An ensemble-based scalable approach for intrusion detection using big data framework

Durga Prasad Mohapatra
National Institute of Technology, India

Statement of the Problem: Big Data analytics plays a vital role in intrusion detection. It provides tools to support structured, unstructured, and semi-structured data for analysis. Also, it offers scalable algorithms for the fast processing of data using machine learning approaches. Most of the AI/ML-based approaches are biased towards the majority class during the learning process. As a result, the minority classes are misclassified. It is a challenging task for researchers of different domains to detect such instances which are commonly known as class imbalanced problems.

Research Objective: To develop a scalable, robust intrusion detection model that can handle imbalanced data and avoid biases during the learning and testing process on the big data framework as shown in [Figure 1].

Tools used: Apache Spark, Hadoop, PySpark, etc.

Methodology & Theoretical Orientation: In this work, scalable machine learning algorithms i.e. RUSBoost and Decision Tree classifier with revised K-Means approaches for threat classification are implemented using the big data framework. The algorithm has been designed to work even with a relatively small training set and support to classify a large volume of testing data as shown in [Figure 2].

Findings: The result of the proposed approach is discussed. The decision tree is used to detect the majority class and RUSBoost for detecting the minority class in the proposed ensemble approach. The popular intrusion datasets are used for training, validation, and model testing during the experiment.

Conclusion: The combination of clustering, boosting, and supervised approach improves the model's robustness to deal with the imbalanced datasets. As per the obtained result, the detection accuracy of the minority classes was drastically enhanced compared to other popular approaches. The classification error is minimized in all the datasets and the confusion matrix and ROC, show the proposed approach's robust detection accuracy.

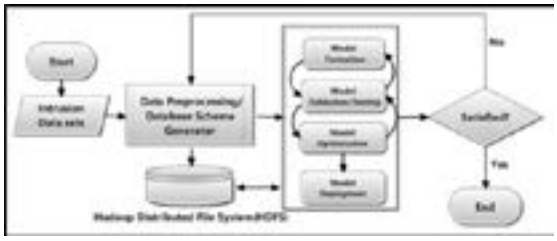

Figure 1. *Flowchart of proposed approach using the big data framework*

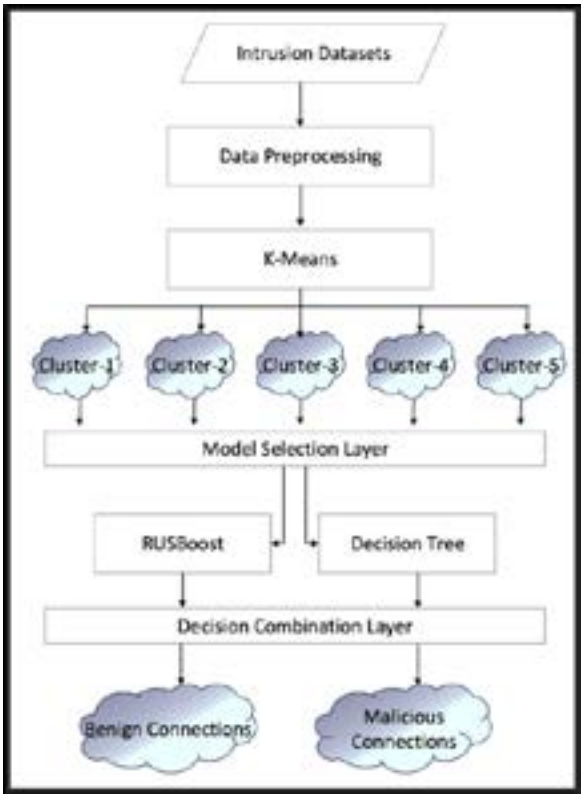

***Figure 2**. Proposed ensemble approach*

Biography

Durga Prasad Mohapatra is a Professor at NIT, Rourkela, India. His research interests include software engineering, discrete mathematics, and distributed computing. He has authored 250+ research articles and co-authored the book Elements of Discrete Mathematics published by Mc-Graw Hill. He has received the Young Scientist Award and Prof. K. Arumugam National Award for outstanding research works in Software Engineering.

Received: June 24, 2022; **Accepted**: June 26, 2022; **Published**: August 03, 2022