Michał Ściubisz
Wojciech Tokarz

Zastosowanie algorytmów rankingowych do wyboru najlepszego klasyfikatora w systemach wykrywania intruzji na podstawie zbioru danych NSL-KDD przy wykorzystaniu różnych metryk oceny.

Wprowadzenie

Systemy wykrywania intruzji (IDS) odgrywają kluczową rolę w ochronie sieci przed atakami i nieautoryzowanym dostępem. Współczesne IDS muszą przetwarzać miliony pakietów danych, co stawia wysokie wymagania w zakresie precyzji i szybkości detekcji anomalii. W związku z tym kluczowe staje się opracowanie efektywnych metod oceny i wyboru klasyfikatorów, które najlepiej spełniają wymagania konkretnego systemu IDS.

W ramach tego projektu wykorzystamy zmodyfikowany zbiór danych NSL-KDD oraz algorytmy rankingowe, takie jak TOPSIS, VIKOR i AHP. Celem jest ocena skuteczności różnych klasyfikatorów w systemach IDS na podstawie wybranych metryk, takich jak dokładność, F1-score, czy czas obliczeń, a także wyznaczenie najlepszego klasyfikatora do wykrywania intruzji w określonym środowisku.

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 **TXT**.

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 algorytmy rankingowe, takie jak TOPSIS, VIKOR oraz AHP, które umożliwiają ocenę i wybór najbardziej efektywnych metod klasyfikacji spośród różnych alternatyw. Algorytmy te pozwalają na uwzględnienie wielu atrybutów jednocześnie, w celu wyboru rozwiązania najbliższego rozwiązaniu idealnemu. Dzięki zastosowaniu kilku technik wielokryterialnej optymalizacji zostanie zbadany wpływ różnych metod klasyfikacji na czas obliczeń i dokładność wykrywania intruzji.

Metodyka

- **Wybór zbioru danych**
  - W projekcie wykorzystamy zbiór NSL-KDD, który zawiera oznaczenia ataków i poziom trudności klasyfikacji.
- **Klasyfikacja**
  - Stosujemy różne rodzaje klasyfikatorów, aby uzyskać metryki, takie jak dokładność, F1-score, i czas obliczeń, które posłużą do dalszej analizy przy użyciu algorytmów rankingowych.
- **Ranking klasyfikatorów**
  - Wyniki klasyfikatorów zostaną ocenione za pomocą algorytmów rankingowych, takich jak TOPSIS, VIKOR oraz AHP.
- **Analiza wyników**
  - Obliczenia algorytmów rankingowych zostaną przeprowadzone w środowisku Python, co pozwoli na rangowanie klasyfikatorów pod kątem ich efektywności i przydatności w systemach IDS.

Oczekiwane efekty

- **Skrócenie czasu obliczeń**: Zastosowanie efektywnych metod oceny klasyfikatorów pozwoli na wybór takich algorytmów, które minimalizują czas detekcji intruzji przy zachowaniu wysokiej efektywności.
- **Poprawa dokładności**: Wybór najlepszego klasyfikatora umożliwi zachowanie odpowiedniego balansu między szybkością działania a precyzją klasyfikacji.
- **Ocena technik rankingowych**: Wytypowanie najlepszych algorytmów rankingowych na podstawie wyników takich metod jak TOPSIS, VIKOR oraz AHP.
- **Zunifikowane wyniki**: Możliwość porównania efektywności różnych 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 oceny i wyboru klasyfikatorów. Wykorzystanie zmodyfikowanego zbioru NSL-KDD, narzędzi takich jak Python oraz algorytmów rankingowych, takich jak TOPSIS, VIKOR i AHP, pozwoli na zidentyfikowanie optymalnych rozwiązań pod względem czasu obliczeń i dokładności w detekcji intruzji. Dzięki temu możliwe będzie opracowanie bardziej efektywnych i precyzyjnych systemów ochrony sieci, które lepiej sprostają współczesnym wyzwaniom związanym z cyberbezpieczeństwem.

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