Dr Antonina Krajewska
Naukowa i Akademicka Sieć Komputerowa - Państwowy Instytut Badawczy (NASK-PIB)
ORCID: 0000-0001-6626-5667
e-mail: antonina.krajewska@nask.pl

ANALIZA I KLASTERYZACJA RUCHU SIECIOWEGO Z ROZPROSZONEGO SYSTEMU PUŁAPEK HONEYPOT

Streszczenie

Algorytmy klasteryzacji pełnią kluczową rolę w analizie dużego wolumenu ruchu i wykrywaniu wzorców nieznanych ataków. W pracy omówiono problem grupowania danych zebranych przez rozproszony system pułapek sieciowych. W proponowanym podejściu, zgodność przepływu sieciowego z sygnaturą ataku jest równoważna nadaniu mu odpowiedniej etykiety. Dzięki temu możliwe jest zastosowanie algorytmów uczenia częściowo nadzorowanego oraz poprawa jakości wyników klasteryzacji. W artykule porównano wyniki algorytmów uczenia przeprowadzonego bez oraz z częściowym nadzorem w zadaniu grupowania przepływów sieciowych.

Słowa kluczowe: cyberbezpieczeństwo, honeypoty, sztuczna inteligencja, uczenie z częściowym nadzorem, klasteryzacja.

WPROWADZENIE

Rozwój i wszechobecność systemów teleinformatycznych powodują, że stają się one coraz częstszym celem ataków cybernetetycznych. W celu ochrony infrastruktury telekomunikacyjnej, inżynierowie i specjaliści ds. bezpieczeństwa stale opracowują nowe metody i narzędzia. Od zaawansowanych algorytmów szyfrowania po systemy detekcji zagrożeń oparte na Sztucznej Inteligencji (SI), rozwiązania z zakresu cyberbezpieczeństwa są niezbędnym elementem walki z coraz bardziej wyrafinowanymi i złożonymi zagrożeniami[1].

Rozproszone systemy honeypotów należą do klasy rozwiązań zwiększających świadomość sytuacyjną. Honeypoty, czyli pułapki sieciowe, są specjalnie zaprojektowanymi systemami, symulującymi infrastrukturę teleinformatyczną w celu zwabienia i monitorowania działań potencjalnych cyberprzestępcow. Rozproszone systemy honeypotów są w stanie analizować

1 E. Niewiadomska-Szynkiewicz, R. Litka, Ataki na urządzenia mobilne i metody ich wykrywania, Cybersecurity and Law 2023, 9(1), s. 95-107.

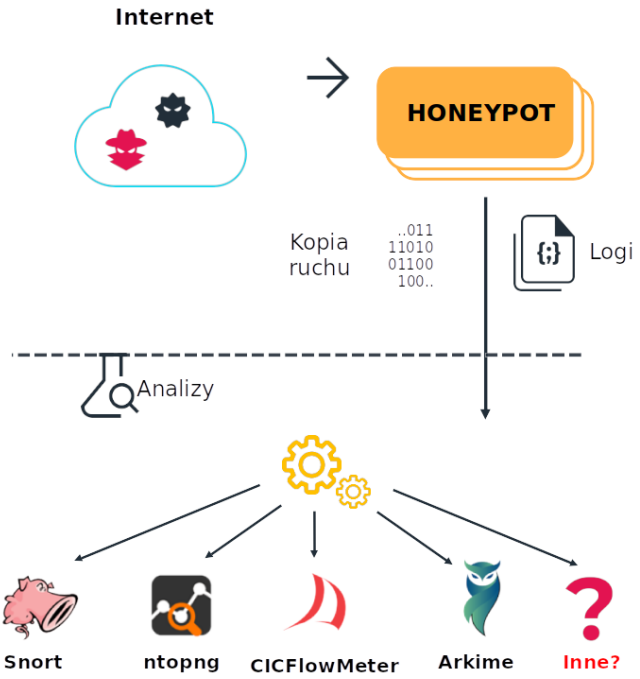
duże ilości danych w czasie rzeczywistym. Dzięki zaawansowanym algorytmom uczenia maszynowego, identyfikują wzorce ataków oraz adaptują się do zmieniających się taktyk przeciwnika. W rezultacie, systemy teleinformatyczne są lepiej chronione, a specjaliści ds. bezpieczeństwa mogą szybciej reagować na nowe i nieznane zagrożenia.

Kluczową rolę w działaniu takich systemów pełnią algorytmy klasteryzacji. Umożliwiają one wykrywanie wzorców nowych ataków, detekcję anomalii, a także poprawiają wydajność systemu. Gwałtowny rozwój technik sztucznej inteligencji (SI) pozwolił na znaczną poprawę efektywności metod klasteryzacji. W pracy przedstawiono przykład zastosowania algorytmu klasteryzacji z częściowym nadzorem w systemie pułapek sieciowych. W pierwszej części, omówiono znaczenie analizy pełnego zapisu ruchu sieciowego odbieranego przez honeypoty, a także omówiono problem reprezentacji danych. Następnie pokazano rolę algorytmów klasteryzacji w rozproszonych systemach pułapek sieciowych. W kolejnej części przedstawiono propozycję wykorzystania dopasowania przepływu sieciowego do sygnatury IDS[2] jako mechanizmu etykietującego dane, a także opisano kilka wybranych algorytmów klasteryzacji. Na zakończenie przedstawiono wyniki eksperymentów porównujących jakość rezultatów klasteryzacji przeprowadzonej bez i z częściowym nadzorem.

ANALIZA RUCHU SIECIOWEGO Z HONEYPOTÓW

Honeypoty rejestrują zdarzenia sieciowe oraz umożliwiają analizę poleceń wykonanych przez atakującego[3]. Skuteczna korelacja i agregacja zebranych przez nie danych, istotnie zwiększa świadomość sytuacyjną operatora systemu. Niemniej jednak, dane rejestrowane przez oprogramowanie honeypot zawierają tylko część informacji zawartych w pełnym zapisie ruchu sieciowego. Co więcej, ich format i treść zależą od wersji honeypota oraz są podatne na błędy w oprogramowaniu. Aby w pełni wykorzystać potencjał rozproszonego systemu pułapek sieciowych, należy również analizować zapis ruchu sieciowego (rysunek 1).

2 Intrusion Detection System – system wykrywający atak na systemy teleinformatyczne
3 J. Skłodowski, P. Arabas, Wykorzystanie drzew sufiksowych do efektywnej prezentacji podobieństw sesji z systemu pułapek honeypot. Cybersecurity and Law, 9(1), 298-315. 2023.



Źródło: opracowanie własne
Rys. 1. Schemat przepływu danych w systemie

Obecnie dostępnych jest wiele technologii umożliwiających wydajną analizę ruchu sieciowego, takich jak Arkime[4], Snort[5], Ntopng[6] czy CicFlowMeter[7]. Ponadto, charakterystyka ruchu sieciowego zebranego z rozproszonego systemu honeypotów sprawia, że stanowi on wartościowy zbiór danych dla algorytmów SI. Honeypoty obsługują wybrane protokoły sieciowe, takie jak HTTP, HTTPS, FTP, czy SSH. Dzięki temu dane mają ściśle określoną, przewidywalną charakterystykę. Ponieważ ruch ten jest z założenia nieprawidłowy, zbiór danych może być wykorzystany przez algorytmy uczenia nienadzorowanego.

Jakość wyników analiz zależy od wybranej reprezentacji ruchu sieciowego (rysunek 2). W literaturze najczęściej spotykane są dwa podejścia[8]. W pierwszym, analizie poddawane są nagłówki oraz payload poszczególnych pakietów sieciowych. W drugim, analizowane są przepływy (ang. flow), czyli zagregowane dane pakietów. Najczęściej spotykanym kryterium agregacji jest zestaw pięciu cech: źródłowy adres IP, port źródłowy, docelowy adres IP, port

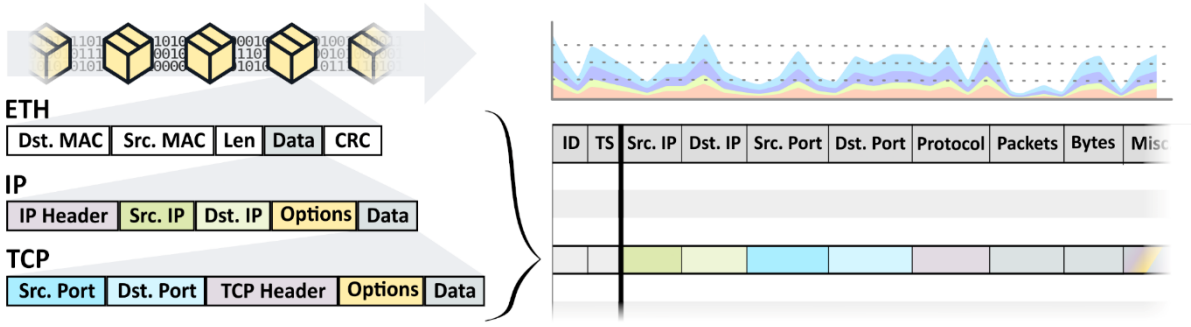
4 https://arkime.com/
5 https://www.snort.org/
6 https://www.ntop.org/
7 https://github.com/ahlashkari/CICFlowMeter
8 Ring M. i in., A Survey of Network-based Intrusion Detection Data Sets. Comput. Secur. 86: 147-167 2019.

docelowy oraz protokół. Rozróżniane są zarówno przepływy jedno jak i dwukierunkowe. Oprócz wymienionych atrybutów, dane zawierają statystyki, takie jak czas trwania komunikacji, liczba przesłanych bajtów na sekundę, długość odstępów czasowych pomiędzy wysyłanymi komunikatami, wielkość i liczba przesłanych pakietów.

Zaletami reprezentacji ruchu sieciowego w postaci przepływów jest mniejszy koszt przetwarzania i przechowywania danych, a także mniejsza podatność na wartości odstające. Ponieważ dane nie zawierają payloadu, system przetwarza i gromadzi mniej danych wrażliwych. W przypadku komunikacji szyfrowanej, nie ma potrzeby odszyfrowywania payloadu pakietów. Taka reprezentacja ruchu sieciowego jest naturalnie odwzorowana w przestrzeni wektorowej. Dzięki temu, zastosowanie algorytmów uczenia maszynowego nie wymaga skomplikowanego wstępnego przetwarzania danych (ang. *data preprocessing*).

Źródło: opracowanie własne

Rys. 2. Reprezentacja ruchu sieciowego – pakiety vs. Przepływy



ROLA ALGORYTMÓW KLASTERYZACJI W ROZPROSZONYM SYSTEMIE PUŁAPEK SIECIOWYCH

Klasteryzacja to metoda uczenia bez nadzoru, polegająca na grupowaniu elementów w jednorodne klasy. Algorytmy klasteryzacji pełnią kluczową rolę w procesie generowania sygnatur, czyli znajdowaniu wzorców ataków sieciowych, zawierających formalny opis cech danej kampanii.

Zaproponowany przez Wernera i in. algorytm Nebula[9], w pierwszym kroku grupuje pakiety złośliwego ruchu sieciowego. Następnie, generowany jest zestaw sygnatur opisujących każdy z klastrów. Utworzone w ten sposób sygnatury składają się z ciągów fragmentów payloadu pakietów. Każdy element ciągu ma zdefiniowane dopuszczalne początkowe i końcowe położenie w payloadzie. Takie reguły można zapisać w formacie obsługiwanym przez system IDS, Snort:

9 Werner, T., i in. Nebula-generating syntactical network intrusion signatures 2009 4th International Conference on Malicious and Unwanted Software (MALWARE). IEEE 2009.

alert udp any any −> $HOME_NET 53 (msg : " "; content :"|01 00 00 00|"; offset : 0 ; depth : 11; content : "|07| in−addr | 0 4 | arpa |00 00 0c 00 01|"; distance : 12; within : 39; sid : 5011237; rev : 3 ;)

Najczęstszym przypadkiem użycia tak generowanych sygnatur jest ich automatyczne wdrożenie w systemach IDS lub IPS[10], w celu powstrzymania epidemii nieznanego wcześniej zagrożenia. Takie zastosowanie wymaga sygnatur bardzo wysokiej jakości, która zależy od wykorzystanych algorytmów klasteryzacji.

Klasteryzacja przepływów lub pakietów sieciowych umożliwia również poprawę jakości sygnatur znanych ataków, a także ułatwia operatorowi systemu klasyfikację zdarzeń sieciowych. Ruch sieciowy odbierany przez system niekoniecznie musi być atakiem. Część połączeń z honeypotami wynika z błędów konfiguracji urządzeń sieciowych. Skuteczne algorytmy klasteryzacji ułatwiają określenie, czy dane zdarzenie sieciowe to atak wymagający dalszej analizy.

KLASTERYZACJA PRZEPŁYWÓW Z CZĘŚCIOWYM NADZOREM

W ostatnich latach coraz większą popularność zyskują algorytmy uczenia z częściowym nadzorem (ang. *Semi-Supervised Learning*)[11]. Stosuje się je, gdy część danych jest etykietowana. Dzięki temu, możliwe jest grupowanie nie tylko na podstawie wzajemnego podobieństwa elementów, lecz również na podstawie relacji między nimi, na przykład:

- obiekty z etykietą A powinny być grupowane razem,
- obiekt z etykietą A i obiekt z etykietą B, powinny należeć do różnych klastrów.

Jednym z algorytmów klasteryzacji z częściowym nadzorem jest wprowadzony przez Lee i in.[12] algorytm Semi-Supervised Nonnegative Matrix Factorization (SSNMF). Opisana metoda jest rozszerzeniem klasycznego algorytmu Nonnegative Matrix Factorization (NMF). Algorytm SSNMF został rozwinięty przez Haddocka i in.[13] o różne miary podobieństwa obiektów.

Wykorzystanie algorytmu SSNMF w zadaniu klasteryzacji przepływów wymaga, aby część z nich była prawidłowo oznaczona. Proponowane podejście zakłada wykorzystanie w tym celu oprogramowania IDS. Zgodność ruchu

10 Intrusion Prevention System – system zapobiegający atakowi na systemy teleinformatyczne.

11 Zhu, X., Goldberg, A. B., Introduction to semi-supervised learning. Springer Nature 2022.

12 Lee H., i in., Semi-Supervised Nonnegative Matrix Factorization. IEEE Signal Processing Letters 17: 4-7 2010.

13 Haddock, J, i in., Semi-supervised nonnegative matrix factorization for document classification. 55th Asilomar Conference on Signals, Systems, and Computers. IEEE 2021.

z sygnaturą znanego ataku, oznaczałaby nadanie mu etykiety reprezentującej ten atak. Należy w tym celu wykorzystać sygnatury wysokiej jakości, czyli te udostępniane w oficjalnych, publicznych repozytoriach dostawców systemu lub opracowane przez ekspertów ds. bezpieczeństwa.

Motywacją opisanej metodologii jest fakt, że część połączeń nawiązywanych z honeypotami, to znane i opisane ataki. Przepływy reprezentujące dany atak powinny być umieszczane w jednym klastrze. Nie powinny się natomiast w nim znaleźć przepływy reprezentujące ruch zgodny z sygnaturą innego ataku.

ANALIZA PORÓWNAWCZA ALGORYTMÓW KLASTERYZACJI

W wykonanych badań przeprowadzono liczne eksperymenty w celu porównania skuteczności klasteryzacji bez nadzoru (NNMF) oraz częściowo nadzorowanej (SSNMF). W obu przypadkach uwzględniono dwie miary podobieństwa obiektów: normę euklidesową i dywergencję Kullbacka-Leiblera.

Kod obliczeń został zaimplementowany w języku Python 3.9 oraz wykonany na stacji roboczej z system Linux, wyposażonej w procesor 11th GenIntel(R) Core(TM) i7-1165G7 @ 2.80GHz (8 rdzeni) oraz 32GB RAM-u. Kod wykorzystuje biblioteki scikit-learn [14] oraz ssnmf[15].

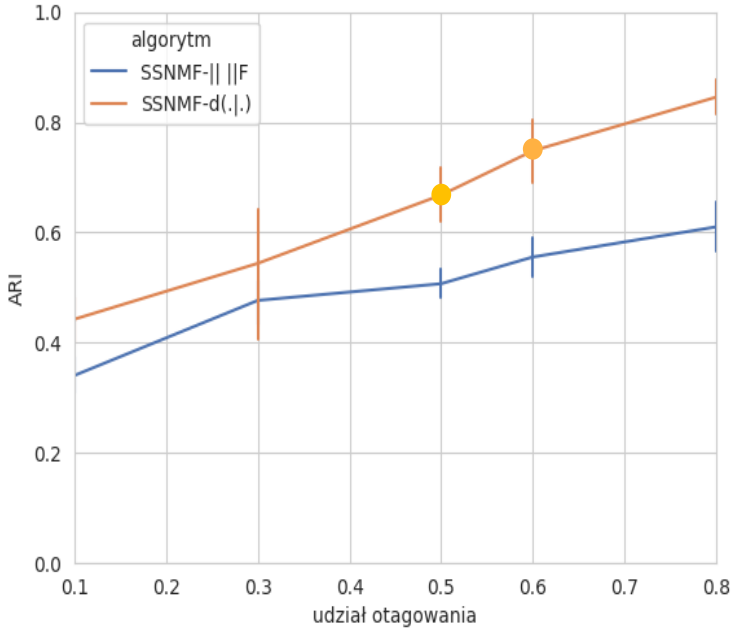
Wybrane algorytmy klasteryzacji były oceniane na danych ze zbioru CIC-IDS2017[16]. Zbiór zawiera zapis ruchu sieciowego w postaci plików PCAP wraz metadanymi otrzymanymi za pomocą narzędzia CICFlowMeter. Etykiety przepływów odpowiadają wybranym 12 atakom (w tym DoS, DoS Slowhttp, Heart-bleed, Brute Force, skanowania) lub oznaczają ruch niezłośliwy. Przepływy zostały scharakteryzowane przez 80 cech, w tym: źródłowy adres IP, port źródłowy, docelowy adres IP, port docelowy, protokół oraz czas trwania komunikacji. Ponadto udostępniono wartości: maksymalną, minimalną, średnią i odchylenie standardowe wielkości takich jak: rozmiar, liczba przesłanych pakietów, czy liczba poszczególnych flag TCP.

W pierwszym etapie ze zbioru danych CICIDS2017 zostały wybrane cechy mające charakter ilościowy. Następnie, odfiltrowano ruch niezłośliwy. Kolumny macierzy danych zostały znormalizowane, aby ich norma euklidesowa wynosiła 1. Poprawność wyników klasteryzacji została zweryfikowana przez miarę zgodności klasteryzacji z etykietami Skorygowany Indeks Randa (ang. Adjusted Rand Index – ARI).

Wartości ARI wyników klasteryzacji NMF wynosiły 0.51 i 0.49 odpowiednio dla miary podobieństwa opartej o normę euklidesową i dywergencję Kullbacka-Leiblera. Rysunek 4. przedstawia wartości ARI dla

14 Scikit-learn: Machine Learning in Python, Pedregosa i in. JMLR 12, pp. 2825-2830 2011.
15 Haddock J., i in., Semi-supervised Nonnegative Matrix Factorization Models for Topic Modeling in Learning Tasks. Submitted 2020.
16 Sharafaldin I., Lashkari A. H., i Ghorbani A. A, Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization, 4th International Conference on Information Systems Security and Privacy (ICISSP) 2018.

dwóch wariantów algorytmu SSNMF: SSNMF-||.||F i SSNMF-d(.|.), w których podobieństwo między obiektami liczone jest odpowiednio na



podstawie normy euklidesowej i dywergencji Kullbacka-Leiblera. W przypadku, etykietowania połowy danych algorytm SSNMF-d(.|.) osiągnął ARI równe 0.67. W przypadku etykietowania 0.6 przepływów ARI osiągnęło wartość 0.75.

Źródło: opracowanie własne

Rys. 3. Wyniki klasteryzacji z częściowym nadzorem

Wyniki testów przedstawione na rysunku 3 potwierdzają pozytywny wpływ etykietowania części przepływów na jakość klasteryzacji. Jednocześnie pokazują znaczenie wyboru miary podobieństwa między obiektami. Należy jednak zaznaczyć, że udział etykietowanych danych, przy których widać poprawę jakości klasteryzacji jest wysoki. Dalsze prace będą obejmowały porównanie innych metod klasteryzacji przeprowadzonej z częściowym nadzorem.

PODSUMOWANIE

Zbieranie i analiza danych z systemów pułapek sieciowych umożliwiają identyfikację zagrożeń, zrozumienie taktyk atakujących oraz doskonalenie strategii obronnych. Klasteryzacja ruchu sieciowego z rozproszonego systemu honeypotów pełni kluczową rolę w identyfikacji nowych ataków. Ponieważ część ruchu odbieranego przez honeypoty stanowią znane ataki, zgodność z sygnaturą może być wykorzystana jako etykietowanie w algorytmie uczenia. Wyniki przeprowadzonych badań pokazują, że jakość wyników klasteryzacji

jest wyższa w przypadku gdy uczenie przeprowadzono z częściowym nadzorem.

Bibliografia

Haddock J., i in., Semi-supervised Nonnegative Matrix Factorization Models for Topic Modeling in Learning Tasks. Submitted, 2020.

Haddock, J, i in., Semi-supervised nonnegative matrix factorization for document classification. 55th Asilomar Conference on Signals, Systems, and Computers. IEEE, 2021.

Lee H., i in., Semi-Supervised Nonnegative Matrix Factorization. IEEE Signal Processing Letters 17, 2010.

Niewiadomska-Szynkiewicz E, Litka R. Ataki na urządzenia mobilne i metody ich wykrywania. Cybersecurity and Law ;9(1):95-107, 2023.

Ring M. i in., A Survey of Network-based Intrusion Detection Data Sets. Comput. Secure. 86, 2019.

Sharafaldin I., Lashkari A. H., Ghorbani A. A., Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization, 4th International Conference on Information Systems Security and Privacy (ICISSP), 2018.

Skłodowski, J., Arabas, P. Wykorzystanie drzew sufiksowych do efektywnej prezentacji podobieństw sesji z systemu pułapek honeypot. „Cybersecurity and Law", 9(1), 2023

Werner, T., i in. Nebula-generating syntactical network intrusion signatures. 2009 4th International Conference on Malicious and Unwanted Software (MALWARE). IEEE, 2009.

Zhu, X., Goldberg, A. B. Introduction to semi-supervised learning. Springer Nature, 2022.

ANALYSIS AND CLUSTERING OF NETWORK TRAFFIC FROM A DISTRIBUTED HONEYPOT SYSTEM

Abstract

Clustering algorithms play a crucial role in detecting patterns of unknown attacks. The paper discusses the problem of clustering data collected by a distributed system of network traps. In the proposed approach, the flow conformity to the attack signature is equivalent to assigning it an appropriate label. This allows for the application of semi-supervised learning algorithms and the improvement of clustering quality. The article compares the results of learning algorithms conducted with and without partial supervision in clustering network flows.

Keywords: cybersecurity, cyberattack honeypot, artificial intelligence, semi-supervised learning, clustering.