Michał Ściubisz
Wojciech Tokarz

Zastosowanie algorytmu TOPSIS do selekcji cech w systemach wykrywania intruzji na bazie zbioru danych NSL-KDD

Wprowadzenie

Systemy wykrywania intruzji (IDS) odgrywają kluczową rolę w ochronie sieci przed atakami i nieautoryzowanym dostępem. Współczesne IDS muszą przetwarzać miliony pakietów danych z licznymi cechami, co znacząco wydłuża czas detekcji anomalii. W związku z tym niezbędne staje się opracowanie efektywnych metod redukcji wymiarowości danych, które pozwolą na szybszą i bardziej precyzyjną identyfikację potencjalnych zagrożeń.

W ramach tego projektu wykorzystamy zmodyfikowany zbiór danych NSL-KDD oraz techniki selekcji cech oparte na algorytmie TOPSIS (ang. *Technique for Order of Preference by Similarity to Ideal Solution*). Celem jest zbadanie wpływu różnych metod selekcji cech na czas obliczeń i dokładność wykrywania intruzji przy użyciu popularnych klasyfikatorów.

Opis bazy danych NSL-KDD

Zbiór danych **NSL-KDD** został opracowany jako ulepszona wersja popularnego, lecz krytykowanego zbioru KDD'99. Główne cechy NSL-KDD to:

- Brak redundantnych rekordów w zbiorze uczącym, co zapobiega uprzedzeniom klasyfikatorów wobec często występujących danych.
- Brak duplikatów w zbiorze testowym, co zapewnia bardziej wiarygodną ocenę efektywności algorytmów.
- Zrównoważony podział danych według poziomu trudności, co pozwala na bardziej wszechstronną ocenę metod klasyfikacji.
- Rozsądna liczba rekordów w zbiorach uczących i testowych, co umożliwia przeprowadzanie eksperymentów na pełnym zbiorze bez konieczności losowej redukcji danych.

Zbiór danych zawiera pliki w różnych formatach, m.in. pełne zbiory uczące i testowe, ich podzbiory oraz dane w formacie **ARFF** i **CSV**.

Cel i problem badawczy

Podstawowym problemem, który staramy się rozwiązać, jest redukcja czasu obliczeń w procesie detekcji intruzji, przy jednoczesnym zachowaniu akceptowalnej dokładności klasyfikacji.

Wielowymiarowość danych oraz duża liczba cech zwiększają złożoność obliczeniową, co wpływa negatywnie na efektywność systemów IDS.

W projekcie zastosujemy algorytm TOPSIS, który umożliwia ocenę i wybór najbardziej efektywnych metod selekcji cech spośród różnych alternatyw. Algorytm TOPSIS polega na wyborze rozwiązania najbliższego rozwiązaniu idealnemu, z uwzględnieniem wielu atrybutów.

Metodyka

1. **Wybór zbioru danych**
    - W projekcie wykorzystamy zbiór **NSL-KDD,** zbiór ten zawiera oznaczenia ataków i poziom trudności klasyfikacji.
2. **Selekcja cech**
    - Zastosujemy dziesięć różnych technik selekcji cech, a wyniki zostaną ocenione za pomocą algorytmu TOPSIS.
3. **Klasyfikacja**
    - Różne rodzaje klasyfikatorów zostaną wykorzystane, do uzyskania metryk pozwalających na dalszą analizę przy użyciu algorytmu TOPSIS.
4. **Analiza wyników**
    - Wyniki TOPSIS zostaną obliczone w środowisku Python, co pozwoli na rangowanie technik selekcji cech pod kątem efektywności.

Oczekiwane efekty

- **Skrócenie czasu obliczeń**: Zastosowanie efektywnych metod selekcji cech pozwoli na redukcję wymiarowości danych, co przełoży się na krótszy czas detekcji intruzji.
- **Poprawa dokładności**: Zachowanie odpowiedniego balansu między redukcją cech a precyzją klasyfikacji.
- **Ocena technik selekcji cech**: Wytypowanie najlepszej metody selekcji cech na podstawie wyników algorytmu TOPSIS.
- **Zunifikowane wyniki**: Możliwość porównania efektywności różnych technik i klasyfikatorów na ujednoliconych danych, co zapewni spójność i powtarzalność badań.

Podsumowanie

Projekt ten ma na celu stworzenie podstaw dla bardziej wydajnych systemów IDS poprzez zastosowanie zaawansowanych metod selekcji cech oraz analizy klasyfikacji. Wykorzystanie zmodyfikowanego zbioru NSL-KDD, narzędzi takich jak Python oraz algorytmu TOPSIS pozwoli na zidentyfikowanie optymalnych rozwiązań dla redukcji czasu obliczeń w detekcji intruzji. Dzięki temu możliwe będzie opracowanie bardziej efektywnych i precyzyjnych systemów ochrony sieci.

Artykuł referencyjny:

https://www.researchgate.net/publication/269399129_TOPSIS_Based_Multi-Criteria_Decision_Making_of_Feature_Selection_Techniques_for_Network_Traffic_Dataset

Baza danych:

https://www.kaggle.com/datasets/hassan06/nslkdd/data