

DRZEWA DECYZYJNE – wstęp teoretyczny

Modele drzew klasyfikacyjnych i regresyjnych (CART, ang. *Classification and Regression Trees*), jak sama nazwa mówi, służą zarówno do rozwiązywania problemów regresyjnych (gdzie zmienną zależną jest cecha ilościowa – ciągła/liczbowa) jak i klasyfikacyjnych (zmienna zależna jakościowa – kategoryczna). Najogólniej, celem analizy z zastosowaniem algorytmu budowy drzew decyzyjnych jest znalezienie zbioru logicznych warunków podziału, typu *jeżeli, to*, prowadzących do jednoznacznego zaklasyfikowania obiektów.

Drzewa decyzyjne służą do wyboru deskryptorów o największym wpływie na modelowaną wielkość (najbardziej znaczących). Technika ta polega na „wzrastaniu drzewa” tj. dzielenia związków na wzajemnie wykluczające się grupy – węzły (ang. *nodes*). Linie łączące węzły nazywa się gałęziami (ang. *branches*). Algorytm rozpoczyna się od węzła głównego – korzenia (ang. *root*) – zawierającego wszystkie związkki, które następni dzielone są na węzły podzielne. Końcowe węzły, które nie podlegają podziałom to liście (ang. *leaves*). Każdy podział określa reguła (próg) uwzględniająca wartości wybranego na danym etapie deskryptora.

Zarówno w przypadku klasycznych modeli jakościowych (SAR, ang. *Structure-Activity Relationships*), jak również modeli ilościowych (QSAR, ang. *Quantitative Structure-Activity Relationships*) związkki dzielone są na dwa zbiory – uczący (wykorzystywany do opracowania drzewa decyzyjnego) oraz walidacyjny (służący do oceny zdolności predykcyjnych drzewa decyzyjnego).

W przypadku **drzew klasyfikacyjnych** deskryptory wybierane są pod kątem najmniejszego prawdopodobieństwa błędnej klasyfikacji, co oznacza, że binarny podział wykonywany z opracowaną regułą powinien prowadzić do maksymalnie dwóch jednorodnych grup związków. Prawdopodobieństwo błędnej klasyfikacji mierzy się za pomocą indexu Giniego, wyrażonego wzorem:

$$G = 1 - \sum_{j=1}^c \left(\frac{n_j}{n} \right)^2$$

gdzie n_j jest liczbą związków z klasy j zawartych w węźle.

Do weryfikacji zdolności predykcyjnych modeli jakościowych służą miary statystyczne:

$$\text{Sensitivity}(\text{recall}, \text{positiverate}) = TP / (TP + FN)$$

$$\text{Specificity} = TN / (FP + TN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$F1 \quad (\text{harmonic mean of precision\&sensitivity}) = (2 \times TP) / (2 \times TP + FP + FN)$$

$$\text{Balanced accuracy} = (\text{Sensitivity} + \text{Specificity}) / 2$$

$$\text{Balanced error} = 1 - \text{Balanced accuracy}$$

		Predicted	
		Active	Inactive
Observed	Active	True positive (TP)	False positive (FP)
	Inactive	False negative (FN)	True negative (TN)

Figure 1. Confusion matrix describing the performance of a classification model (or ‘classifier’) on a set of test data for which the true values are known.

Wybór deskryptorów w przypadku **drzew regresyjnych** dokonywany jest przy pomocy metody najmniejszych kwadratów, czyli tak aby suma kwadratów różnic pomiędzy wartościami przewidywanymi przez model a zmierzonymi eksperymentalnie (tzw. rezyduałów) była jak najmniejsza.

Sprawozdanie:

Regresja (DT):

1. Wykonać model drzewa decyzyjnego z całą pulą deskryptorów.
2. Wykonać feature selection z podstawowymi parametrami dla algorytmu Backward.
3. Przygotować optymalizację maksymalnej głębokości drzewa z wykresem statystyk generowanych przez model dla różnych głębokości.
4. Przygotować optymalizację minimalnej liczby próbek w liściu (`min_samples_leaf`) z wykresem statystyk generowanych przez model dla różnych wartości.
5. Wykonać jednoczesną optymalizację głębokości drzewa i minimalnej liczby próbek w liściu z odpowiadającymi im statystykami.
6. Powtórzyć pkt 1-5 dla drzewa decyzyjnego z przycinaniem (pruning) lub ustawieniem minimalnego kryterium czystości liści.
7. Napisać wnioski wynikające z powyższych analiz z uwzględnieniem tego, które deskryptory są najczęściej wybierane do różnych modeli.

Klasyfikacja:

1. Wykonać model drzewa decyzyjnego z całą pulą deskryptorów.
2. Wykonać feature selection z podstawowymi parametrami dla algorytmu Backward.
3. Przygotować optymalizację maksymalnej głębokości drzewa z wykresem statystyk generowanych przez model dla różnych głębokości.
4. Przygotować optymalizację minimalnej liczby próbek w liściu (min_samples_leaf) z wykresem statystyk generowanych przez model dla różnych wartości.
5. Wykonać jednoczesną optymalizację głębokości drzewa i minimalnej liczby próbek w liściu z odpowiadającymi im statystykami.
6. Powtórzyć pkt 1-5 dla drzewa decyzyjnego z przycinaniem (pruning) lub ustawieniem minimalnego kryterium czystości liści.
7. Napisać wnioski wynikające z powyższych analiz z uwzględnieniem tego, które deskryptory są najczęściej wybierane do różnych modeli.