Metody AI w badaniu zagrożeń w systemach komputerowych

Wprowadzenie do
- Tematyki kursu
- Projektu
- Seminarium

Autor: **Henryk Maciejewski**
Mateusz Gniewkowski
Tomasz Walkowiak

CYBERBEZPIECZEŃSTWO 2.0

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego, Program Operacyjny Wiedza Edukacja Rozwój, Priorytet III Szkolnictwo Wyższe dla gospodarki i rozwoju, Działanie 3.5 Kompleksowe programy szkół wyższych w ramach konkursu nr POWR.03.05.00-IP.08-00-PZ3/18 na Zintegrowane Programy uczelni – Ścieżka III, nr umowy POWR.03.05.00-00-Z308/18-00
Tytuł projektu: „Cyberbezpieczeństwo dla gospodarki przyszłości”

Organizacja kursu

- Wykład – 2h co tydzień, 30h
- Projekt – 3h co tydzień, 30 h
- Seminarium – 1h co tydzień, 15h

- Wykład prowadzony przez
  - Henryk Maciejewski
    Mateusz Gniewkowski
    Tomasz Walkowiak
- Projekt + seminarium
  - 3 strumienie, studenci w tych samych grupach P+S
  - każdy strumień prowadzony przez innego prowadzącego

Zasady zaliczeń

Ocena z kursu:

- średnia ocen z poszczególnych form = 1/3 * (P+S+W), o ile P>2 i S>2 i W>2

Ocena z wykładu:

- kolokwium po ostatnim wykładzie

Ocena z seminarium:

- grupy 2-3 osobowe, ocenianie wspólnie, każda grupa wykona 3 prezentacje związane z wykonywanym projektem
- średnia ocena z prezentacji, przy **max. 1** nieobecności

Ocen z projektu:

- ta sama grupa co na seminarium
- kamienie milowe i spotkania w grupach projektowych (update, 2 tyg)
- raport końcowy:
  - opis problemu, potok przetwarzania, ewaluacja metod, czas działania (predykcja), porównanie metod, analiza wyników, wnioski, literatura + załączony kod eksperymentów
- możliwa ocena 5.5: próba wdrożenia, przykład użycia, przetestowanie na innym zbiorze

Cele przedmiotu / tematyka kursu

- Metody sztucznej inteligencji (AI) i metod uczenia maszynowego (ML) wykorzystywane w wykrywaniu zagrożeń / ataków na systemy komputerowe

- Metody wykrywania anomalii / nietypowych profili w oparciu o dane z monitoringu ruchu sieciowego, monitoringu zdarzeń i obciążenia urządzeń i z innych źródeł

Przykłady źródeł danych:

tekst (sms, email), potencjalnie niebezpiecznie URL, ciągi czasowe, transakcje kartą płatniczą, połączenia w sieci LAN, …

Tematyka kursu – metody / zagadnienia AI/ML

- Metody uczenia nadzorowanego
- Uczenie nienadzorowane
- Metody redukcji wymiaru danych / wyboru cech
- Metody modelowanie szeregów czasowych (ARIMA)
- Uczenie głębokie w analizie zagrożeń
- Wykrywanie anomalii (nadzorowane, nienadzorowane, w szeregach czasowych)
- Uczenie w oparciu o dane *class-imbalanced*
- Metody reprezentacji tekstu dla potrzeb uczenia maszynowego (NLP)
- Interpretowalność modeli

Cel **projektu**:

Poznanie i praktyczne wykorzystanie metod i algorytmów uczenia maszynowego w kontekście danych związanych z cyberbezpieczeństwem.

Realizacja projektu w zespole 2-3 osobowym.

Seminarium: prezentacja (i) tematyki związanej z zadaniem projektowym, (ii) przebiegu i wyników zadania projektowego

Projekt – etapy realizacji

1. Zapoznanie się ze zbiorem danych (zrozumienie danych, cech i typów ataków/etykiet jaki dany zbiór zawiera)
2. Zapoznanie się z literaturą dot. zbioru danych
3. Eksploracyjna analiza danych (jakość, kompletność, zrównoważenie klas, …)
4. Przygotowanie danych (preprocessing, generacja cech, wybór cech, redukcja wymiaru, zbalansowanie, …)

Projekt – etapy realizacji

5. Zbudowanie kilku klasyfikatorów, badanie wpływu metaparametrów (*model selection*)

   a. Do ewaluacji zbiorów danych należy użyć metryk odpowiednich dla zbioru
   b. Porównanie czasu uczenia i inferencji klasyfikatorów
   c. Do analizy można użyć również metod nienadzorowanych
   d. Jeżeli jest to możliwe, należy porównać wyniki z opisanymi w literaturze
   e. Analiza istotności cech - które cechy są istotne dla anomalii

6. Zbadanie metod nienadzorowanych dla AD

   - dla danych o różnych charakterystykach (wymiar, typ danych) – zbiory pkt. 8
   - metody oparte na sąsiedztwie/gęstości, oparte na klasteryzacji, metody dla szeregów czasowych)

Uwaga – wybieramy zadanie 5 albo 6

7. Opracowanie raportu – opis wyników powyższych punktów

Każda grupa wygłasza ~~trzy~~ dwie prezentacje:

1. **Prezentacja wstępna**: opis zadania i zbioru danych; opis typów ataków/anomalii jakie występują w danym zbiorze, wyjaśnienie ich w kontekście cyberbezpieczeństwa

2. **Prezentacja śródokresowa:** wstępne wyniki, eksploracja danych, wstępne wyniki klasyfikacji, plan dalszych badań (*sensitivity studies*)

3. **Końcowa prezentacja:** podsumowanie projektu, wyniki badań, analiza cech (interpretowalność modeli), wnioski

Projekt – narzędzia

1. Sugerowanym środowiskiem pracy jest Python i Jupyter Notebook
2. Można korzystać z innego języka programowania/środowiska, ale kod, który dostarczamy powinien być możliwy do uruchomienia
3. Ważnym elementem prac jest ich reprodukowalność

Notatniki Jupyter z przykładami metod (omówimy je na pierwszych zajęciach):

1. Klasyfikacja danych tabelarycznych / *class-imbalanced*
2. Klasyfikacja danych tekstowych
3. Klasyfikacja – szeregi czasowe
4. Użycie metod nienadzorowanych

0. Python w przykładach (na dobry początek)

Projekt – proponowane zbiory danych

1. Analiza niebezpiecznych adresów URL

Zbiór: https://www.unb.ca/cic/datasets/url-2016.html
Uwaga: zbiór zawiera również surowe dane (adresy URL) - zachęcamy, aby z nich skorzystać.

2. Analiza botów z Twittera

Zbiór: https://www.kaggle.com/c/utkmls-twitter-spam-detection-competition

3. System wykrywania intruzów (IDS)

Zbiór: https://www.unb.ca/cic/datasets/ids-2017.html

4. Analiza działania Malware w systemie Android

Zbiór: https://www.unb.ca/cic/datasets/andmal2017.html

Projekt – proponowane zbiory danych

5. Wykrywanie ataków DDoS:
Zbiór: https://www.unb.ca/cic/datasets/ddos-2019.html

6. Analiza ruchu DoH
Zbiór: https://www.unb.ca/cic/datasets/dohbrw-2020.html

7. Analiza fałszywych transakcji kartą płatniczą
Zbiór: https://www.kaggle.com/mlg-ulb/creditcardfraud

8. Zbiory danych z różnych dziedzin do badania metod nienadzorowanych wykrywania anomalii
Repozytorium: https://odds.cs.stonybrook.edu/

Projekt – proponowane zbiory danych

9. Adversarial Attacks - ataki na klasyfikatory obrazów

Artykuły: https://robustbench.github.io/#div_cifar10_corruptions_heading
Zbiory: https://www.cs.toronto.edu/~kriz/cifar.html

10. Adversarial Attacks - ataki na klasyfikatory tekstowe

Artykuły:
https://arxiv.org/abs/1812.05271
https://www.researchgate.net/publication/374307223_Do_Not_Trust_Me_Explainability_Against_Text_Classification

Kod: https://github.com/thunlp/OpenAttack
Zbiory: https://huggingface.co/datasets/rotten_tomatoes
https://huggingface.co/datasets/imdb

11. Inny zbiór zaakceptowany przez prowadzącego