ABSTRACT

Analyzing network traffic logs to identify cyberattacks is crucial for upholding cybersecurity. Conventional approaches to anomaly detection may falter with high-dimensional and noisy datasets, resulting in a compromise between accuracy and recall. This paper introduces an improvement to current ensemble anomaly detection techniques by integrating the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm into the pruning procedure. The suggested approach for anomaly identification integrates Isolation Forest (iForest) with Local Outlier Factor (LOF). Furthermore, DBSCAN enhances the pruning procedure by adeptly managing dense clusters of normal data while disregarding noisy spots. A comprehensive assessment was conducted on the effectiveness of the improved pruning technique utilizing the NSL-KDD and HIKARI2021 datasets, which are recognized standards in cybersecurity research. The experimental results demonstrate that the integration of DBSCAN enhances the recall rate to a perfect 100% by minimizing false negatives and improving the overall effectiveness of the anomaly detection process. The proposed strategy significantly improves the F1-score for the NSL-KDD dataset from 0.867 to 0.915, indicating a better balance between precision (which declined from 0.838 to 0.843) and recall (which grew from 0.890 to 1.000) compared to existing methods. For the HIKARI2021 dataset, the F1-score increased from 0.785 to 0.816, but precision marginally declined from 0.693 to 0.689, and recall improved to a flawless 1.000 from 0.905. Moreover, the improved pruning method utilizing DBSCAN markedly diminishes the dataset size, achieving pruning rates of up to 70.29% and 58.14% for the NSL-KDD and HIKARI2021 datasets, respectively. Each dataset can preserve anomaly proportions of up to 84.26% and 69.34%, respectively. The diminishment in dataset size results in a 15.74% and 30.66% decline in identification outcomes, consequently enhancing computational efficiency. This research enhances cybersecurity by offering a more dependable and effective method for detecting cyberattacks in intricate datasets.

Keywords: Anomaly detection, DBSCAN, isolation forest, local outlier factor, cybersecurity, network intrusion.