Podstawy Sztucznej Inteligencji
Zespół: BT.TM: Karbownik Patrycja, Grabowski Kamil, Aksamit Sebastian

Dokumentacja wstępna

Zadanie:

Stworzenie własnego algorytmu (ewolucyjnego lub genetycznego) do rekonstrukcji drzewa filogenetycznego i porównanie wyników do innych algorytmów rekonstrukcji.

Informacje wstępne:

Drzewo filogenetyczne - graf acykliczny prezentujący drogę rozwoju, pochodzenie i zmiany ewolucyjne grupy organizmów (zazwyczaj gatunków); graf przedstawiający ewolucyjne zależności pomiędzy sekwencjami lub gatunkami wszystkich organizmów.

Porównywanie gatunków na podstawie sekwencji kwasów nukleinowych jest uznawane za wiarygodny i precyzyjny wyznacznik stopnia pokrewieństwa. Pozwala też wyznaczyć czas powstania mutacji / specjacji (proces biologiczny w wyniku którego powstają nowe gatunki).

Opis problemu:

Ilość możliwych drzew zależy od ilości taksonów/liści (ich ilość oznaczymy jako *n)*. Istnieją drzewa ukorzenione, gdzie korzeń reprezentuje “praprzodka” (jest ich $\frac{(2n-3)!}{2^{n-2}*(n-2)!}$), a także drzewa nieukorzenione, których jest $\frac{(2n-5)!}{2^{n-3}*(n-3)!}$. Taka ilość możliwych drzew uniemożliwia sprawdzenia wszystkich możliwości w celu wybrania najlepszego rozwiązania już przy liczbie kilkunastu taksonów. Problem stworzenia takiego drzewa jest NP-zupełny.

W ramach pierwszego projektu zbadamy skuteczność wykorzystywania algorytmu ewolucyjnego do tworzenia zadowalającego drzewa końcowego.

Założenia implementacji:

- Sposób przechowywania danych:
  - sekwencje genetyczne - pliki *.txt* / *.fasta* / *.fastq* / inne (decyzja zostanie podjęta we wstępnym etapie implementacji)
  - podobieństwo sekwencji - symetryczna macierz o rozmiarze *n x n*, gdzie *n* - ilość sekwencji
- Sposób prezentacji drzewa - (decyzja zostanie podjęta we wstępnym etapie implementacji)
- Język programowania: Python

Wstępny opis algorytmu (ewolucyjnego):

1. Tworzymy *k* losowych macierzy substytucji.
2. Na podstawie stworzonych macierzy i sekwencji wejściowych, tworzymy *k* różnych drzew filogenetycznych:
   a. Obliczamy podobieństwo każdej pary sekwencji - *algorytm Needlemana-Wunscha*, i zapisujemy je w macierzy podobieństwa.
   b. Sprawdzamy ile zostało sekwencji do złączenia:

   i. Jeśli została 1 - kończymy
   ii. W przeciwnym razie - łączymy ze sobą dwie najbardziej podobne sekwencje, wprowadzamy korekcje w macierzy podobieństwa, obliczamy długości gałęzi, redukujemy macierz i wracamy do punktu 2b
3. Oceniamy stworzone drzewa filogenetyczne:
   a. Jeśli został spełniony warunek stopu, kończymy działanie algorytmu
   b. Jeśli nie, modyfikujemy macierze substytucji wg ustalonej funkcji (ustalenie przewidziane w następnych etapach implementacji) i przechodzimy do punktu 2.

Ocena drzewa:

Proponowanym sposobem oceny drzewa jest tzw. “*bootstraping*”. W statystyce metoda ta polega na permutacji danych wejściowych na np. 100 różnych sposobów, wygenerowanie danych wyjściowych i obliczeniu prawdopodobieństwa, z jakim dane się powtórzyły. Na bazie wyników takiego zabiegu jesteśmy w stanie przeanalizować i ocenić drzewo.

Jest to metoda docelowa, może się jednak okazać nieefektywna lub zbyt złożona do implementacji, wtedy zostanie zastąpiona odpowiednią inną metodą lub narzędziem.