ANALIZA WYJAŚNIALNOŚCI ALGORYTMÓW UCZENIA MASZYNOWEGO W ZADANIACH DETEKCJI ATAKÓW W PROGRAMOWALNYCH SIECIACH KOMPUTEROWYCH

STRESZCZENIE

W rozprawie doktorskiej przedstawiono aktualne spojrzenie na współczesne wyzwania związane z detekcją ataków DDoS w środowiskach sieci definiowanych programowo, a także skoncentrowano się na ważnym zagadnieniu, jakim jest wyjaśnialność modeli uczenia maszynowego. W niniejszej rozprawie doktorskiej teza brzmi: **"Autorskie propozycje i implementacje kryteriów wyjaśnialnej sztucznej inteligencji umożliwiają skuteczne wyjaśnienie podejmowanych decyzji przez modele uczenia maszynowego w procesie detekcji rozproszonych ataków odmowy usługi w sieciach definiowanych programowo."**

W związku z tą tezą celem dysertacji było wykazanie, iż odpowiednia integracja algorytmów klasyfikacji, technik selekcji cech oraz metod wyjaśnialności może nie tylko skutecznie pomóc wykrywać ataki DDoS, ale również przyczynić się do głębszego zrozumienia, które elementy zaproponowanego systemu bezpieczeństwa SDN są kluczowe.

W ramach badań przeprowadzono serię eksperymentów z wykorzystaniem zaawansowanych modeli uczenia maszynowego tj. XGBoost, LightGBM, NGBoost, Decision Tree oraz Random Forest z naciskiem na ich wyjaśnialność. Istotnym wkładem jest:

- Opracowanie autorskiej metody selekcji cech SelectDVC oraz opracowanie autorskiego zbioru danych DVCDDoS2022.
- Analiza porównawcza algorytmów klasyfikacji zaadaptowanych do zadania detekcji ataków DDoS w sieciach definiowanych programowo.
- Analiza porównawcza wpływu cech sieciowych oraz ich redukcji na predykcje modeli.
- Opracowanie autorskich implementacji metryk D, F, S zdefiniowanych przez A. Rosenfelda oraz autorskich propozycji ich modyfikacji.
- Opracowanie autorskich implementacji metryk monotoniczności, wierności, niewierności oraz niekompletności.
- Opracowanie autorskich algorytmów ataków Flip-Label oraz oczyszczania zainfekowanych etykiet label-sanitization.

Łącznie w pracy przedstawiono osiem autorskich algorytmów.

Rozprawa doktorska została podzielona na dziesięć rozdziałów. Rozpoczyna się od wstępu, w którym zdefiniowano tezę, cel i zakres badań. Przedstawiono także strukturę pracy oraz omówiono oryginalne rozwiązania będące rezultatem badań. Rozdział 2 poświęcono porównaniu architektury i

zastosowania tradycyjnych sieci komputerowych z sieciami definiowanymi programowo. Omówiono w nim także zagrożenia związane z atakami DoS i DDoS, a także rolę wyjaśnialnej sztucznej inteligencji i koncepcji Data Centric AI, które zyskują w ostatnim czasie na znaczeniu. Na koniec przedstawiono problematykę Adversarial Machine Learning oraz Data Poisoning. W rozdziale 3 dokonano przeglądu literatury, analizując dotychczasowe badania nad detekcją ataków DDoS przy użyciu metod uczenia maszynowego. Szczególną uwagę poświęcono zastosowaniu metod wyjaśnialnej sztucznej inteligencji tj. SHAP w cyberbezpieczeństwie oraz wpływowi ataków zatruwania danych na skuteczność działania modeli. Rozdział 4 opisuje oryginalne środowisko badawcze, które zostało zbudowane z wykorzystaniem narzędzi Mininet, Ryu i Scikit-learn. Przedstawiono w nim również zaproponowaną architekturę systemu detekcji ataków DDoS oraz szczegóły dotyczące procesu zbierania danych i utworzenia autorskiego zbioru DVCDDoS2022. W rozdziale 5 skoncentrowano się na szczegółowej analizie zbioru DVCDDoS2022, uzasadniając jego powstanie w związku z ograniczeniami istniejących i dostępnych publicznie zbiorów danych. Przedstawiono również zastosowane metody selekcji cech tj. SelectKBest z zastosowaniem testu chi-kwadrat (chi2) i testu Anova, algorytm Recursive Feature Elimination (RFE), Principal Component Analysis (PCA) i narzędzie AutoML o nazwie FeatureWiz, a także zaproponowaną autorską metodę nazwaną SelectDVC. Rozdział 6 poświęcono ocenie algorytmów detekcji ataków DDoS, ze szczególnym uwzględnieniem algorytmu NGBoost. Dokonano jego porównania z innymi metodami klasyfikacji, analizując wpływ selekcji cech na wydajność oraz wyjaśnialność modeli. Zaadaptowano i wprowadzono metrykę D, służącą do oceny kompromisu między skutecznością a wyjaśnialnością modeli. W rozdziale 7 skupiono się na analizie wyjaśnialności modeli klasyfikacyjnych wobec ataków DDoS z wykorzystaniem metody SHAP oraz metryki F, badających wpływ poszczególnych cech na predykcje modeli oraz efekty stopniowego usuwania cech o niskiej ważności. Rozdział 8 rozszerza temat o analizę wpływu perturbacji danych wejściowych. Wprowadzono autorskie implementacje metryki monotoniczności, wierności, niewierności i niekompletności, które pozwalają ocenić stabilność wyjaśnień modeli. W rozdziale 9 przeanalizowano odporność modeli detekcji na ataki zatruwania danych i manipulacje etykietami. Przedstawiono autorskie modyfikacje algorytmów Flip-Label oraz metody label sanitization, a także zaproponowano metrykę S do oceny stabilności modeli w obliczu tego rodzaju zagrożeń. W zakończeniu dysertacji dokonano podsumowania, uwzględniając odpowiedzi na zasadnicze pytania badawcze oraz główne wnioski formułowane we wcześniejszych częściach pracy. Wskazano także kierunki dalszych eksploracji w podjętej problematyce. Podkreślono w nim skuteczność zaproponowanych autorskich algorytmów i koncepcji oraz znaczenie analizy wyjaśnialności modeli w zakresie cyberbezpieczeństwa sieci definiowanych programowo.

Zaprezentowane propozycje i badania oczywiście nie wyczerpują tematu, gdyż wciąż pozostaje on aktualny i otwarty na kontynuację. Teza przedstawiona na początku rozprawy doktorskiej jak najbardziej znajduje swoje potwierdzenie w niniejszej pracy. Badania wykazały, że:

- Zaproponowane metody selekcji cech w połączeniu z algorytmami klasyfikacji osiągają wysoki poziom efektywności w identyfikacji ataków typu DDoS, co potwierdza ich przydatność w praktycznych zastosowaniach sieciowych.
- Implementacja technik wyjaśnialnej sztucznej inteligencji umożliwiła dogłębną analizę i interpretację procesów decyzyjnych modeli, co przyczynia się do zwiększenia zaufania do stosowanych rozwiązań.
- Stwierdzono zróżnicowaną odporność modeli na ataki i perturbacje danych, co implikuje konieczność prowadzenia dalszych badań w celu zwiększenia ich stabilności i niezawodności w rzeczywistych warunkach operacyjnych.
- Wykazano istnienie kompromisu pomiędzy wydajnością, liczbą cech a wyjaśnialnością modeli, co wskazuje na zasadność stosowania kontekstowego podejścia i priorytetów w procesie selekcji cech i optymalizacji algorytmów.

W przyszłości planowane jest rozszerzenie badań o wdrożenie modelu detekcji w rzeczywistym środowisku akademickiej sieci definiowanej programowo. Kolejnym etapem będzie opracowanie i wdrożenie mechanizmów mitygacji i prewencji ataków, zintegrowanych z kontrolerem SDN, a także dalsza kontynuacja badań związana z wyjaśnialnością modeli. Ponadto rozważane są dalsze oceny środowiska detekcji na różnych zbiorach danych, innych algorytmach (np. sieciach typu transformer), ich hybrydowych połączeniach oraz w warunkach rzeczywistych.

14.03.2025