Title: Synthetic Data Generation with Machine Learning for Network Intrusion Detection System

Authors: Marvin Newlin, Mark DeYoung, Mark Reith

Abstract: Machine learning is becoming an integral part of cybersecurity today, particularly in the area of network anomaly detection. However, machine learning techniques require large volumes of data to be effective. Although some real and semi-real data sets exist, many are outdated or do not contain enough useful information for training/classification of network intrusion detection systems (NIDS), particularly considering the new types and variants of communication protocols in today's environment. Generating sufficiently realistic synthetic network traffic is imperative for training effective NIDS. In this paper, we discuss the available datasets, the features that make up a sufficiently realistic dataset, and how to measure the degree of realism. Furthermore, we outline a basic approach to developing synthetic data by leveraging machine learning and suggest opportunities to tailor the dataset for specific research objectives. We believe this work might have broad applicability to other researchers developing their own NIDS framework.

Key Words: Synthetic Data Generation, Network Intrusion Detection System, Machine Learning