

ABSTRACT

Due to the growing amount and variety of network traffic, traditional intrusion detection systems (IDS) have become insufficient due to inadequate labeled data, severe class imbalances, and the rigidity of confidence thresholds in semi-supervised learning. To address these challenges, this paper proposes a two-stage, cascaded co-training framework that utilizes Naive Bayes and Random Forest algorithms with graph-based label propagation. In recognition of the fact that static confidence cut-offs either introduce noisy labels or restrict learning excessively, adaptive thresholds derived from Flexible Confidence (FlexCon) variants are applied to govern the inclusion of pseudo-labels. Additionally, three feature representations—original, IPCA, and random projection—are compared to ensure equal trade-offs between detection accuracy and computational overhead. By utilizing both labeled and unlabeled data, this design balances class imbalance while dynamically adjusting labeling criteria to enhance the performance of multi-class anomaly detection in large-scale IDS.

Keywords: Semi-Supervised, CICIOT2023, FlexCon, Co-Training, Dimensionality Reduction