Politechnika Wrocławska

Wydział Informatyki i Telekomunikacji

Rozprawa doktorska

ADAPTACYJNA METODA WYKRYWANIA ANOMALII W RUCHU TELEINFORMATYCZNYM Z UWZGLĘDNIENIEM PRZYPADKÓW NIEPRZYJAZNYCH

Arkadiusz Warzyński

Promotor: dr hab. inż. Grzegorz Kołaczek, prof. uczelni

Wrocław 2024

Streszczenie

Systemy wykrywania zagrożeń wykorzystujące metody detekcji anomalii są obecnie ważnym elementem infrastruktury systemów teleinformatycznych służącym zapewnieniu bezpieczeństwa. Rosnąca liczba zróżnicowanych ataków wymaga poszukiwania coraz skuteczniejszych metod wykrywania zagrożeń. Jednym z obiecujących kierunków rozwoju systemów wykrywania zagrożeń jest zastosowanie metod sztucznej inteligencji i uczenia maszynowego w celu rozpoznawania i klasyfikacji monitorowanego ruchu teleinformatycznym. Dotychczasowe badania udowodniły, że obecnie wykorzystywane algorytmy klasyfikacji są podatne na ataki polegające na dodawaniu niewielkich perturbacji do oryginalnych danych, co ostatecznie prowadzi do błędnej klasyfikacji. Istnienie przypadków nieprzyjaznych (*ang. adversarial examples*) stanowi poważane zagrożenie dla bezpieczeństwa systemów wykrywania zagrożeń wykorzystujących metody detekcji anomalii.

W niniejszej rozprawie dokonano analizy bezpieczeństwa sieci neuronowej przygotowanej w celu wykrywania zagrożeń poprzez detekcję anomalii w ruchu teleinformatycznym. W tym celu dokonano wyboru metody generowania przypadków nieprzyjaznych oraz przygotowano zbiór danych charakteryzujący ruchu teleinformatyczny, który posłużył do przeprowadzenia badań nad skutecznością klasyfikacji przypadków nieprzyjaznych przez systemy wykrywania zagrożeń. Przeprowadzone badania potwierdziły, że dotychczas stosowane metody uczenia pozostawiają model podatnym na ataki polegające na unikaniu (*ang. evasion attacks*). Otrzymane wyniki udowodniły wysoką skuteczność tego rodzaju ataku w celu wymuszenia błędnej klasyfikacji danych.

Celem pracy było opracowanie metody wykrywania zagrożeń w ruchu teleinformatycznym wykorzystującej wykrywanie anomalii zdolnej do identyfikacji przypadków nieprzyjaznych. Realizacja celu poprzedzona została analizą dotychczas stosowanych metod obrony przed wpływem przypadków nieprzyjaznych na jakość klasyfikacji. Dokonana została charakterystyka najpopularniejszych rozwiązań i opisano najważniejsze wymagania dla metod obrony wykorzystujących koncepcję uczenia przypadkami nieprzyjaznymi (*ang. adversarial training*): konieczność poprawnej klasyfikacji obserwacji w zbiorach danych niezawierających przypadków nieprzyjaznych oraz odporność na przeprowadzanie ataku na zabezpieczony uczeniem przypadkami nieprzyjaznymi model.

W dalszej części pracy przedstawiono autorską metodę zabezpieczania modelu klasyfikacji poprzez uczenie przypadkami nieprzyjaznymi, zdolną do adaptacji do zmieniających się wartości parametrów przeprowadzanego ataku oraz modyfikacji samego ataku. Ocena skuteczności opracowanej metody została dokonana dla czterech zaproponowanych scenariuszy testowych i zestawieniu otrzymanych wartości miar jakości klasyfikatora z wynikami uzyskiwanymi przez dotychczas stosowną, referencyjną metodę uczenia przypadkami nieprzyjaznymi dla tych samych danych.

Przeprowadzone badania obejmowały analizę różnych wartości parametrów zarówno dla opracowanej metody, jak i parametrów ataku. Uzyskane wyniki dla optymalnych konfiguracji parametrów uznano za satysfakcjonujące. W każdym z przeprowadzonych testów zaproponowana metoda uczenia przypadkami nieprzyjaznymi okazała się bardziej skuteczna od metody referencyjnej. Zweryfikowano również skuteczność klasyfikacji danych pochodzących z zestawów danych niezawierających przypadków nieprzyjaznych oraz w przypadku wystąpienia ataku na zabezpieczony model.

W ostatnim rozdziale przedstawiono potencjalne kierunki rozwoju i dalszych badań nad zastosowaniem uczenia przypadkami nieprzyjaznymi do ochrony systemów wykrywania anomalii.

18.06.24r. Arkadiusz Warzyński