International Journal for Innovative
Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IJIEMR Transactions, online available on 26th Apr 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue= Spl Issue 04

DOI: 10.48047/IJIEMR/V11/SPL ISSUE 04/05

Title IDENTIFICATION OF TYPES OF INTRUSION ATTACKS USING SPARK AND GRADIENT BOOSTED TREE CLASSIFIER

Volume 11, SPL ISSUE 04, Pages: 49-55

Paper Authors
K.Sudhakar, S.Chandana Sree, L.Meghana, S.Bhavana, B.Naga Durga

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IDENTIFICATION OF TYPES OF INTRUSION ATTACKS USING SPARK AND GRADIENT BOOSTED TREE CLASSIFIER

[1]K.Sudhakar, [2]S.Chandana Sree, [3]L.Meghana, [4]S.Bhavana, [5]B.Naga Durga
[1]Associate Professor in CSE, PSCMRCET, Vijayawada.
[2,3,4,5] Student, CSE, PSCMRCET, Vijayawada.
[1]ksudhakar@pscmr.ac.in, [2]chandanasreesunkara@gmail.com,
[3]lolugu.meghana@gmail.com, [4]somarajubhavana@gmail.com,
[5]nagadurgabanka2001@gmail.com

Abstract
Big data has become a phenomenon in recent years, coinciding with the technological revolution. Big data is having a greater impact on nearly every field, from mathematics to medicine. To properly explore this data, we need a tool that can handle large amounts of data. Apache Spark is a tool that provides a real-time model for big data analysis. Spark includes memory operations. In addition to this data, data security is an issue. We must take appropriate measures to protect data from intrusion attacks. Knowing the types of attacks is critical in this situation. The main goal of this work is to use the libraries provided by Spark to pre-process the large amounts of data and identify the types of intrusion attacks. We classified the attacks using four different classifiers into two types of labels: benign attacks and infiltration attacks. As a result, we can determine the type of intrusion attack.

Keywords: Intrusion attacks, Apache Spark, Gradient Boosted Tree Classifier, Big Data, Linear SVM, Logistic Regression, Decision Tree Classifier, Benign attack, Infiltration attack.

I. INTRODUCTION

Intrusion attacks occur when unauthorized individuals attempt to gain access to a secure system and its data [3]. These can cause the system to crash or slow down. Private data can be viewed by intruders. To protect confidential data from others, we must take precautions, and in order to know what type of protection to provide, we must first understand the type of attack. Malicious and non-malicious intrusion attacks can be distinguished. Malicious attacks are those that disrupt system functionality, whereas non-malicious attacks are those that do not disrupt system operations. In our research, we classified intrusion attacks as benign or infiltration, with benign attacks being non-malicious and infiltration attacks being malicious.

In general, when it comes to network and security data, we are presented with an abundance of data. So we used Apache Spark to pre-process and select the appropriate features for attacks. Spark provides various utilities for working with big data[6], as well as several libraries for working with machine learning algorithms. In our investigation, we first used the selectkbest library to classify the type of intrusion attacks, followed by data processing[7]. Then we use various classifiers, such as logistic regression, decision tree classifier, gradient boosted tree classifier, and linear SVM. And then we compared the results to see which one best suited our needs.

II. RELATED WORK

[1] Using the PySpark Python library, Abhijeet Urunkar et al. created a machine learning model to detect insurance fraud. They built the model using KNN, Logistic Regression, SVM, and Decision Trees and compared their performance metrics. They compared these algorithms in terms of confusion matrix, accuracy, precision, recall, and so on. The PySpark and Scikit-Learn libraries were used by the authors. To convert categorical values to numerical values, they used a single hot encoding. During the training phase, they extracted the best features and removed unnecessary columns such as insured educational level, insured occupation, and so on. Finally, they discovered that their machine learning models identified fraud cases with a low false positive rate, resulting in low exactness. The authors concluded that different characteristics of different datasets are not suitable for using feature engineering to improve performance.

[2] Mohamed Haggag et al. proposed an Apache Spark Platform technique for developing an intrusion detection deep learning model. In today's world, the expansion of the internet has resulted in massive amounts of data. The NSL-KDD dataset has a class imbalance problem. The NSL-KDD dataset is used to train and test the DLS-IDS model, which is made up of four main building blocks. They are select and investigate, pre-process, and class imbalance solution. Two factors that slow down machine learning model training are the size of the dataset and the optimization parameters used to design the best-fitting model. We use an Apache Spark tool to solve these problems. It is one of the fastest Clustering frameworks, as well as an open-source cluster distributed programming tool. It might be useful in resolving the issues. Apache Spark also performs memory operations. If there is an imbalance in class, we should use the Synthetic Minority Over-Sampling Technique (SMOTE) whenever we work with a large dataset. The procedure SMOTE is a pre-processing step that improves model detection accuracy while decreasing false positives. Deep Learning overfitting has a positive effect on the DLS-IDS because LSTM and SMOTE work well together and increase detection precision to 83.57 percent.

[3] VenkatramaphanikumarSstla et al. investigated the effectiveness of two supervised machine learning algorithms, SVM and Deep Neural Networks, on Network Intrusion Detection Systems. They used SVM and DCNN to create a Network Intrusion Detection System and tested its performance with various kernels and activation functions. They used DNN classifiers to identify the unknown attacks because DNN adapts to new patterns as well as previously defined ones. Convolutional layers were used to extract features, and

pooling layers were used to improve feature generalisation. They extracted 13 features from a total of 38 features after reducing them. They divided the data into batches in order to complete computational challenges. The performance of the proposed method is evaluated using the NSL-KDD dataset. In terms of accuracy, DCNN outperformed SVM in the testing.

[4] Ali Mostafaeipour et al. created a model that compares the performance of Hadoop and Spark frameworks for distributed data processing using the KNN algorithm. They ran several experiments on Hadoop and Spart platforms to evaluate parameters such as runtime, CPU utilisation, memory utilisation, and network utilisation. In both platforms, KNN with k=5 was developed for each dataset. The authors recorded the performance parameters using Ganglia monitoring software. The results were analysed in separate sections. According to the comparison study, Spark outperforms Hadoop in terms of runtime. It is 40% more advantageous than Hadoop, but Hadoop outperforms it in memory usage. Spark is also dominant in terms of CPU and network utilisation. They came to the conclusion that the growth and expansion of data volume has an impact on the performance of both platforms. When the dataset is small, Spark outperforms Hadoop by 4.5 to 5 times. Spark can store data in memory, making it ideal for iterative algorithms.

[5] An intrusion detection system based on Apache Spark and machine learning was proposed by Otmane Azeroual and Anastasija Nikiforova. The volume of data in both industries and research organisations has grown dramatically in recent years. Traditional software solutions are used to process this massive amount of data. This paper discusses a security flaw in the Apache Spark big data analytics engine. An intrusion detection system's goal is to detect intrusions using the K-Means approach for clustering analysis implemented in sparks MLlib to discover data anomalies using machine learning. The authors proposed a system that uses big data analysis methodologies and machine learning algorithms to improve IT system security by utilising the K-Means approach for clustering analysis implemented in sparks MLlib to detect data anomalies using machine learning.

III. PROPOSED SYSTEM

This paper aims to handle large amounts of data using machine learning algorithms and PySpark libraries. PySpark is a tool designed to work with Python and Spark simultaneously. We classify intrusion attacks using the network dataset. In the spark environment, the data is pre-processed using SQL queries. To analyse statistics, we worked with spark data frames and then converted them to pandas data frames. To extract the best features, we used selectkbest. To determine the type of attack, we used four different classifiers. Logistic regression, decision trees, gradient boosted tree classifiers, and linear SVM are the classifiers used. The gradient boosted tree classifier outperformed the remaining classifiers.

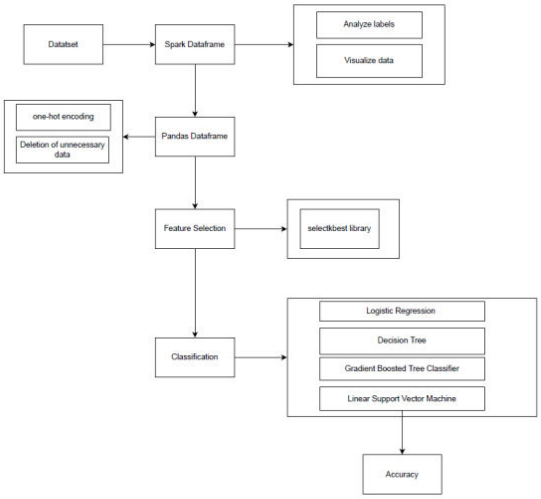
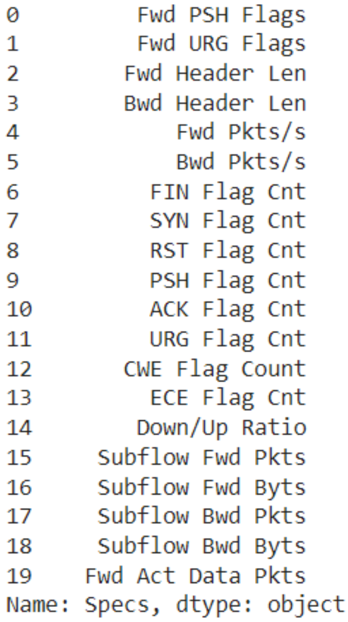


Fig 1: System Architecture

Figure 1 depicts the proposed system's system architecture. The network dataset is imported as a spark data frame, then SQL is used to analyse and visualise it. Following the calculation of descriptive statistics, it is converted into a pandas data frame and one hot encoding is performed. The processed data is then used by selectkbest to select the top ten features. Finally, four classifiers are used to classify the attacks, and their accuracy is measured.

A. FEATURE SELECTION

The process of selecting our data features that contribute the most to the target variable/label is known as feature selection. Classes from sklearn.feature selection are used here. Selectkbest is a task that involves extracting the best features from available data. It chooses the features based on the highest k score. The application of these extracted relevant features to the label variable improves the precision of our ML model.



```
0         Fwd PSH Flags
1         Fwd URG Flags
2        Fwd Header Len
3        Bwd Header Len
4            Fwd Pkts/s
5            Bwd Pkts/s
6          FIN Flag Cnt
7          SYN Flag Cnt
8          RST Flag Cnt
9          PSH Flag Cnt
10         ACK Flag Cnt
11         URG Flag Cnt
12       CWE Flag Count
13         ECE Flag Cnt
14        Down/Up Ratio
15     Subflow Fwd Pkts
16     Subflow Fwd Byts
17     Subflow Bwd Pkts
18     Subflow Bwd Byts
19    Fwd Act Data Pkts
Name: Specs, dtype: object
```

Fig 2: Extracted Features

B. CLASSIFICATION

Apache Spark is highly fast while pre-processing data and it is very easy to use. PySpark is an APL provided in python around Apache Spark. It is used for data mining and data pre-processing. Here we used PySpark to classify the attacks into benign and infiltration attacks. Classification refers to a task of separating given data into classes. These classes are referred to as labels or categories.

Our system involves:

- Logistic Regression
- Decision Tree
- Gradient Boosted Tree Classifier
- Linear SVM

1.LOGISTIC REGRESSION

Logistic regression is used to predict a categorical response. It is a popular method in case of generalized linear models. In spark.ml we can used logistic regression model to predict a binary outcome or a multiclass outcome. It can also be used to describe data. This is an classification that says the dependency of data in PySpark ML model.

2. DECISION TREE

Decision trees are widely used methods for classification tasks as they are easy to handle and interpret. They can easily be extended to multi-classifiers and they do not require feature scaling. MLlib provides decision tress for binary and multiclass classification. It is powerful supervised machine learning to extract rules from training data.

3. GRADIENT BOOSTED TREE CLASSIFIER

Gradient Boosted Tree Classifiers (GBTs) generally train one tree at a time. Each new tree improves upon drawbacks of preceding tree. They take longer than random forest to train tress. We can call summary to get summary of the fitted GBT model. GBT currently supports only classification tasks.

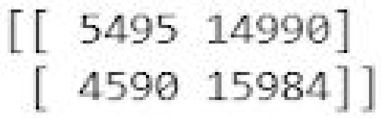
4. LINEAR SUPPORT VECTOR MACHINE

Linear SVM is basic method for large-scale classification tasks. It outputs a SVM model. Here all the data must be numeric. It provides feature scaling to reduce the condition numbers before training. The scaling correction is transparent to the users as it will be translated back to model weights.

IV. EXPERIMENTAL RESULTS

A. Confusion Matrix

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.



```
[[ 5495 14990]
 [ 4590 15984]]
```

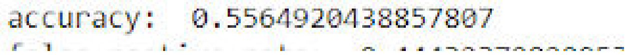
Fig 3: Confusion Matrix for GBT Classifier

B. Accuracy

Machine learning model accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training data.

Formula:
(Total Number of correct predictions) Accuracy
=---
(Total number of samples)

<u>Accuracy for GBT Classifier:</u>

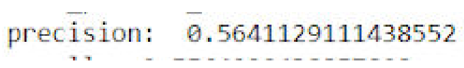


```
accuracy:  0.5564920438857807
```

C. Precision

Precision is one indicator of a machine learning model's performance the quality of a positive prediction made by the model. Precision refers to the number of true positives divided by the total number of positive predictions (i.e., the number of true positives plus the number of false positives).

Precision for GBT Classifier:



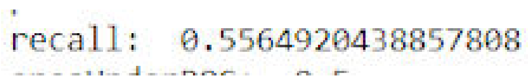
```
precision:  0.5641129111438552
```

D. Recall

Recall literally is how many of the true positives were recalled (found), i.e. how many of the correct hits were also found.

Formula:
True Positives
Recall = ---
(True Positives + False Negatives)

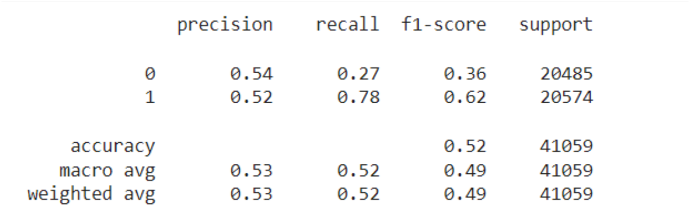
Recall for GBT Classifier:



```
recall:  0.5564920438857808
```

Table 2: Comparison of Performance Metrics

Algorithm Name	Accuracy	Recall	Precision
Logistic regression	50.0	50.0	25.0
Decision Tree	53.6	56.5	53.6
Gradient Boosted Tree Classifier	55.6	56.4	55.6
Linear Support Vector Machine	52.5	52.2	52.5



```
              precision    recall  f1-score   support

           0       0.54      0.27      0.36     20485
           1       0.52      0.78      0.62     20574

    accuracy                           0.52     41059
   macro avg       0.53      0.52      0.49     41059
weighted avg       0.53      0.52      0.49     41059
```

Fig 4: Analysis of GBT Classifier

V. CONCLUSION

Finally, the dataset is analysed and handled in the spark environment to remove unnecessary attributes before being converted to a pandas data frame. Following the selection of the best ten features, classification algorithms are run to classify benign and infiltration attacks. The gradient boosted tree classifier outperformed the other classifiers tested. We would like to improve this work by having the system notify the user when an intrusion is detected.

REFERENCES

[1] Urunkar, Abhijeet &Khot, Amruta & Bhat, Rashmi &Mudegol, Nandini. (2022). “Fraud Detection and Analysis for Insurance Claim using Machine Learning”. 406-411. 10.1109/SPICES52834.2022.9774071.

[2] Haggag, Mohsen M, Magdy (2020). “Implementing a deep learning model for intrusion detection on apache spark platform”. ECE, Cairo University, Egypt. doi: 10.1109/ACCESS.2020.3019937

[3] Sstla, V., Kolli, V.K.K., Voggu, L.K., Bhavanam, R., Vallabhasoyula, S. (2020). “Predictive model for network intrusion detection system using deep learning”.

Revue d'IntelligenceArtificielle, Vol. 34, No. 3, pp. 323-330. doi.org/10.18280/ria.340310

[4] Mostafaeipour, Ali &Jahangard, Amir & Ahmadi, Mohammad &Arockia Dhanraj, Joshuva. (2021). "Investigating the performance of Hadoop and Spark platforms on machine learning algorithms". The Journal of Supercomputing. 77. 10.1007/s11227-020-03328-5.

[5] Azeroual, Otmane, and Anastasija Nikiforova. 2022. "Apache Spark and MLlib-Based Intrusion Detection System or How the Big Data Technologies Can Secure the Data" *Information* 13, no. 2: 58. doi: 10.3390/info13020058

[6] Gopala Krishna, K Sudhakar, Sandeep S R, Karthik, Navaraja (2021). "Implications of big data analytics for pharmaceutical industry and transforming healthcare". PSCMRCET, Andhra Pradesh.

[7] J S V K Gopala Krishna, K Sudhakar, B Loveswara Rao (2020). "Meta data management and its governance using big data tools". ISSN NO : 1006-6748

[8] Lekha, Sujala, Siddharth (2018). "Applying spark based machine learning model on streaming big data for health status prediction". doi: 10.1016/j.compeleceng.2017.03.009

[9] Jinlong Liu, Christopher, Emil Dumitrescu (2020). "Random forest machine learning model for predicting combustion feedback information of a natural gas spark ignition engine". doi: 10.1115/1.4047761

[10] A. Esmaeilzadeh, M. Heidari, R. Abdolazimi, P. Hajibabaee and M. Malekzadeh, (2022) "Efficient Large Scale NLP Feature Engineering with Apache Spark," doi: 10.1109/CCWC54503.2022.9720765.