Abstract No: 312 ***ICT, Mathematics and Statistics***

EFFECTIVE MACHINE LEARNING ALGORITHMS FOR ANOMALY INTRUSION DETECTION SYSTEM

D.M.C. Dissanayake*

Department of Computing and Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka
**charithdissanayake1@gmail.com*

Cyber security is an important and highly considerable area in the technology due to increment of various kinds of security attacks. Network intrusion detection system is a software tool which monitors malicious activities of the network traffic combining with hardware components of the network. Mainly, there are two types of network intrusion detection approaches which are signature-based detection and anomaly-based detection. The signature-based detection method attempts to detect attack patterns, while the anomaly-based detection method classifies network traffic into "attack" or "normal" using techniques like machine learning. This study is designed to compare five discrete machine learning algorithms for the anomaly-based detection, and further to identify the most effective algorithm. The dataset NSL-KDD which contains 125,973 records for training and 22,544 records for testing was used for the study. Initially, pre-processing was carried out for the training and testing data and during the pre-processing, data were encoded first, in order to convert categorical variables to numerical values. Then features were reduced up to 27, using Principal Component Analysis (PCA) dimensionality reduction technique. Finally, they were normalized in order to change feature values to specific range. After the pre-processing step, 10-Fold cross validation was carried out on the training data. The mean values of 10-Folds cross validation for Decision Tree, Random Forest, K-Nearest Neighbours, Logistic Regression and Support Vector Machine are 99.70%, 99.84%, 99.67%, 96.81% and 99.47%, respectively. Among these results, Random Forest algorithm was observed with the best score. Therefore, hyperparameters were tuned of this algorithm and fit the model. Thereafter, it obtained the best score as 99.84%. Finally, it was validated with the test data and scored 99.99% against test data. Logistic Regression provides the lowest performance and Random Forest provides the highest performance while other three algorithms also provide satisfactory performances.

Keywords: Anomaly based detection, Classification, Cyber security, Network intrusion detection, Random forest