Master thesis proposal

Balancing data for anomaly-based intrusion detection systems

Arkadiusz Warzyński

November 29, 2024

# Background

Network anomaly detection is currently one of the main directions in the development of intrusion detection systems (IDS) and a way to deal with the growing number of network attacks on information systems. The need to process complex data about network activity and the growing volume of data make the use of traditional rule-based methods ineffective to ensure a high level of security. However, it should be noted that anomaly detection is by design based on the detection of outlier observations, which are underrepresented in the processed data. The limited representation of the selected class in the training sets for classification methods can cause degradation in the quality of the classifier, which must be counteracted by using data balancing methods. The problem of detecting threats in network traffic is particularly challenging in this regard due to many limitations on the range of observed values and the ability to observe attacks in real network traffic. This necessitates research to determine the effectiveness of data set balancing methods used to date for the intrusion detection problem.

# Objectives

The project focuses on conducting an analysis of the impact of dataset balancing methods on the ability to detect anomalies in network traffic to identify potential threats. It is necessary to consider constraints arising from the specifics of network traffic: network layer model, flows and devices. The project should analyze modern oversampling and undersampling algorithms and solutions like Synthetic Minority Oversampling Technique (SMOTE).

# Data description

The CICIDS collection contains network traffic generated in a laboratory environment by profiling 25 abstract network users. User behavior was modeled using the B-profile system by simulating the use of HTTP, HTTPS, FTP, SSH and e-mail services. The data collection period covered 5 days. At the same time, it is one of the newest and most widely used datasets for the problem of intrusion detection in network traffic, which shows great imbalance due to the small representation of data on attacks carried out.