Optymalizacja próbki danych do klasyfikacji ruchu sieciowego przy użyciu uczenia maszynowego

Streszczenie

Liczba cyberataków z roku na rok rośnie, tylko w zeszłym roku liczba ta wyniosła ponad 6 miliardów na całym świecie. Najlepszą z metod przeciwdziałania tym zagrożeniom jest skuteczna detekcja i blokada zagrożeń już na poziomie ruchu sieciowego. W tym celu stosuje się algorytmy sztucznej inteligencji. Niniejsza rozprawa bada jak zoptymalizować próbkę ruchu sieciowego i nadal otrzymywać akceptowalnie wysoką wydajność klasyfikacji. Badania przeprowadzane są na dwóch najczęstszych formach zapisu ruchu sieciowego: ruchu surowym oraz ruchu zapisanym jako cechy ruchu sieciowego.

Badania optymalizacji surowego ruchu skupiają się na wpływie wybranych segmentów pliku PCAP na wynik klasyfikacji ruchu sieciowego. Jest to szczególnie ważne podczas używania zarejestrowanego wcześniej ruchu sieciowego do trenowania modeli uczenia maszynowego, jako że te modele będą później zastosowane do klasyfikacji ruchu w innych środowiskach. Stąd, segmenty ruchu, które generują wysokie wyniki na zbiorze treningowym ale nie prowadzą do takich wyników na nowych, nieznanych jeszcze danych, są identyfikowane. Następnie mają miejsce testy dotyczące optymalizacji wielkości próbki ruchu surowego. Testy wykonane zostały na plikach PCAP zawierających ruch sieciowy bezpieczny i niebezpieczny z następujących zbiorów danych: USTC-TFC2016, ISCX VPN-nonVPN i ISCX Tor-nonTor. Otrzymane wyniki potwierdzają, iż można analizować małe próbki ruchu sieciowego i nadal skutecznie klasyfikować ruch niebezpieczny.

W dalszej części badań ma miejsce optymalizacja zbioru cech ruchu sieciowego. Zaproponowana została nowa metoda selekcji cech oparta na zespole klasyfikatorów. Testy skupiają się na zmniejszeniu liczebności zbioru cech, pozostawiając jedynie te cechy, które są najważniejsze dla klasyfikacji ruchu sieciowego. Badania nad klasyfikacją wykorzystują ruch sieciowy, w formie cech, pochodzący z wybranych zbiorów danych: NF-UNSW-NB15-v2, NF-CSE-CIC-IDS2018-v2, NF-BoT-IoT-v2 , NF-ToN-IoT-v2 i NF-UQ-NIDS-v2. Wyniki badań wykazują, które cechy ruchu sieciowego mają kluczowe znaczenie przy detekcji konkretnych ataków. Otrzymane wyniki są obiecujące do codziennego użytku przez analityków cyberbezpieczeństwa.

Celem badań jest również weryfikacja zdolności wspomnianych zbiorów danych do uogólnień. Prace badawcze koncentrują się zarówno na bezpiecznym jak i na niebezpiecznym ruchu. Prowadzą do ważnych z punktu widzenia bezpieczeństwa sieci komputerowej wniosków, wskazując słabą zdolność do uogólnień badanych zbiorów.

Słowa kluczowe: cyberbezpieczeństwo, analiza ruchu sieciowego, wykrywanie włamań sieciowych, uczenie maszynowe