

DNA C4.5

Rafał Kwiatkowski, Franciszek Sioma

25 maja 2020

1 Opis projektu

1.1 Cel projektu

Celem projektu *DNA C4.5* było stworzenie aplikacji wykorzystującej algorytm C4.5 do klasyfikacji sekwencji DNA na podstawie zadanych zbiorów.

1.2 Przyjęte założenia

1.3 Wkład autorów

- Algorytm C4.5 - Rafał Kwiatkowski
- Testy i eksperymenty - Franciszek Sioma
- Dokumentacja - Franciszek Sioma

1.4 Decyzje projektowe

1.5 Wykorzystane narzędzia i biblioteki

Do napisania aplikacji użyliśmy języka Python w wersji: 3.8, dokumentacja została stworzona przy użyciu języka Latex, a IDE z którego korzystaliśmy to Visual Studio Code. Użyliśmy również systemu kontroli wersji Git. Link do repozytorium: <https://github.com/Rolfrider/C4.5-Gene-Splicing>

2 Uruchamianie aplikacji i odtworzenie wyników testów

W przypadku tego projektu pracowaliśmy na konkretnym zbiorze danych i nasza aplikacja służy tylko do przeprowadzenia testów, dlatego jest tylko jeden sposób jej uruchomienia. W celu odtworzenia przeprowadzonych testów należy wykonać komendę:

```
python app.py
```

3 Eksperymenty

W naszym projekcie przeprowadziliśmy walidacji krzyżowej dla parametru k równego 10 dla drzewa utworzonego przez algorytm ID3 oraz dla drzewa utworzonego przez C4.5. Test został powtórzony 100 razy, a uśrednione wyniki zostały przybliżone poniżej.

3.1 Wyniki

Algorytm	Dopasowanie
ID3	81,5%
C4.5	81,4%

Tablica 1: Wyniki walidacji krzyżowej dla akceptorów

Algorytm	Dopasowanie
ID3	83,2%
C4.5	81,7%

Tablica 2: Wyniki walidacji krzyżowej dla donorów

Z wyników walidacji wynika, że drzewo przed zastosowaniem algorytmu C4.5 daje lepszą dokładność wyników niż po. Algorytm C4.5 przyspiesza działanie programu na utworzonym drzewie oraz pozwala uniknąć nadmiernego dopasowania. Wyniki świadczą o tym, że w naszym przypadku zjawisko to nie występowało na tyle często by polepszyć końcowy wynik, a wręcz przeciwnie.