JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Intrusion Detection System using Machine Learning Approach

Ruchi Singh
M. Tech Scholar
Department of Computer Science and Engineering
VITS Satna M.P. India
ruchisinghc2@gmail.com

Mr. Chhatrapani Gautam
Assistant Professor
Department of Computer Science and Engineering
VITS Satna M.P. India
acpg.77@gmail.com

Abstract: Random harmful actions for a single machine or for the whole network may be seen on the internet from time to time. As computer connection continues to expand at an unprecedented rate, it is becoming more difficult to keep up. Security risks may be seen on the internet, just as they can be seen in person. The intrusion detection system (IDS) is designed to identify and investigate such hostile actions occurring across a network. The intrusion detection system (IDS) aids in the detection of assaults on the system and the identification of attackers. Various machine learning (ML) methods have been used to intrusion detection systems in the past, with the goal of improving the results for intruder detection and increasing the accuracy of the IDS. In this article, we present a method for developing an efficient IDS that makes use of the principle component analysis (PCA) and the CNN classification algorithm. PCA may be used to organise data by decreasing its dimensionality, whereas random forest can be used to classify data. The tests will be carried out using the suggested system over the KDD (Knowledge Discovery Dataset). When compared to other methods such as SVM, Naive Bayes, and Decision Tree, it is certain that the suggested methodology would perform more efficiently in terms of accuracy. We got the following findings using our suggested method: performance time (min) is 3.24 minutes, accuracy rate (percentage) is 96.78 percent, and error rate (percentage) is 0.21 percent.

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

I. INTRODUCTION

With the fast advancement of technology, the internet's presence in everyday life is becoming more prevalent. Almost everyone's life is now dependent on the internet to some degree or another. Using the internet has become more important for everyone these days. It is thus becoming more important to protect the system against harmful activity as the number of people using the internet for personal activities continues to climb.

On the system or the network, several types of assaults are seen. It is the goal of these attacks to steal information from a system or alter the data that is present on any system[1]. A variety of assaults are used by attackers to gain access to and abuse data from the system; they include denial of service (DoS), probe, sniff, r2l, and other similar techniques. Consequently, an intrusion detection system was implemented to protect the system from such assaults. System intrusion detection systems (IDS) maintain track of system assaults and work to keep the system safe from these threats.

1.1 Intrusion Detection System :

Introduction: Intrusion is a word that refers to the act of accessing a system without authorization and causing damage to the information contained inside the system[1]. This infiltration into any system has the potential to cause damage to the system's hardware. The word "intrusion" has evolved into a highly significant phrase in terms of protecting the system from being compromised. The intrusion detection system (IDS) may be used to regulate or keep track of any intrusions that occur inside a system, depending on the situation. Although the different kinds of intrusion detection systems have been utilised in the past, the accuracy of each technique has been questioned in recent years. The two terms, such as the detection rate and the false alarm rate, are examined in order to determine the accuracy of the system[2]. The system should be designed in such a way that the false alarm rate is kept to a bare minimum while the detection rate is increased. As a result, the IDS employs the random forest in conjunction with the PCA.

In nature, the IDS may be of two kinds, for which it is effective, and they are as follows:

In this system, the network traffic is analysed, and any intrusions that occur as a result of the traffic are identified and investigated.

System files that are accessed via the network are tracked by host-based intrusion detection systems (HIDS), which are used to detect network intrusions.

In addition, there is a subset of IDS types. The most often seen variations are those that rely on signature detection and anomaly detection.
Signature-based: In this case, the system discovered certain particular patterns that malware uses to hide its identity. Signatures are the patterns that have been discovered. This is effective in identifying existing assaults; but, when it comes to detecting new attacks, it falls short in the signature detection process.
Anomaly-based: This is a kind of detection that has been specifically designed for the detection of unknown assaults. The model is constructed using this system, which makes use of ML.

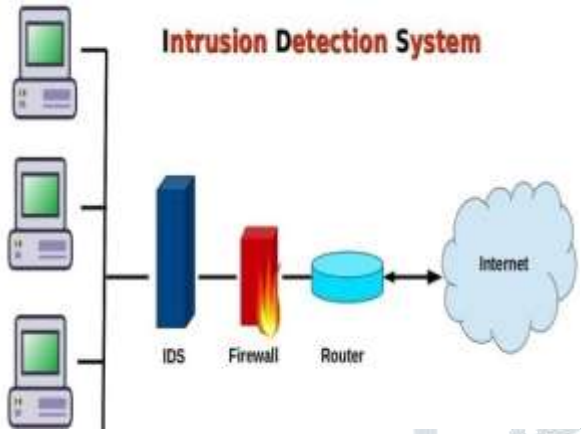

Figure 1. Intrusion Detection System[2]

1.2 Random Forest:

RF is one of the most powerful methods that is used in machine learning for classification problems. The random forest comes in the category of the supervised classification algorithm[2]. This algorithm is carried out in two different stages the first one deals with the creation of the forest of the given dataset, and the other one deals with the prediction from the classifier that obtained in the very first step.

Pseudocode for the creation of a random forest is as follows:

1. Select some features k from total m as k<<m
2. By applying split point from k features get node d
3. By applying **best split** get the daughter nodes
4. Repeat 3 steps till we reach 1 node
5. Create forest by repeating the steps from 1 to 4 for the creation of forest.

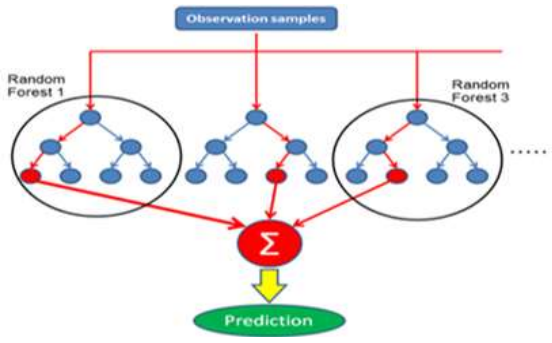

Figure 2. Random Forest Model.

1.3 PCA:

The principal component analysis method is one that is used, particularly for the purpose of reducing the size of a given dataset's dimensions. It is one of the most efficient and precise techniques available for decreasing the dimensionality of data, and it produces the required results[3]. The characteristics of a given dataset are reduced to a desired number of attributes, which are referred to as main components, using this technique.

This technique treats the whole input as a dataset, which has a large number of characteristics and therefore has a large dimension due to the large number of attributes. By aligning all of the data points on the same axis, this technique helps to decrease the size of the dataset. After the data points have been relocated to one of the axes, the main components analysis is carried out.

II. LITERATURE REVIEW

According to the authors, a solution for the IDS was found via application of both SVM and Nave Bayes algorithms, with SVM proving to be superior to both of the other methods. Their experiment was conducted out using data from the KDD dataset, and they also provide findings in terms of detection and false alarm rate. [4]
The authors of this article conducted three separate experiments, which they describe in detail. They used feature selection in the analysis as well as in the design. In addition, the naive Bayes, adaptive boost, and partial decision tree were shown. They looked at all of the intrusion detection methods available. [5]
With the help of this article, the authors have determined that when compared to the Support vector machine method, artificial neural networks with feature selection would provide superior outcomes. The experiment was carried out using the NSL-KDD dataset. The method that was provided was successful. [6]

This paper presents an overview of intrusion detection systems that make use of a machine-learning method, as described by the authors. The authors presented a performance comparison of different machine learning algorithms based on the results they obtained. In order to assess the survey, they looked at its detection rates and false alarm rates. [7]

The authors have developed a method for intrusion detection that makes use of logistic regression and belief propagation techniques. As previously stated, the suggested approach has shown that it offers a faster average detection time when compared to previous methods. [8]

The feature extraction from the dataset was accomplished via the use of an in-depth learning method developed by the authors. They attempted to extract characteristics from a dataset in order to make a dataset more efficient for usage, and in doing so, they came to the conclusion that they could offer better input to the intrusion detection system. [9]

They have conducted a study of intrusion detection systems using a machine learning method in this section. In their study, they looked at all of the machine learning algorithms that have been utilised up to this point and came to the

conclusion that the algorithms provided by Md Nasimuzzaman Chowdhury and ANN submitted by Alex Shenfield, Aladdin Ayesh, and David Day were the most effective. [10]

In this paper, the authors investigated a number of machine learning techniques for use in an intrusion detection system. They compared a number of methods, including SVM, Extreme learning machine, and the random forest, among others. In their conclusions, the authors claim that the Extreme machine learning technique outperforms all other methods by a wide margin. [11]

The authors of this paper attempted to enhance the quality of the dataset in order to make it available to the intrusion detection system for analysis. They have utilised a fuzzy rule-based feature selection method to enhance the dataset, which they have described before. They utilised the KDD dataset, and the results of the IDS showed a dynamic increase in the number of results.[12]

III. PROBLEM DOMAIN

The systems that operate via the internet are subjected to a wide range of harmful actions. The most serious issue that has been seen in this area is the infiltration into the system for the purpose of breaching the information. This intrusion is detected via the development of an intrusion detection system; however, this system must be precise and efficient in its detection of intruders in order to be effective. Machine learning methods were employed for intrusion detection, including SVM, Naive Bayes and other variants of the technique. However, the findings indicate that there may be some room for improvement in terms of accuracy, detection rates, and the incidence of false alarms, among other things. Some additional methods, like as SVM and Nave Bayes, may be used to replace techniques that have been previously used. Additionally, the research claims that by applying certain techniques to the dataset, it may be made better. In order to enhance the quality of the input to the proposed system, it is necessary to.

IV. PROPOSED SOLUTION

The intrusion detection system strives to enhance the overall performance of the system, which is adversely impacted by intruders. This device has the capability of detecting intruders on the premises. The suggested system makes an attempt to resolve the issues that have arisen as a result of the prior work. It is suggested that the system be comprised of two techniques, one of which is principal component analysis and the other of which is random forest. The principle component analysis technique is used to reduce the size of a dataset's dimension; the quality of the dataset will be enhanced as a result of this approach since the dataset will include the right characteristics as a result of this method. For intruder detection, the random forest algorithm will be used, which has a higher detection rate and a lower false alarm rate than the SVM method.

4.1 Algorithm for the proposed solution:

The attribute compatibility replaces the coordination degree of the original attribute for the split node standard.

1. Attribute compatibility
Let the modulus be | Pr | for the main decision set, secondary set be | Se |, and attribute compatibility is defined as:

$$CO(X \rightarrow D) = \frac{|P_r| - |S_e|}{|X|} \quad (1)$$

here X is the subset for non-empty C. Strict compatibility is called when the influence of the secondary set over the main set is seen. A contradiction is seen between the main and the second set. The secondary set is rounded off by the expression.

$$CO(X \rightarrow D) = \frac{|P_r|}{|X|} \quad (2)$$

here X is the subset for non-empty C. In this, the wide compatibility of the second set is seen.

Algorithm for The Base Classifier Improvement:

Step 1: initialisation of data set active attribute by marking all condition attribute.
Step 2: calculate the modulus for every condition attribute in both primary and secondary set.
Step 3: By using equation (1) compatibility calculation of all conditional attribute is done in this step. Use equation (2) if more characteristic with similar compatibility is seen.
Step 4: To separate the sample, select the most extensive compatibility for splitting as the split node and delete the active tag.
Step 5: go on selecting the active attribute for splitting till we get the active attribute ar we reach up to leaf node.
Step 6: At last, we generate the base classifier.

4.2 Flowchart for The Proposed Algorithm:

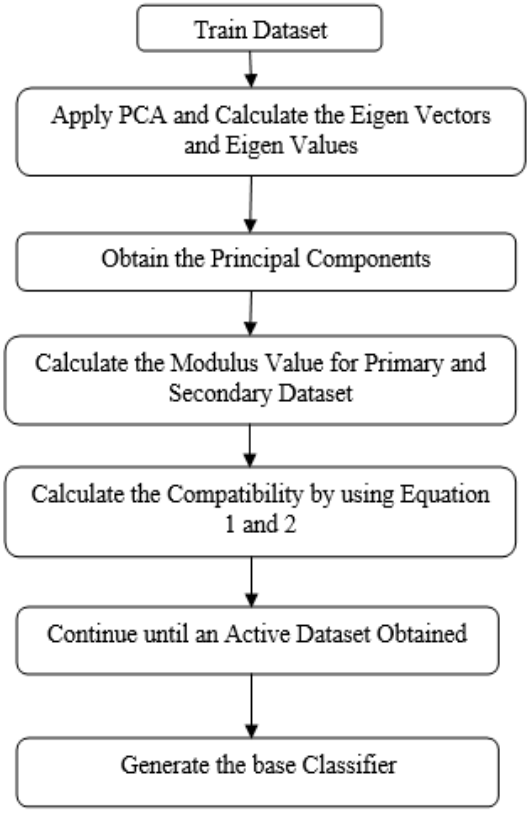

Figure 3. Flowchart for The Proposed Approach

V. RESULTS

With the KDD dataset as the basis for the experiment carried out to test the suggested method, it was possible to achieve satisfactory results. In contrast to current methods such as SVM, Naive Bayes, Decision tree, and CNN, the use of PCA in conjunction with CNN performed very well. The performance time (in minutes), accuracy rate (percentage), and error rate (percentage) for various methods are given in tabular form in the next section:

Table 2. Result Comparison with other Classifiers

Method	Performance time (min)	Accuracy rate (%)	Error rate (%)
SVM	4.57	84.34	2.67
Naïve Bayes	9.12	80.85	3.49
Decision Tree	12.36	89.91	0.78
PCA with CNN	3.42	96.78	0.21

So we can conclude here that the presented approach works well in comparison with the previous algorithms like SVM, Naïve Bayes and Decision Tree. PCA with Random Forest better on the bases of three parameters. Its represent on table 1.

VI. CONCLUSION

As the participation of systems via the internet has grown in recent years, so have the security issues that have accompanied it. The suggested method effectively deals with the detection of intruders via the internet and is thus cost-effective. When compared to previously used algorithms such as SVM, Naive Bayes, and Decision Tree, the suggested method outperformed them all. The suggested method has the potential to significantly increase both the detection rates and the false error rates in a number of ways. The dataset that has been utilised in this example is the knowledge discovery dataset. We obtained the following results using our proposed method: performance time (min) is 3.24 minutes, accuracy rate (percentage) is 96.78 percent, and error rate (percentage) is 0.21 percent. Performance time (min) is 3.24 minutes, accuracy rate (percentage) is 96.78 percent, and error rate (percentage) is 0.21 percent.

Reference:

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
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. 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
4. 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."
5. 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."
6. 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."
7. 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"
8. Nimmy Krishnan, A. Salim, 2018 International CET Conference on Control, Communication, and Computing (IC4) " Machine Learning-Based Intrusion Detection for Virtualized Infrastructures"
9. Mohammed Ishaque, Ladislav Hudec, 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS) "Feature extraction using Deep Learning for Intrusion Detection System."
10. 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."
11. 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."
12. B. Riyaz, S. Ganapathy, 2018 International Conference on Recent Trends in Advanced Computing (ICRTAC)" An Intelligent Fuzzy Rule-based Feature Selection for Effective Intrusion Detection."