_scaled)
```

```
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier()
rfc.fit(X_train_scaled_pca, y_train)
print(rfc.score(X_train_scaled_pca, y_train))
```

```
0.9775459287820367
```

Fig 6. Results Screenshot

```
y_pred = rfc.predict(X_test_scaled_pca)
```

```
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix

print(accuracy_score(y_pred, multi_val_y))
```

```
0.9766617273497037
```

```
print(confusion_matrix(y_pred, multi_val_y))
```

```
[[20053    58    28     7    19]
 [  133 13740   588     0     0]
 [   20    14  2809     1     5]
 [    7     0     0     6     0]
 [    0     0     2     0   302]]
```

Fig 7. Results Screenshot

V. CONCLUSION

As the involvement of the systems over the internet increasing rapidly, the security concerns have also seen. The proposed approach deals with the detection of intruders over the internet efficiently. The proposed algorithm has performed well as compared to the previously applied algorithms such as SVM, Naïve Bayes, and Decision Tree. The detection rates and the false error rates can be improved at a great extent by the proposed approach. The dataset used here is the knowledge discovery dataset. The results obtained by our proposed method having the values for Accuracy rate (%) is 96.78 %, and the Error rate (%) is 0.21 %.

VI.FUTURE WORK

In future, we will focus on applying some more machine learning algorithm with combination of PCA algorithm so we can check the deviations in result. With help of hyper-parameter optimization we can try to achieve somewhat more accuracy and thereby trying to minimize the false alarm rate.

II. REFERENCES

[1]. Jafar Abo Nada; Mohammad Rasmi Al-Mosa, 2018 Internat ional Arab Conference on Informat ion Technology (ACIT), A Proposed Wireless Intrusion Detect ion Prevent ion and Attack System

[2]. Kinam Park; Youngrok Song; Yun-Gyung Cheong, 2018 IEEE Fourth Internat ional Conference on Big Data Comput ing Service and Applicat ions (BigDataService), Classificat ion of Attack Types for Intrusion Detect ion Systems Using a Machine Learning Algorithm

[3]. S. Bernard, L. Heutte and S. Adam “On t he Select ion of Decision Trees in Random Forest s” P roceedings of Internat ional Joint Conference on Neural Networks, At lanta, Georgia, USA, June 14-19, 2009, 978-1-4244-3553- 1/09/$25.00 ©2009 IEEE

[4]. A. T esfahun, D. Lalitha Bhaskari, ” Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduct ion” 2013 Internat ional Conference on Cloud & Ubiquitous

Comput ing & Emerging Technologies, 978-0-4799-2235-2/13 $26.00 © 2013 IEEE

[5]. Le, T.-T.-H., Kang, H., & Kim, H. (2019). The Impact of PCA-Scale Improving GRU Performance for Intrusion Detect ion. 2019 International Conference on Platform Technology and Service (PlatCon). Doi:10.1109/platcon.2019.8668960

[6]. Anish Halimaa A, Dr K.Sundarakantham: Proceedings of the Third Internat ional Conference on Trends in Elect ronics and Informat ics (ICOEI 2019) 978-1-5386-9439-8/19/$31.00 ©2019 IEEE "MACHINE LEARNING BASED INTRUSION DETECTION SYSTEM."

[7]. Mengmeng Ge, Xiping Fu, Naeem Syed, Zubair Baig, Gideon Teo, Antonio Robles-Kelly (2019). Deep Learning- Based Intrusion Detect ion for IoT Networks, 2019 IEEE 24th Pacific Rim Internat ional Symposium on Dependable Comput ing (PRDC), pp. 256-265, Japan.

[8]. R. Patgiri, U. Varshney, T. Akutota, and R. Kunde, "An Investigation on Intrusion Detection System Using Machine Learning" 978-1-5386-9276-9/18/$31.00 c2018IEEE.

[9]. Rohit Kumar Singh Gautam, Er. Amit Doegar; 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence) " An Ensemble Approach for Intrusion Detection System Using Machine Learning Algorithms."

[10]. Kazi Abu Taher, Billal Mohammed Yasin Jisan, Md. Mahbubur Rahma, 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)"Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection."

[11]. L. Haripriya, M.A. Jabbar, 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)" Role of Machine Learning in Intrusion Detection System: Review"

[12]. Nimmy Krishnan, A. Salim, 2018 International CET Conference on Control, Communication, and Computing (IC4) " Machine Learning-Based Intrusion Detection for Virtualized Infrastructures"

[13]. Mohammed Ishaque, Ladislav Hudec, 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)"Feature extraction using Deep Learning for Intrusion Detection System."

[14]. Aditya Phadke, Mohit Kulkarni, Pranav Bhawalkar, Rashmi Bhattad, 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)"A Review of Machine Learning Methodologies for Network Intrusion Detection."

[15]. Iftikhar Ahmad , Mohammad Basheri, Muhammad Javed Iqbal, Aneel Rahim, IEEE Access (Volume: 6) Page(s): 33789 – 33795 "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection."

[16]. 98 B. Riyaz, S. Ganapathy, 2018 International Conference on Recent Trends in Advanced Computing (ICRTAC)" An Intelligent Fuzzy Rulebased Feature Selection for Effective Intrusion Detection."