IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116100

Network Intrusion Detection System Using Random Forest and PCA

Kapil Sachan[1], Akshay Pratap Singh[2], Sheela S[3]

[1,2]Final Year B.E.,CSE, Global Academy of Technology, Benagluru, India

[3]Assistant Professor, GAT, Bengaluru, India

Abstract: Due to the advancement of wireless communication, there are several online security risks. The importance of **intrusion detection systems** (IDS) in computer and network security cannot be overstated. The experiment dataset in this research was the **KDDCUP'99** (Knowledge Discovery Dataset) intrusion detection dataset. Due to intrusion detection's fundamental properties, there is still a significant imbalance between the classifications in the dataset, which makes it more difficult to apply machine learning to intrusion detection efficiently. IDS techniques come in a wide variety and yield results with varying degrees of precision. This calls for the creation of an efficient and reliable intrusion detection system. In this paper, a method for creating effective IDS that makes use of the random forest classification algorithm and **principal component analysis** (PCA) is proposed. While **Random Forest** (RF), an ensemble classifier that outperforms other standard classifiers for the accurate classification of attacks, PCA will assist in organizing the information by reducing its dimensionality. Together with the Confusion Matrix, a performance evaluation tool, we have also employed other approaches for model evaluation and selection, including as accuracy, precision, recall, and f-score.

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

I. INTRODUCTION

The increase in network utilization is causing an increase in intrusion activities. Attacks on computer systems have risen in recent years, necessitating the use of an effective intrusion detection system. IDS seeks to detect intrusion in routine data.

Host-based Intrusion Detection System (HIDS) and Network-based Intrusion Detection System are the two forms of IDS:

- network-based systems, which analyze data shared among computers,
- host-based intrusion detection systems look at the data stored on a single computer system.

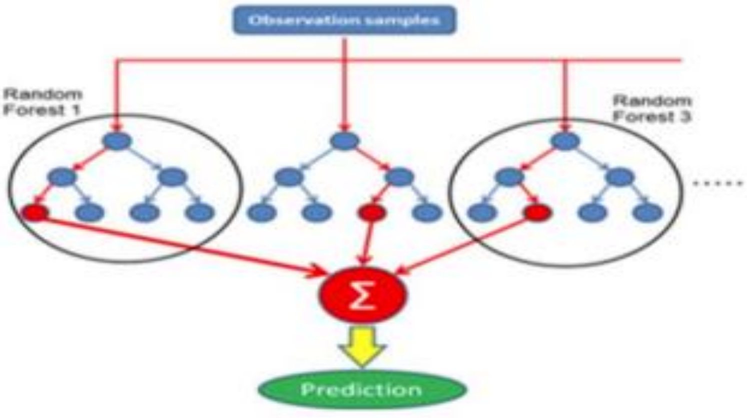


Fig. Random Forest Classifier

As new types of attacks are found on the network, an expert system must be maintained to cope with them. The Intrusion Detection System (IDS) monitors system threats and protects the system from them. Here, the random forest method is combined with an intrusion detection system that makes use of principal component analysis.

To enhance IDS performance, numerous data dimensionality reduction strategies have been used. The most well-known of them is principal component analysis (PCA). By using a few linear combinations of the original variables, PCA aims to describe the variance-covariance structure. Data minimization and interpretation are its main goals. The obtained low-

DOI: 10.17148/IJARCCE.2022.116100

dimensional linear subspace is as free of useless information as feasible. Additionally, it keeps the original space's data structure intact.
The random forest is an ensemble classifier. Various IDS have been created for categorizing attacks. The methods, however, yield lower levels of intrusion detection accuracy. As a result, the abuse component of our suggested system leverages the random forests algorithm for classification in intrusion detection, while the anomaly component is based on the algorithm's outlier detection mechanism. For a more precise and reliable forecast, random forest constructs many decision trees and combines them.

II. RELATED WORK

The term "intrusion" actually refers to trying to access a machine that has data on it. Any machine's hardware may also be harmed by this intrusion. It's now a great opportunity to keep the machine for you. With the aid of the IDS, this infiltration inside of any machine might be controlled or perhaps maintained track of. Although many types of incursion structures were employed earlier, at the end it was clear that each technique had accuracy issues. The phrases, including detection price and phony alarm price, are genuinely examined to determine the machine's level of accuracy. These words must be used in such a way as to minimize the rate of false alarms and ensure that the machine's detection rate is increasing. As a result, the random forest at the edge of the PCA is actually carried out.

Some related works :
1. By P. Natesan et al., multistage filtering for network IDS is proposed.
To identify frequent attacks in networks, authors used enhanced adaboost with a decision tree algorithm and naive bayes.

2. Hybrid IDS based on feature selection was proposed by Ujwala Ravale et al. utilizing K-means and Radio basis function. K-means and SVM are combined in the hybrid approach suggested by the authors. Results from experiments are produced using the KDD Cup 99 dataset.

3.IDS using Random forest and SVM was proposed by Md Al Mehedi Hasan et al.7 Authors developed two models for IDS using SVM and Random forest. The performance of these two approaches are compared based on their accuracy, precision and false negative rate.

4.Network IDS utilizing PSO and Random forest was proposed by Arif Jamal Malik et al. in 2012. To choose the proper attributes for categorizing incursions, binary PSO is utilized. They employ the Random Forest algorithm as a classifier, and their process has two steps. 1) Selection of features 2) Categorization. The suggested strategy was used by the authors in MATLAB.

5.The idea for IDS was put up by Mrutyunjaya Panda et al. as a Hybrid Intelligent Approach.To enhance the performance of the final model, authors combined many classifiers. They employed a 10-fold cross-validation classification approach. The NSL-KDD dataset is used to perform the experimental outcomes.

III. PROPOSED WORK

3.1 Intrusion Detection System :

A malicious, externally produced operational problem is what is meant by IDS. IDS is crucial for spotting different kinds of assaults. IDS's primary objective is to discover intrusions, which is a categorization issue10. IDS can be categorized into a number of attacks, including DOS, probe, U2R, and R2L.

3.2 Random Forest Classifier

An ensemble classifier used to increase accuracy is called random forest (RF). Numerous decision trees make up the random forest. When compared to other conventional classification algorithms, random forest has a low classification error rate. Node splitting criteria include the number of trees, the smallest node size, and the amount of features. The optimal node to split on is chosen randomly when building individual trees in random forest. Where A is the number of characteristics in the data set, this value is equal to √A.

3.3 Principal Component Analysis

The method used to considerably reduce the dimension of the given dataset is principal component analysis. One of the most effective and thorough approaches for minimizing the amount of information is the principal component analysis, which yields the desired outcomes. Using this technique, the given dataset's characteristics are broken down into the desired number of primary components. Because the dataset has a large number of attributes and a large number of

DOI: 10.17148/IJARCCE.2022.116100

dimensions, this approach requires all of the input. By placing the data points on the same axis, this technique lowers the size of the information set. The information points are distributed across the primary components as a result of shifting them along one axis.

The PCA will be performed using the subsequent steps:
1. Take the dataset with all dimensions d.
2. Calculate the mean vector for every dimension d.
3. Calculate the covariance matrix for the entire dataset.
4. Calculate the eigen vectors (e1, e2, e3 … .ed), and eigen values (v1,, v2, v3, ….vd).
5. Perform sorting of eigenvalue in decreasing order and choose n eigenvector with the best eigenvalues to induce a matrix of d*n= M.
6. By using this M form a replacement sample space.
7. The obtained sample spaces are the principal components

Algorithm:
Input: NSL-KDD dataset
Output: Classification of different type of attacks
Step 1: Load the dataset
Step 2: Apply pre-processing technique
Step 3: Partition the data set into training and test
Step 4: Select the best set features using PCA
Step 5: Data set is given to Random forest for training
Step 6: The test data set is then fed to random forest for classification
Step 7: Calculate accuracy, Detection rate, False alarm rate

IV. RESULT AND DISCUSSION

Intrusion Intrusion Detection Systems (IDS) are one of the defences against these attacks, according to an experiment done using unique machine learning models and the KDD dataset. Furthermore, new technologies for next-generation networks, such wireless networks (also known as Wi-Fi), have arisen, necessitating a thorough understanding of the major challenges and limitations associated with the design and deployment of an IDS for such techniques. IDS frequently needs to improve its performance in terms of increasing precision and reducing false alarms. Random Forest Algorithm is able to detect DOS, probe R2L and U2R attacks.

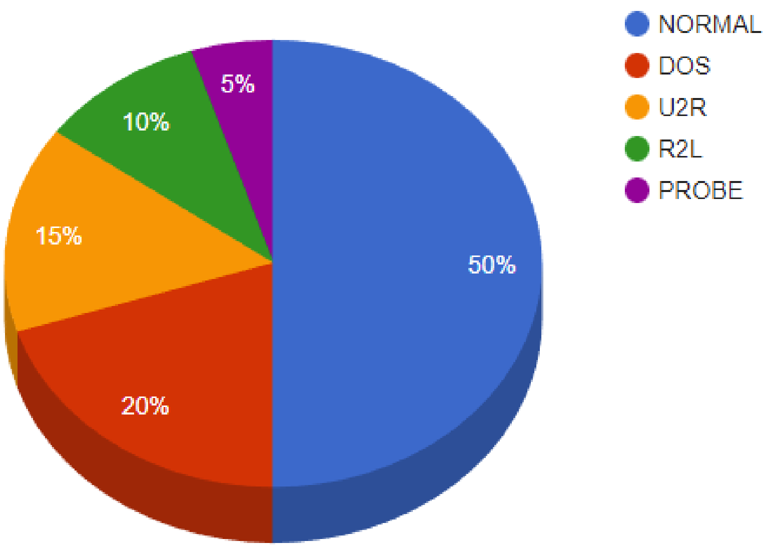
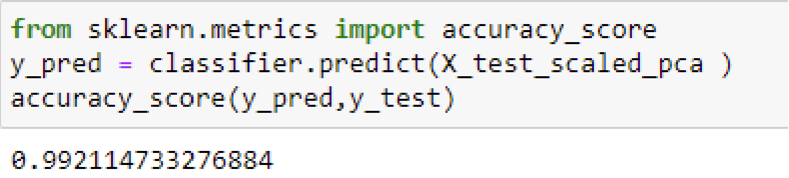


Fig. Piechart for Classification

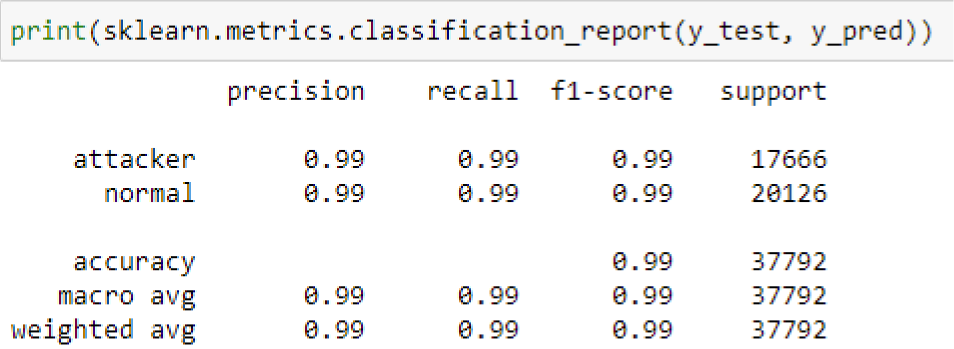


```
from sklearn.metrics import accuracy_score
y_pred = classifier.predict(X_test_scaled_pca )
accuracy_score(y_pred,y_test)

0.992114733276884
```

Fig. Accuracy of Random Forest Classifier

Results from our suggested method have Performance time (min) values of 3.24 minutes, Accuracy rate (percent) values of 99.21 percent, and Error rate (percent values of 0.21 percent.



```
print(sklearn.metrics.classification_report(y_test, y_pred))

              precision    recall  f1-score   support

    attacker       0.99      0.99      0.99     17666
      normal       0.99      0.99      0.99     20126

    accuracy                           0.99     37792
   macro avg       0.99      0.99      0.99     37792
weighted avg       0.99      0.99      0.99     37792
```

Fig. Accuracy, Recall, F-1, Support Values

V. CONCLUSION

Security risks have been observed as internet-connected systems are used more frequently. This paper deals with the Random Forest (RF) algorithm to detect four types of attack like DOS, probe, U2R and R2L. The suggested method effectively handles the detection of online intrusions. Compared to previously used algorithms like SVM, Naive Bayes, and Decision Tree, the suggested approach has performed admirably. The suggested method can significantly increase both the detection rates and the false error rates. The knowledge discovery dataset was employed in this instance.

VI. REFERENCES

1.Jafar Abo Nada; Mohammad Rasmi Al-Mosa, 2018 International Arab Conference on Information Technology (ACIT), A Proposed Wireless Intrusion Detection Prevention and Attack System www.jespublication.com PageNo:99 Vol 12, Issue 7, July/2021 ISSN NO:0377-9254

2. Kinam Park; Youngrok Song; Yun-Gyung Cheong, 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService), Classification of Attack Types for Intrusion Detection Systems Using a Machine Learning Algorithm

3. S. Bernard, L. Heutte and S. Adam "On the Selection of Decision Trees in Random Forests" Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, June 14-19, 2009, 978-1-4244-35531/09/$25.00 ©2009 IEEE

4. A. Tesfahun, D. Lalitha Bhaskari, " Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction" 2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, 978-04799-2235-2/13 $26.00 © 2013 IEEE

5. Le, T.-T.-H., Kang, H., & Kim, H. (2019). The Impact of PCA-Scale Improving GRU Performance for Intrusion Detection. 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 International Conference on Trends in Electronics and Informatics (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 LearningBased Intrusion Detection for IoT Networks, 2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 256-265, Japan.

DOI: 10.17148/IJARCCE.2022.116100

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. G. Geetha Vaishnavi, Somaraju Ram Prasad, Pothina Ambika, K. Mounika; 2022 International ResearchJournal of Engineering and Technology “Intrusion Detection System Using PCA with Random Forest Approach.”
12. Nabila Farnaaz, M. A. Jabbar; Twelfth International Multi-Conference on Information Processing 2016, “Random Forest Modeling for Network Intrusion Detection System.”