Paweł Świątkowski

Grzegorz Aleksiuk

**Podstawy sztucznej inteligencji – Projekt 2**

**MM.DD.R3 Random Forest – Forest Fires**

1. Treść

Zaimplementować algorytm lasu losowego (ang. Random Forest) do predykcji. Zbiór danych do użycia: Forest Fires - https://archive.ics.uci.edu/ml/datasets/Forest+Fires. Uzyskane rezultaty porównać z wynikami dla wybranej implementacji algorytmu ML z dostępnych bibliotek np. Scikit-learn, WEKA, MLlib, Tensorflow/Keras etc.

1. Przyjęte założenia, doprecyzowanie treści

Projekt został zaimplementowany z wykorzystaniem Javy, natomiast dla porównania jakości rozwiązania została zastosowana biblioteka pythona Scikit-learn (RandomForestRegressor).

1. Podział odpowiedzialności
   * Paweł Świątkowski:
     + Stworzenie struktur danych do implementacji zadania w Javie
     + Implementacja tworzenia drzewa decyzyjnego
     + Zapisanie struktury zbudowanego drzewa
   * Grzegorz Aleksiuk:
     + Implementacja budowy lasu losowego
     + Implementacja przycinania drzewa decyzyjnego
     + Implementacja k-krotnej walidacji krzyżowej
     + Implementacja zadania z wykorzystaniem pythona (Scikit-learn)
2. Opis algorytmu
   * Budowanie drzewa decyzyjnego – algorytm CART – jako wejście algorytmu przyjmowane są próbki danych oraz argumenty względem których będzie odbywał się podział.

W trakcie budowania drzewa decyzyjnego wybierany jest argument oraz jego wartość względem jakiego będzie odbywał się podział – ten wybór następuje poprzez sprawdzenie każdego możliwego podziału i wybraniu takiego, dla którego dostajemy najmniejszy błąd średniokwadratowy. Takie podziały następują aż do czasu kiedy osiągnęliśmy maksymalną zdefiniowaną głębokość drzewa, bądź w danym wierzchołku drzewa mam już tylko 3 próbki danych. Dodatkowy warunek wynika z implementacji przycinania drzewa w trakcie jego budowy tj. jeżeli błąd średniokwadratowy przed podziałem był mniejszy niż po podziale.

* + Budowanie lasu losowego – jako wejście przyjmujemy wszystkie próbki danych.

Tworzonych jest zbiorów. W każdym ze zbiorów znajduje się B próbek danych, losowanych z zbioru początkowego ze zwracaniem (próbki w danym tworzonym zbiorze mogą się powtarzać). Następnie dla każdego z tych zbiorów losowane różnych argumentów (losowanie bez zwracania).

Dla każdego ze stworzonych zbiorów budowane jest drzewo decyzyjne.

Zbiór stworzonych drzew stanowi las losowy.

Wynik dla nowo wprowadzonej próbki uzyskuje się, poprzez wzięcie średniej wartości z wyników zwracanych przez drzewa decyzyjne.

* + K-krotna walidacja krzyżowa – jest stosowana do oceny jakości stworzonego modelu przy zachowaniu jak największej ilości danych do zbioru trenującego.

Dane są dzielone losowo na K możliwie równych zbiorów T1, T2 … Tk. Następnie model lasu losowego jest uczony na zbiorze T – Ti (zbiór danych treningowych) natomiast zbiór danych testowych to Ti. Na podstawie zbioru testowego wyliczany jest błąd średniokwadratowy modelu Ei który zapisujemy. Taka operacja jest powtarzana dla i = 1, 2 … K. Po tym model jest uczony na pełnym zbiorze danych T, a jego błąd średniokwadratowy to średnia wyliczonych błędów E1,E2 … Ek.