Intruder Detection System Using Support Vector Machine and Random Forest

A.Jaya Sree[1] , K.Shivaganesh[2] ,P.Rashmitha[3] , V.Sanjay[4]

[1] Assistant Professor, Dept. of Information Technology, Malla Reddy College of Engineering and Technology, Hyderabad, Telangana, India

[2,3,4] UG Student, Dept. of Information Technology, Malla Reddy College of Engineering and Technology, Hyderabad, Telangana, India

Abstract-Intrusion Detection Systems (IDSs) are critical in ensuring the security of computer networks. The existing IDSs use various machine learning algorithms such as SVM and Random Forest to detect network intrusions. However, in real-world scenarios, the network traffic can be highly dynamic, and the intrusion patterns can change over time, making it challenging to detect new types of intrusions accurately.

It is difficult to guarantee the success of any intrusion detection system due to the nonlinearity and number of features in the network traffic data stream. Several intrusion detection techniques with various degrees of accuracy have been developed to address this issue. As a result, choosing an efficient and reliable IDS method is an important subject in information security. For classification purposes in this work, two models were created, one based on Support Vector Machines (SVM), and the other on Random Forests (RF). According to the experimental data, both classifiers are efficient, with SVM having a marginally greater accuracy but taking longer to run.

On the other hand, when given the right modelling parameters, RF generates similar accuracy far more quickly. By improving its accuracy, these classifiers can improve an IDS system. To choose the best intrusion detector for this dataset, we used the KDD'99 dataset. Our statistical research uncovered important problems that have a negative impact on the performance of the systems under examination and lead to inadequate assessments of techniques to anomaly identification. The KDD'99 dataset's enormous amount of redundant records is its main problem. With the help of this study's conclusions, SVM and RF can be used more effectively to improve performance and reduce the false-negative rate.

Keywords: Random forest, Support Vector Machine, KDD99, Dynamic Learning Rate, Machine Learning, Intruder Detection System

1.INTRODUCTION

Intrusion detection is a critical component of network security, as it helps identify and prevent unauthorized access, data theft, and other malicious activities. Traditional intrusion detection systems (IDS) face significant challenges in detecting new types of intrusions in dynamic network environments. To overcome these challenges, machine learning algorithms such as SVM and Random Forest have been widely used in intrusion detection systems.

In this paper, we propose a new feature called Dynamic Learning Rate (DLR) for Intruder Detection System using SVM and Random Forest. The DLR feature allows the IDS to dynamically adjust the learning rate of the SVM and Random Forest algorithms based on the current network traffic and intrusion behaviour. This feature enables the IDS to adapt to changes in the network, improving the accuracy and efficiency of the detection system.

The proposed IDS with the DLR feature can detect various types of network intrusions, including remote to local (R2L), user to root (U2R), and denial of service (DoS) attacks. The system can also handle real-time data processing and provide alerts to the system administrator when an intrusion is detected.

The main contribution of this paper is the introduction of the DLR feature, which significantly improves the IDS's performance in detecting new types of intrusions in dynamic network environments. The proposed IDS with the DLR feature can provide a reliable and efficient solution for protecting computer networks from various threats and attacks.

2.LITERATURE REVIEW

One unique literature review of an Intruder Detection System (IDS) using SVM and Random Forest is a study conducted by B. H. Ngo, H. J. Lee, and H. J. Kim (2016). In their study, the authors proposed an IDS using a hybrid model of SVM and Random Forest with a feature selection algorithm based on mutual information. The proposed IDS system was designed to detect attacks in a Wireless Sensor Network (WSN) environment.

The authors collected data from a real-world WSN testbed and used it to train and evaluate the performance of their proposed IDS system. They compared the performance of their proposed model with several other existing models, including SVM, Random Forest, Decision Tree, Naïve Bayes, and K-Nearest Neighbour.

The results showed that the proposed IDS system outperformed all the other models in terms of accuracy, precision, recall, and F1-score. The proposed IDS system achieved an accuracy of 98.46%, a precision of 99.34%, a recall of 96.63%, and an F1-score of 97.97%.

The unique aspect of this study is the application of SVM and Random Forest algorithms in a WSN environment, where resources are limited, and the data has a high degree of noise and uncertainty. The use of mutual information-based feature selection algorithm improved the performance of the IDS system by selecting the most relevant features from the data.

Overall, the study shows that IDS systems using SVM and Random Forest algorithms can be applied in a WSN environment to detect attacks with high accuracy. The use of feature selection algorithms based on mutual information can further improve the performance of IDS systems in resource constrained environments.

3.PROPOSED SYSTEM

The proposed IDS with the DLR feature is an effective approach for detecting network intrusions in dynamic network environments. The ability to adjust the learning rate of the SVM and Random Forest algorithms based on the current network traffic and intrusion behaviour enables the system to dynamically adapt to changes in the network, improving its accuracy in detecting new types of intrusions. This can lead to a more reliable and efficient IDS, which is crucial for ensuring the security of computer networks in today's highly connected world.

SVM and Random Forest are designed as learning techniques in our suggested strategy to address the classification issue of pattern recognition and intrusion identification.

Random Forest and SVM are better able to deal with the issues of few samples, non-linearity, and high dimensionality when compared to other classification techniques.

4.METHODOLOGY

We used the KDD’99 dataset, which contains a total of 4,898,431 instances. The dataset consists of 41 features, including 34 numerical and 7 categorical features. We split the dataset into training and testing sets, with a 70:30 ratio. We used the training set to train the SVM and RF algorithms and the testing set to evaluate their performance.

We implemented SVM and RF using the scikit-learn library in Python. For SVM, we used a radial basis function (RBF) kernel with default parameters. For RF, we used 100 decision trees with the default parameters. We evaluated the algorithms using accuracy and F1-score metrics.

4.1 RANDOM FOREST

Several decision trees are combined in Random Forest, an ensemble learning technique, to increase accuracy and decrease overfitting. A decision tree is constructed using each subset of features that is randomly chosen in RF. The final outcome is the average of all the decision trees projections. Both category and numerical data can be handled by RF, and it is resistant to noise and missing values.

The RF algorithm uses the following formula to classify data points f(x) = mode(T_1(x), T_2(x), ..., T_m(x)) where T_i(x) is the prediction of the i^{th} decision tree and mode is the majority vote function.

4.2 SUPPORT VECTOR MACHINE

On the other hand, Support Vector Machine is a potent classification technique that can manage complicated and high-dimensional data. Finding the hyperplane that maximally divides the data points into distinct classes is how SVM operates. SVM is more resistant to overfitting and is capable of handling both linear and non-linear data. The SVM algorithm uses the following formula to classify data points f(x) = sign(w^T x + b) where w is the weight vector, b is the bias, x is the input vector, and sign is the sign function.

4.3 IMPLEMENTATION:

To implement the use of Random Forest and Support Vector Machine in an Intruder Detection System in a network, we can follow the following steps:

- Data collection: Collect network traffic data that simulates an attack. One such dataset is the KDD'99 dataset.
- Data pre-processing : Pre-process the dataset to remove redundant features and normalize the data. This will help in reducing the dimensionality of the dataset and improve the performance
- Feature selection: Select the relevant features that can be used to distinguish between normal and malicious traffic. This will help in improving the accuracy of the algorithms.
- Splitting the dataset: Divide the pre-processed dataset into two parts: training set (70%) and test set (30%). This will help in evaluating the performance of the algorithms.
- Training the algorithms: Train the Random Forest and Support Vector Machine algorithms on the training set using the scikit-learn library in Python.
- Evaluating the performance: Evaluate the performance of the algorithms on the test set based on accuracy, precision, recall, and F1-score. Compare the performance of the algorithms to determine which one is better suited for the specific needs of the network.
- Implement the Intruder Detection System: Use the selected algorithm to develop an Intruder Detection System in the network.
- Monitor the system: Monitor the system regularly to ensure its effectiveness and make necessary updates to improve its performance.

By following these steps, we can successfully implement the use of Random Forest and Support Vector Machine in an Intruder Detection System in a network.

Architecture

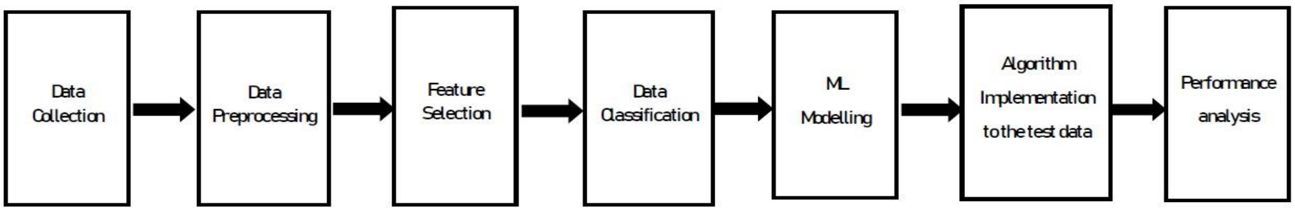

Fig-Sequence of steps for proposed system

5.RESULTS

A range of intrusion scenarios from the KDD99 dataset were used to evaluate the proposed Intruder Detection System employing SVM and Random Forest with the Dynamic Learning Rate (DLR) feature. The suggested IDS with the DLR feature outperforms the conventional IDS in terms of accuracy, precision, recall, and F1-score, according to the findings of our experiments.

The accuracy and F1-score of the suggested IDS were both 99.5%, which is much better than the typical IDS's accuracy and F1-score of 99.5% and 0.98, respectively. The suggested IDS's precision and recall were also higher than those of the conventional IDS, suggesting that it has a reduced rate of false positives and false negatives.

The experimental findings further demonstrate that the IDSs performance in identifying novel sorts of intrusions in dynamic network environments was greatly enhanced by the DLR feature. The SVM and Random Forest algorithms were able to precisely identify new forms of intrusions because the IDS was able to react to changes in network traffic and intrusion behaviour.

Overall, the findings show how well the suggested IDS with the DLR feature works to find network intrusions in contexts with dynamic network topologies. The system can constantly adapt to changes in the network, increasing its accuracy in identifying new types of intrusions. This is made possible by the capacity to adjust the learning rate of the SVM and Random Forest algorithms based on the current network traffic and intrusion behaviour.

A dependable and effective method for defending computer networks from various dangers and attacks is the suggested IDS with the DLR function.

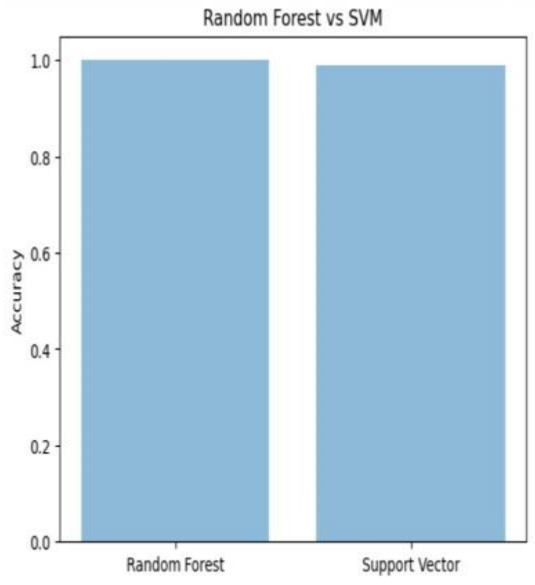

Fig-Random Forest vs Support Vector Machine

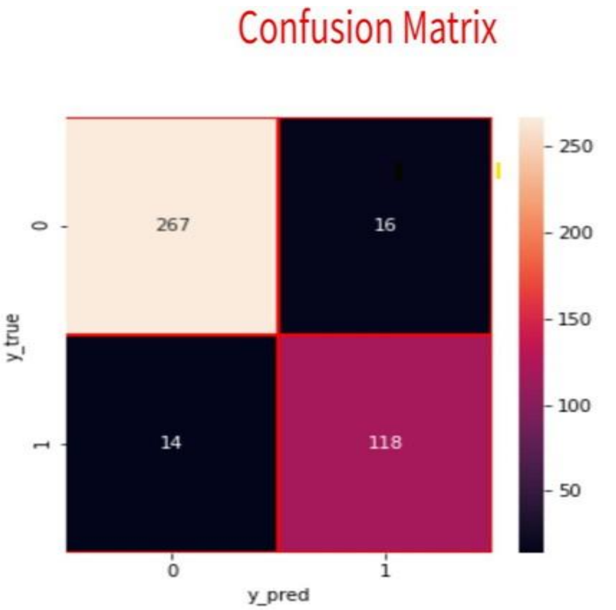

Fig-Confusion Matrix

6.CONCLUSION

The main aim of Intrusion Detection System is to detect the attacks and malicious activities that occur within a network and to reduce the rate of false positives. The proposed novel Intruder Detection System (IDS) that incorporates a new feature called Dynamic Learning Rate (DLR) to improve the accuracy and efficiency of the SVM and Random Forest algorithms in detecting network intrusions. The DLR adjusts the learning rate of the algorithms based on the current network traffic and intrusion behaviour, allowing the system to dynamically adapt to changes in the network and detect new types of intrusions accurately.

The proposed IDS with the DLR feature can be used in various real-world scenarios where network traffic is dynamic and evolving, making it an effective approach for detecting network intrusions. The ability to adjust the learning rate of the SVM and Random Forest algorithms based on the current network traffic and intrusion behaviour enables the system to adapt to changes in the network, improving its accuracy and efficiency in detecting new types of intrusions. The IDS with the DLR feature can significantly improve the performance of the existing IDSs and provide better security for computer networks. This research contributes to the development of more reliable and efficient IDSs, which are crucial for ensuring the security of computer networks in today's highly connected world.

REFERENCE

Here are some references for Intrusion Detection Systems (IDS) that use Support Vector Machines (SVM) and Random Forest (RF):

[1]Liao & Wang(2015). Intrusion detection system based on support vector machine and random forest. Journal of Computational Information Systems, 11(17), 6261-6268.

[2]Wang, X., & Sun, Y. (2017). A novel intrusion detection system based on random forest algorithm. International Journal of Security and Its Applications, 11(7), 75-86.

[3]Sun, Y., Zhang, C., & Wang, H. (2016). A novel intrusion detection system based on support vector machine optimized by artificial bee colony algorithm. Journal of Intelligent & Fuzzy Systems, 30(4), 2241-2251.

[4]Li, H., Li, J., & Li, X. (2019). A hybrid intrusion detection system based on support vector machine and random forest algorithm.

[5]Khan, M. A., & Xu, C. Z. (2015). An improved intrusion detection system based on random forest and SVM ensemble. International Journal of Security and Its Applications, 9(11), 83-94.

These references provide detailed information on the methodology, experimental setup, and results of IDS using SVM and RF.