

Metoda BRF

zbalansowane lasy losowe

Cel projektu

Zbudowanie modelu prognostycznego (tj. klasyfikatora) zdolnego do rozróżniania połączeń “złych”, zwanych atakami lub włamaniami, oraz “dobrych” (normalnych) połączeń

Trudności

Uczenie klasyfikatora z **niezrównoważonych danych**

Niezrównoważone dane – klasy nie są w przybliżeniu równoliczne

Dlaczego problem nierównoważonych
danych jest ważny?

Jak rozwiązać problem? 😊

Las losowy

- Metoda łączenia wielu drzew klasyfikacyjnych
 - Losujemy K prób bootstrapowych – dla każdej konstruujemy drzewo klasyfikacyjne.
 - W każdym węźle losujemy m cech, które będą uczestniczyły w metodzie najlepszego podziału.
 - Drzewa budowane są bez przycinania
 - Ostateczna obserwacja klasyfikowana jest przez metodę głosowania
 - Jedynym parametrem metody jest współczynnik m

Problemy lasu losowego

- Lasy losowe skupiają się bardziej na przewidywaniu trafności klas większościowych, co skutkuje słabą dokładnością klas mniejszościowych

Rozwiązanie - zbalansowany las losowy

- Dla każdego pojedynczego drzewa losuje się podpróbę bootstrapową - połowa przypadków jest losowana w sposób niezależny z klasy mniej licznej, a druga połowa – również w sposób niezależny – z klasy bardziej licznej
- Buduje się model drzewa przy pomocy algorytmu CART bez opcji przycinania; na każdym etapie podziału wykorzystuje się losowo dobrany zestaw predyktorów poddawanych ocenie.
- Pierwszy i drugi etap powtarza się zadaną liczbę razy
- Agreguje się pojedyncze drzewa w celu uzyskania finalnego modelu predykcyjnego.