Uczenie Maszynowe - projekt

Tomasz Owienko Wojciech Zarzecki

18.11.2023

1 Cel projektu

Celem projektu jest implementacja zmodyfikowanej wersji algorytmu generowania lasu losowego, w której do generowania kolejnych drzew losowane są częściej elementy ze zbioru uczącego, na których dotychczasowy model się mylił.

Istotą metody lasu losowego w problemach regresji i klasyfikacji jest redukcja wariancji i nadmiernego dopasowania osiąganego przez pojedyncze drzewa decyzyjne. W klasycznych algorytmach generowania lasu losowego każde z B drzew generowane jest na podstawie $\sqrt{B}$ przykładów ze zbioru trenującego wylosowanych ze zwracaniem, zazwyczaj ograniczonych (w problemach klasyfikacji) do $\sqrt{|D|}$ atrybutów wylosowanych bez zwracania, gdzie D jest zbiorem atrybutów. Proces ten odbywa się w jednej iteracji - algorytm kończy pracę po wygenerowaniu B drzew.

Realizowany projekt zakłada wykorzystanie metod boostingowych do iteracyjnego poprawiania wyniku algorytmu generowania lasu losowego. Mechanizm losowania atrybutów do generowania kolejnych drzew zostanie zmodyfikowany przez wprowadzenie preferencji dla tych przykładów, na których dotychczasowy model się mylił.

// i co dalej?? Nowe drzewa dołączamy do lasu (de facto AdaBoost) czy tworzymy nowy las? (w drugim przypadku to nie jest boosting)

2 Opis algorytmu

TODO wzorki i coś o AdaBoost

Algorithm 1 TrainRandomForest

Input: $U \neq \emptyset$: zbiór przykładów trenujących, C: klasy przykładów, D: zbiór atrybutów wejściowych

// przy założeniu, że każda iteracja to nowy las
$F \leftarrow \emptyset$
$\hat{C} \leftarrow \emptyset$
for $i \leftarrow 1$ to N **do**
    $F_i \leftarrow \emptyset$
    for $b \leftarrow 1$ to B **do**
        $U_b \leftarrow$B elementów wylosowanych z U z preferencją dla $\{u_j \in U : \hat{C}(u_j) \neq C(u_j)\}$
        $D_b \leftarrow \sqrt{|D|}$ atrybutów wylosowanych z D bez zwracania
        $f_b \leftarrow$ drzewo decyzyjne wygenerowane na podstawie C, U_b i D_b
        $F_i \leftarrow F_i \cup \{f_b\}$
    end for
    $F \leftarrow$ las losowy F ulepszony za pomocą F_i
    for $u \in U$ **do**
        $\hat{C} \leftarrow \hat{C} \cup$ PredictRandomForest(u, F)
    end for
end for
return F

3 Planowane eksperymenty

Algorithm 2 PredictRandomForest

Input: x: wektor wejściowy, F: nauczony las losowy

1: $\hat{C} \leftarrow \emptyset$
2: **for** $f_b \in F$ **do**
3: $\quad \hat{C} \leftarrow \hat{C} \cup f_b(x)$
4: **end for**
5: **return** najczęstsza klasa z $\hat{C}$

4 Wykorzystywane zbiory danych