Politechnika Łódzka
Instytut Informatyki

Łódź, 15 kwietnia 2025 roku

prof. dr hab. inż. Adam Wojciechowski
Instytut Informatyki
Wydział Fizyki Technicznej, Informatyki i Matematyki Stosowanej
Politechnika Łódzka
Al. Politechniki 8, 93-590 Łódź

RECENZJA ROZPRAWY DOKTORSKIEJ

Tytuł rozprawy: **Analiza wyjaśnialności algorytmów uczenia maszynowego w zadaniach detekcji ataków w programowalnych sieciach komputerowych**

Autor rozprawy: **mgr inż. Stanisław Lota**

Promotor rozprawy: **prof. dr hab. inż. Leszek Rutkowski**

Jakie zagadnienie naukowe jest rozpatrywane w pracy (teza rozprawy) i czy zostało ono dostatecznie jasno sformułowane przez Autora? Czy tematyka rozprawy jest aktualna lub dostatecznie ważna? Jaki charakter ma rozprawa (teoretyczny, doświadczalny, inny)?

Pan mgr inż. Stanisław Lota podjął w swojej pracy doktorskiej zagadnienie wykrywania rozproszonych ataków odmowy usługi (DDoS) w sieciach komputerowych. Z racji na możliwość konfiguracji eksperymentów badawczych prace zostały przeprowadzone w środowisku sieci definiowanych programowo (SDN). Zagadnienie naukowe zostało precyzyjnie sformułowane i wyjaśnione.

W badaniach postawiono szereg pytań badawczych, które doprowadziły do postawienia hipotezy badawczej brzmiącej: „Autorskie propozycje i implementacje wyjaśnialnej sztucznej inteligencji umożliwiają skuteczne podejmowanie decyzji przez modele uczenia maszynowego w procesie detekcji rozproszonych ataków odmowy usługi w sieciach definiowanych programowo".

Tematyka rozprawy jest bardzo aktualna, ważna oraz dobrze postawiona. Rosnąca liczba cyberataków, szczególnie w czasach konfliktów i wojen hybrydowych, przy powszechnej informatyzacji usług oraz serwisów, staje się niezwykle ważnym i aktualnym problemem. Równoczesny rozwój technik sztucznej inteligencji w analizie wielowymiarowych danych sieciowych kieruje uwagę badaczy w stronę tworzenia rozwiązań wyręczających i zabezpieczających człowieka. Prace badawcze Doktoranta wpisują się w cenny nurt wykrywania ataków i zagrożeń w sieciach komputerowych.

Rozprawa ma charakter eksperymentalny i nosi w sobie istotny potencjał praktyczny. Przedstawione w pracy rozważania koncentrują się zarówno na weryfikacji popularnych rozwiązań literaturowych, w słabo zbadanym obszarze wykrywania ataków DDoS w sieciach definiowanych programowo, jak i wyjaśnialnej selekcji cech w zadaniu optymalizacji modeli klasyfikujących ataki odmowy usług. Przedstawione analizy są wystarczające i przekonywujące. Analiza wyników jest inspirująca i otwiera wiele kontekstów dla kolejnych badań. Praca z racji podjętej problematyki, zakresu oraz metodologii badawczej jest pełnoprawnym osiągnięciem o charakterze naukowym w dziedzinie nauk inżynieryjno-technicznych, w dyscyplinie Informatyka Techniczna i Telekomunikacja.

Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł (w tym literatury światowej i stanu zagadnień w przemyśle) świadczącą o dostatecznej wiedzy Autora? Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywujący?

W przedstawionej do oceny rozprawie doktorskiej, przegląd literatury został opisany głównie w rozdziałach 2 i 3. Rozdział 2 poświęcony został wyjaśnieniu podstawowych pojęć i terminów wykorzystywanych w pracy oraz uzupełniony wieloma rzeczywistymi przykładami obrazującymi rangę problemu. Moim zdaniem, trochę zabrakło w opisie ilościowej analizy typowych ataków DDoS, aby zbudować motywację dla wyboru cech opisujących ruch sieciowy i ich kluczowych charakterystyk. Kluczowa, z punktu widzenia oceny wyjaśnialności modeli sztucznej inteligencji, była metoda SHAP oraz prace Rosenfelda, który zaproponował ciekawe metryki oceny wyjaśnialności modeli AI, zaadaptowane później przez Doktoranta. Właściwy przegląd metod wykrywania ataków DDoS znajduje się w rozdziale 3. Myląca może być reporterska narracja towarzysząca przeglądowi, która przypisuje niektórym rozwiązaniom literaturowym blisko 100% skuteczność wykrywania ataków, co może podważać nieco zasadność prowadzenia badań w przedmiotowym zakresie. Przydałoby się nieco bardziej krytyczne spojrzenie na rozwiązania referencyjne. Ciekawe i inspirujące są rozważania na temat zastosowania metod XAI oraz wykorzystania ataków zatruwania danych w detekcji ataków DDoS. Być może warte rozwinięcia są niektóre stwierdzenia umieszczone w podsumowaniu przeglądu literaturowego (sekcja 3.4). Pojawił się w nim zarzut stosowania w referencyjnych badaniach analizy danych offline, co może ograniczać ich przydatność w warunkach rzeczywistych, podczas gdy w dalszej części rozprawy mamy w pewien sposób do czynienia z danymi offline (sekcja 4.1 - Wyniki zapisywane były w pliku CSV). Pojawia się też deklaracja zidentyfikowania luk badawczych w zadaniu detekcji ataków DDoS w sieciach SDN, które nie zostały wyraźnie sprecyzowane.

Bibliografia jest obszerna i liczy 140 pozycji. Doktorant nie raportuje wprawdzie prac opublikowanych z Jego udziałem, ale w bazie ORCID figurują dwie pozycje z udziałem Doktoranta, które ukazały się w latach 2015-2023. W dorobku znajduje się praca opublikowana na uznanej międzynarodowej konferencji *International Conference on Artificial Intelligence and Soft Computing*. Całokształt dokonań publikacyjnych Doktoranta jest formalnie wystarczający w świetle obowiązujących uregulowań.

Czy Autor rozwiązał postawione zagadnienie, czy użył właściwej do tego metody i czy przyjęte założenia są uzasadnione?

W obszarze podjętej problematyki, Doktorant systematycznie przeprowadził szereg eksperymentów, które były synergiczne z podjętymi pytaniami badawczymi.

W pierwszej kolejności Doktorant zaproponował środowisko badawcze oraz narzędzia, z których udziałem opracowano referencyjny, autorski zbiór danych. Architektura oprogramowanej topologii sieci SDN jest odpowiednia do badań w ramach postawionego problemu. Wątpliwości budzi jednak brak dokładnego opisu czyszczenia zebranych danych, co zdaniem Doktoranta miało „wyeliminować wszelkie niedoskonałości" (str. 44). Nie do końca precyzyjna jest też metodologia badawcza, bo zgodnie z deklaracją „dane z symulowanych ataków DDoS były przekształcane w zbiór treningowy, podczas gdy dane z bieżącego ruchu sieciowego stanowiły zbiór testowy" (str. 45). Jak w tym kontekście przebiegała pięciokrotna walidacja krzyżowa? Jak „zagwarantowano, że dane wykorzystywane do treningu są wolne od anomalii i błędów, które mogłyby negatywnie wpłynąć na skuteczność detekcji"? Według jakiego schematu przeprowadzano ataki DDoS i jaki to miało wpływ na balans klas w zebranych danych?

W konsekwencji Doktorant stworzył autorski zbiór danych o akronimie DVCDDoS2022. Deklaratywnie zbiór odzwierciedla przepływy typowe dla hybrydowych ataków DDoS oraz normalnego ruchu sieciowego. Nie wiadomo jednak co to dokładnie oznacza? Jak „typowość" przekłada się na zbalansowanie klas? Warto również mocniej uzasadnić zdystansowanie się od istniejących referencyjnych zbiorów danych (CICDDoS2019, CIC-IDS2017, CSE-CIC-IDS2018), gdyż powszechnie uznanym podejściem, ze względu na porównywalność wyników, jest testowanie opracowanych

rozwiązań zarówno na referencyjnych zbiorach danych jak i na autorskich. Co miał na myśli Autor pisząc „Ekstremalne manipulacje tymi zbiorami mogą ujawnić luki w systemach detekcji DDoS, co w konsekwencji obniży ich skuteczność w rzeczywistych warunkach"? Opracowany zbiór danych oryginalnie obejmował sekwencję wektorów cech, odzwierciedlających wartości 85 parametrów opisujących ruch w sieci SDN. Szkoda, że nie opisano precyzyjnie schematu ataków DDoS oraz nie wyjaśniono precyzyjnie dlaczego np. znacznik czasowy, odzwierciedlający sekwencję cech, nie miał znaczenia w przedmiotowym badaniu, skoro na str. 57 Doktorant podaje, że cechy czasowe znajdują się wśród czterech głównych kategorii cech. Wielkość wektora cech została słusznie poddana optymalizacji poprzez redukcję wymiarowości. Zastosowano w tym celu różne metody selekcji: eliminacji dziedzinowej cech nieistotnych, oceny istotności cech gotowymi narzędziami (*FeatureWiz*), czy systemowej selekcji według kryteriów ważności (m.in. *SelectKBest*, *RFE*, in). Wśród metod selekcji cech znajduje się rozwiązanie autorskie o akronimie *SelectDVC*. Ze względu na brak precyzyjnej definicji pojęć „zbiorów" i „list" w Algorytmie 1 nie są klarowne kryteria wyboru 19 cech. Czy jest to po prostu najbardziej reprezentatywna część wspólna zbiorów cech wyselekcjonowanych przez *FeatureWiz*, *SelectKBest_f_classif*, *SelectBest_chi2*, *RFE*, *PCA*?

Główne eksperymenty badawcze odnoszą się do weryfikacji 19 wyselekcjonowanych cech, lub ich podzbioru, w kontekście popularnych klasyfikatorów. Wyróżnione badania dotyczą możliwości użycia modelu *NGBoost* w detekcji ataków DDoS. Szkoda, że oprócz małej popularności algorytmu w przedmiotowej dziedzinie nie pojawiło się więcej argumentów za wyborem modelu *NGBoost*. Został on zestawiony z klasyfikatorami tj. *DecisionTree*, *XGBoost*, *LightGBM*, czy *Random Forest*, dla których to modeli zrezygnowano z dostrajania hiper-parametrów. Do ewaluacji metod selekcji cech i klasyfikatorów wykorzystano popularne metryki wydajności. Problem był zatem dwuwymiarowy, gdyż skuteczność wykrywania ataków DDoS zależy zarówno od selekcji analizowanych cech sieci jak i od samego klasyfikatora. Intencją Doktoranta było poszukiwanie optymalnego zestawu cech oraz dobór klasyfikatora, zapewniającego odpowiednią skuteczność wykrywania ataków.

Wyniki eksperymentów stanowią pogłębione studium przypadku i zostały wzbogacone analizą wyjaśnialności cech modelu. Przeprowadzono eksperyment referencyjny dla częściowo zredukowanego zbioru 77 cech oraz wybranych klasyfikatorów, po czym oceniono przydatność klasyfikatora *NGBoost* po ograniczeniu liczby cech za pomocą wybranych metod selekcji cech. Wyniki są ciekawe, choć nie zawsze spójne. Przykładowo wyniki z wykresów na rysunku 6.4 nie zgadzają się z wykresami z rysunku 6.5 (np. podpunkt b i c). Badania zostały przeprowadzone również w kontekście autorskiej implementacji metryki D, zaproponowanej przez Rosenfelda. Badania są inspirujące, chociaż zakres bezwzględnych wartości, zastosowanej metryki D, plasuje się na trudno interpretowalnych poziomach. Inspirująca jest również sekcja 6.6 poświęcona złożoności obliczeniowej. Porównania są ciekawe, ale mają jedynie charakter względny i trudno to przełożyć na realne konsekwencje praktyczne. Ciekawe byłoby odniesienie otrzymanych wartości do wielkości referencyjnych (literaturowych).

Najciekawsza, moim zdaniem, część badań poświęcona była wykorzystaniu wartości Shapleya (technika *TreeSHAP*) do analizy ważności cech wpływających na klasyfikację ataków DDoS oraz poszukiwań objaśnień modeli za pomocą wartości autorskiej metryki F, zainspirowanej pracami Rosenfelda. Stopniowe zawężanie listy cech według kryterium ich ważności, połączone z analizą dokładności modeli, analizą wartości metryki F oraz analizą kryteriów informacyjnych Akaike (AIC) i Bayesa (BIC) pozwala świadomie poszukiwać kompromisu pomiędzy liczbą cech i dokładnością modelu klasyfikatora. Ciekawa jest też analiza pozytywnych i negatywnych cech modeli.

Inspirujące są również badania związane z perturbacją lub zatruciem danych wejściowych, które bezsprzecznie wpływają na skuteczność klasyfikacji i jakość samego modelu. Wyjaśnienie wpływu

konkretnych schematów zaburzania/zmiany danych (dodanie losowego szumu, zamiana etykiet, zmiana etykiety 0 na 1, zmiana etykiety 1 na 0) na skuteczność klasyfikacji jest wartościowe i oryginalne. Doktorant wykorzystał w celu oceny zachowania modelu dedykowane adaptacje metryk monotoniczności, wierności, niekompletności oraz niewierności. Szkoda, że opis metryk nie został sformalizowany a opis eksperymentu nie jest do końca precyzyjny. Przykładowo, jak liczymy naruszalność i metrykę monotoniczności (str. 105)? Podobny niedosyt pojawia się przy omawianiu metryki wierności i niewierności. W części poświęconej zatruwaniu danych nie ma pewności na jakim etapie i w jakim zakresie przeprowadzano zmianę etykiet. Czy dotyczy to całości danych, czy tylko zbioru treningowego, a jeśli tak to jak przeprowadzano walidację krzyżową? Ciekawy jest w tym kontekście algorytm sanityzacji danych, bazujący na analizie skupień cech (k-NN z różnymi wariantami funkcji przeszukiwania), ale nieintuicyjny jest dobór wartości k (3,5,6,7,9), gdyż mamy do czynienia z zadaniem binarnej klasyfikacji. Warto byłoby doprecyzować również zasady tworzenia i zarządzania zbiorem próbek zaufanych w praktycznych scenariuszach. Przegląd wyników, uzupełnionych analizą macierzy pomyłek oraz analizą wydajności modeli, w oparciu o autorską metrykę S, dostarcza ciekawych konkluzji, na przykład w zakresie zamiany etykiet 1 na 0, które szczególnie obciążały „zatruty" model. Jednocześnie, sformułowanie ze strony 132, które sugeruje, że „klasa mniejszościowa jest bardziej podatna na zakłócenia" nasuwa ponownie pytanie o zbalansowanie procedowanego zbioru danych, które nie zostało jednoznacznie wyjaśnione.

Dysertację wieńczy rozdział podsumowujący wkład Doktoranta w dyscyplinę naukową. Zaproponowana została teza badawcza, którą w świetle wyników i analiz przeprowadzonych eksperymentów należy uznać za udowodnioną. Doktorant zaproponował również perspektywy dalszych badań.

Podsumowując, przedstawione rozwiązania, eksperymenty oraz wyniki są ciekawe i nowatorskie chociaż nie wszystkie wybory poczynione na etapie konfigurowania eksperymentów zostały umotywowane. Bezwzględne wartości skuteczności wykrywania ataków są, moim zdaniem, bardzo wysokie (ok. 98%) i mogą budzić wątpliwość na ile zaproponowane rozwiązania są generalizowalne oraz na ile konkretny typ ataków DDoS jest praktycznie trudny w wykryciu.

Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek Autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy, czy poziomu techniki reprezentowanych przez literaturę światową?

Oryginalność rozprawy doktorskiej leży przede wszystkim w szeroko ujętym zbiorze eksperymentów i analiz ukierunkowanych na skuteczne wykrywanie ataków DDoS w programowalnych sieciach komputerowych. W szczególności najważniejszymi, moim zdaniem, osiągnięciami są:

- opracowanie autorskiego zbioru danych pozwalającego rozszerzyć badania nad atakami typu DDoS w programowalnych sieciach komputerowych;

- analiza porównawcza metod selekcji cech wraz z wiarygodną oceną ich wpływu na skuteczność wykrywania ataków DDoS w wybranych algorytmach klasyfikacji, co przekłada się na wyjaśnialność modeli klasyfikacji i precyzyjny dobór znaczących cech ruchu sieciowego;

- adaptacja metryk D, F, S, zdefiniowanych przez A. Rosenfelda oraz metryk monotoniczności, wierności, niewierności oraz niekompletności na potrzeby oceny wydajności i złożoności modeli klasyfikacyjnych w kontekście wyselekcjonowanych cech;

- badania nad wpływem perturbacji danych oraz zmian etykiet na skuteczność wykrywania ataków typu DDoS;

Wymienione osiągnięcia stanowią oryginalny wkład Doktoranta w rozwój dyscypliny Informatyka Techniczna i Telekomunikacja.

Czy Autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?

Praca stanowi pogłębione opracowanie z zakresu systemowej, podbudowanej istotnością, selekcji cech ruchu sieciowego wraz z analizą ich wpływu na skuteczność wykrywania rozproszonych ataków odmowy usług w programowalnych sieciach komputerowych. Przeprowadzone eksperymenty zostały przeprowadzone prawidłowo, zaś wyniki zostały przeanalizowane wieloaspektowo i dogłębnie. Pewien niedosyt pozostawia brak ilościowego osadzenia badań na tle współczesnych osiągnięć literaturowych, ale same badania są przekonywujące i na pewno warte opublikowania. Układ pracy jest spójny i logiczny. Pracę czyta się bardzo dobrze. Pod względem edycyjnym praca została zredagowana starannie, choć nie udało się uniknąć drobnych błędów językowych: „w wstecz" (str. 58), „dorównują wydajnościom" (str. 83), osie wykresu z rysunku 6.2 nie zostały opisane; we wzorze 7.1 nie objaśniono zmiennej „n", in.

Jakie są słabe strony rozprawy i jej główne wady?

Rozprawa doktorska nie zawiera istotnych uchybień, a poniższe uwagi mają przede wszystkim charakter dyskusyjny. Przykładowe kwestie, które mogłyby zostać omówione, lecz nie ujmują pozytywnej oceny rozprawy, dotyczą:

- omówieniu realnych trudności i wyzwań przy wykrywaniu ataków DDoS oraz przedyskutowaniu realnych przewag zaproponowanego rozwiązania nad rozwiązaniami literaturowymi;

- próby przedyskutowania bezwzględnych wartości zastosowanych metryk w kontekście wymagań dziedzinowych wykrywania ataków;

- doprecyzowania metodologii konstruowania zbioru danych, szczególnie w zakresie scenariusza symulowania ataków DDoS i zbalansowania zbioru danych oraz metodologii prowadzenia eksperymentów, w czasie których dane były poddawane dodatkowym modyfikacjom/perturbacjom;

- przedyskutowania doboru parametru k w algorytmie czyszczenia (sanityzacji) etykiet oraz ogólnie doprecyzowania etapów czyszczenia danych i wpływu tego procesu na wiarygodność wyników;

Konkluzja

Uważam, że cele rozprawy doktorskiej zostały w pełni zrealizowane, a postawiona teza udowodniona. Doktorant przedstawił w rozprawie nowe i oryginalne rozwiązania wyjaśniające wykrywanie rozproszonych ataków odmowy usług. Uzyskane przez Doktoranta wyniki uważam za oryginalne, wartościowe i ciekawe poznawczo. Zakres prowadzonych badań był szeroki, analiza opracowanych rozwiązań wszechstronna, a warsztat badawczy Doktoranta jest odpowiedni.

Tym samym rozprawa prezentuje wartościowe osiągnięcia naukowe w obszarze dyscypliny Informatyka Techniczna i Telekomunikacja oraz potwierdza umiejętność samodzielnego prowadzenia przez Doktoranta pracy naukowej.

Bez najmniejszej wątpliwości stwierdzam, że recenzowana rozprawa doktorska Pana mgra inż. Stanisława Lota spełnia wymagania stawiane rozprawom doktorskim, przez obowiązującą ustawę. Wnoszę o jej przyjęcie i dopuszczenie rozprawy do publicznej obrony.

Adam Wojciechowski